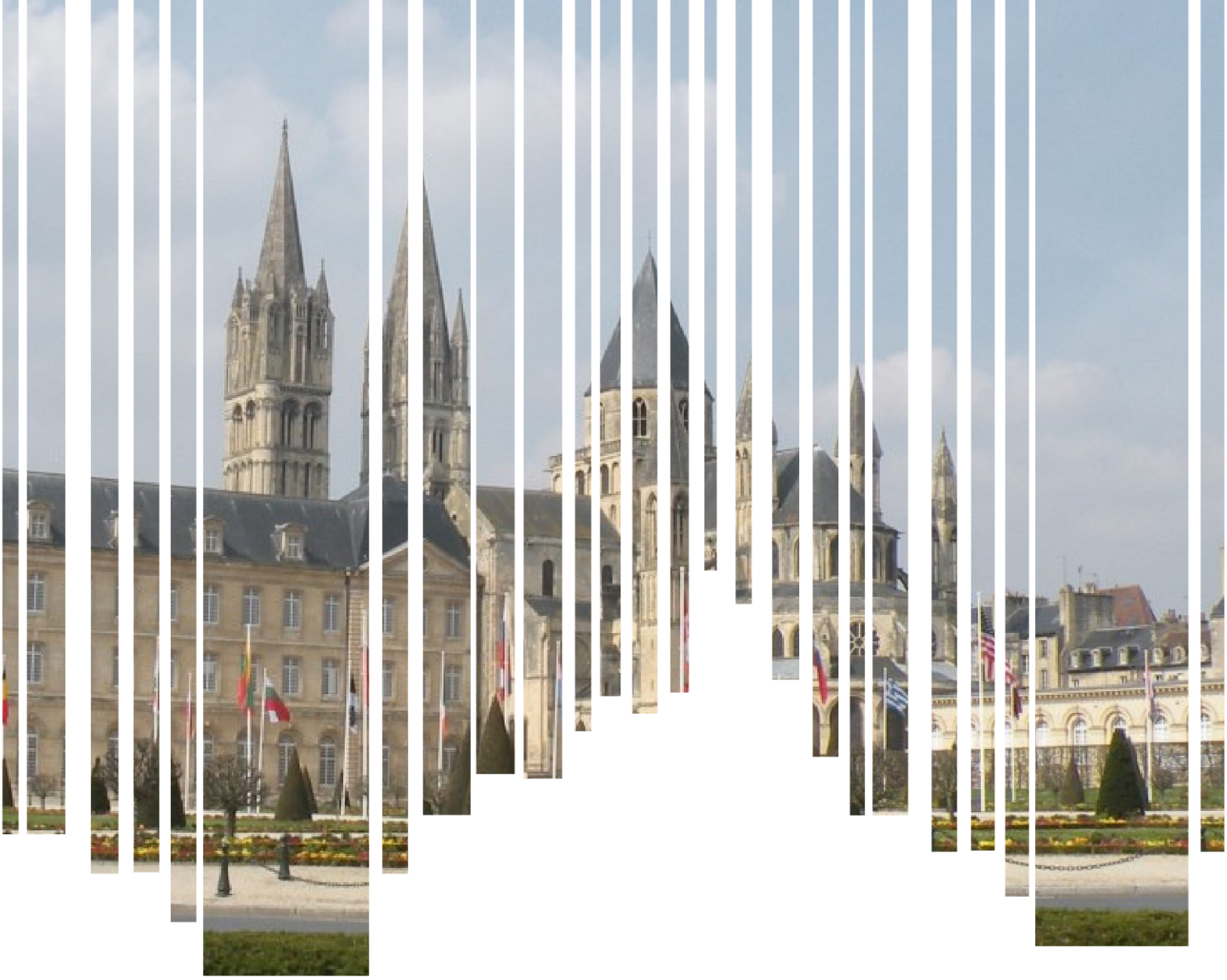


Actes du colloque CORESA 2017

20 & 21 novembre 2017
Caen

GREYC

Information Signal Image viSion

Pré-GDR Sécurité Informatique

UNIVERSITÉ CAEN NORMANDIE

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

Table des matières

Mots des co-présidents du colloque

Pour sa 19ème édition, le colloque CORESA s'installe cette année en Normandie les 20 et 21 novembre 2017. La conférence a lieu au château de Caen (l'un des plus grands d'Europe) construit par Guillaume le Conquérant qui devint Roi d'Angleterre en 1066.

Comme pour les éditions précédentes, CORESA porte sur des recherches et des études en cours dans le domaine de la représentation des signaux audiovisuels et plus spécifiquement sur le traitement des images, des objets graphiques, de la vidéo et du son pour les applications multimédia.. Cette année, l'accent est mis sur le thème de la biométrie avec pas moins de trois sessions orales dédiées. Des applications telles que la reconnaissance ou l'authentification des individus, la protection de la donnée biométrique, la protection du contenu, le traitement des données de vidéo-protection seront exposées.

L'édition 2018 est labellisée par le GdR-ISIS et le Pré-Gdr Sécurité Informatique du CNRS.

Deux conférences invitées mettront en avant deux domaines dans ses développements les plus récents : Christoph Busch de l'Université de Darmstadt traitera de biométrie et Alan C. Bovik abordera les problématiques de qualité d'images et vidéos.

Tous ces éléments permettent d'avoir un programme de qualité qui repose sur le travail important des auteurs mais aussi des experts pour les relectures. Chaque soumission a été évaluée par deux ou trois membres du comité de programme. Parmi la trentaine de soumissions, 28 ont été retenues dans des sessions orales . Nous tenons donc, au nom du comité d'organisation, à remercier toutes les personnes (invités, auteurs, membres des différents comités, partenaires) qui ont contribué à la qualité de cette 19ème édition du colloque CORESA

Christophe Charrier (UNICAEN, GREYC)
Christophe Rosenberger (ENSICAEN, GREYC)

Comité de Pilotage

W. Puech (LIRMM, Université Montpellier)
V. Charvillat (IRIT, ENSEEIHT)
M. Daoudi (TELECOM Lille1, LIFL)
F. Dupont (LIRIS, Université Claude Bernard Lyon 1)
M. Antonini (I3S, Université de Nice Sophia-Antipolis, CNRS)
L. Morin (IETR, Université de Rennes)
J. Jung (Orange Labs)

Comité local d'organisation

Co-responsables

C. Charrier (GREYC, Université de Caen Normandie)
C. Rosenberger (GREYC, ENSICAEN)

Membres

M. Barbier (BREYC, ENSICAEN)
S. Bougleux (GREYC, Université de Caen Normandie)
E. Cherrier (GREYC, ENSICAEN)
B. Hemery (GREYC, ENSICAEN)
P. Lacharme (GREYC, ENSICAEN)
J-M Le Bars (GREYC, Université de Caen Normandie)
O. Lézoray (GREYC, Université de Caen Normandie)

Comité de Programme

M. Antonini
O. Auberton
A. Bartoli
B. Ben Amor
J. Benois-Pineau
S. Bouakaz
M. Cagnazzo
A. Caplier
C. Cavaro-Menard
F. Cayre
R. Chaine
M. Chaumont
C. Charrier
V. Charvillat
F. Davoine
E. Dellandréa
G. Doërr
F. Dufaux
F. Dupont
C. Fernandez-Maloigne
C. Fontaine
Y. Gaudeau
G. Gesquière
F. Ghorbel
D. Hamad
J. Jung
G. Lavoué
S. Marchand
J. Martinet
E.M. Mouaddib
M.C. Larabi
L. Morin
J-M. Moureaux
H. Nicolas
A. Ouled Zaid
F. Payan
D. Petrovska
W. Puech
J. Ronsin
C. Rosenberger
G. Subsol
A. Taleb-Ahmed
J. Tierny
A. Trémeau
J-P. Vandeborre
S. Valette
B. Vozel

Analyse d'image, de la vidéo, et des données 3D

Session 1

Recalage de nuages de points 3D issus d'acquisitions LiDAR de scènes structurées fondé sur des a priori géométriques

J. Sanchez[1], P. Checchin[2], F. Denis[1], F. Dupont[1], L. Trassoudaine[2]

[1]Univ. Lyon-LIRIS UMR 5205 CNRS
Université Claude Bernard Lyon 1
43, bd du 11 novembre 1918
F. 69622 VILLEURBANNE CEDEX

[2]Institut Pascal UMR 6602 CNRS
Campus universitaire des Cézeaux
4 Avenue Blaise Pascal
F. 63178 AUBIERES CEDEX

julia.sanchez@univ-lyon1.fr, paul.checchin@uca.fr, florence.denis@liris.cnrs.fr,
florent.dupont@liris.cnrs.fr, laurent.trassoudaine@univ-bpclermont.fr

Résumé

L'utilisation de capteurs LiDAR pour obtenir des données 3D implique l'acquisition de scans suivant différents points de vue. Dans les systèmes actuels, l'algorithme d'ICP (Iterative Closest Point) est largement utilisé pour recaler les scans entre eux. Cependant, cette méthode se heurte à des problèmes de minima locaux et ne fonctionne que pour de faibles mouvements. Cet article développe une nouvelle méthode de recalage adaptée aux environnements structurés et basée sur des caractéristiques géométriques. La rotation et la translation de la transformation totale recherchée sont calculées de manière successive en utilisant respectivement l'image Gaussienne des nuages de points et une corrélation d'histogrammes. L'évaluation de notre algorithme sur deux ensembles de scans 3D comparé à six méthodes existantes montre que la méthode proposée est plus robuste à de faibles résolutions de scans, à la complexité de la scène et au bruit du capteur. De plus la faible durée de notre algorithme permet une implémentation temps réel du recalage.

Mots clefs

LiDAR ; Recalage ; Image gaussienne ; Nuages de points.

1 Introduction

Les acquisitions 3D sont utilisées dans de nombreux domaines industriels tels que l'inspection de bâtiments [1], ou la navigation autonome [2]. Les données sont collectées sous forme de nuages de points mesurés dans le repère capteur dans des poses différentes. Le recalage est le processus ayant pour but d'estimer la transformation rigide qui aligne un nuage appelé source sur un nuage appelé cible. Cela permet de lier les scans afin de ne former qu'un seul nuage de points qui pourra être analysé par la suite. La phase d'acquisition doit être rapide, nécessiter peu de mémoire et consommer un minimum d'énergie afin que le système puisse être embarqué. De nombreux capteurs sont apparus sur le marché récemment pour répondre à ces critères [2], mais les algorithmes de recalage limitent encore les performances du système complet. En environnement intérieur, ce qui constitue le contexte majeur de cet article, les algorithmes existants présentent un manque de précision et un taux de réussite du recalage trop bas. La plupart sont très coûteux en temps et nécessitent une qualité de données généralement élevée.

L'algorithme le plus utilisé dans les chaînes de traitement commercialisées est ICP (Iterative Closest Point) [3]. Cet algorithme attribue itérativement des correspondances entre les points de la source et les points de la cible, filtre ces correspondances et minimise la distance entre les points de chaque paire créée. Le processus itératif de l'algorithme rend l'implémentation en temps réel compliquée à mettre en place. De plus, la fonction de coût n'étant pas convexe, la minimisation peut mener à un minimum local. Pour répondre à ce problème, une solution est d'assurer un mouvement faible entre les scans en entrée de l'algorithme, soit en adaptant le processus d'acquisition, soit en cherchant un alignement approximatif initial avant le traitement par ICP.

Cependant, aucune méthode proposée jusqu'à aujourd'hui ne permet de résoudre des problèmes de recalage quel que soit le capteur, le chevauchement des nuages de points et leur position initiale. Dans cet article, nous cherchons à développer une nouvelle méthode pour réaliser un recalage sans utiliser de correspondances locales ni d'alignement initial et qui pourrait être utilisé sur des nuages présentant un faible chevauchement et des motifs répétitifs. La section 2 décrit les méthodes connexes au travail réalisé dans cet article. La section 3 développe la méthode proposée. Enfin, l'algorithme est évalué sur deux jeux de données dans la section 4.

2 Etat de l'art

La première méthode proposée pour recaler des nuages de points était fondée sur la projection sur la sphère gaussienne [4]. Cette projection correspond à la représentation

des normales comme des points sur une sphère de rayon unité. L'idée principale est d'échantillonner respectivement l'image gaussienne et l'espace des rotations et d'évaluer la similarité des sphères après transformation par chacune des rotations. Cette méthode reste très approximative, la difficulté majeure résidant dans l'échantillonnage de l'espace des rotations qui doit être le plus uniforme possible et couvrir un maximum de rotations. Les méthodes EGI (Extended Gaussian Sphere) et CEGI (Complex Extended Gaussian Sphere) [5] permettent de retrouver successivement la rotation comme décrit précédemment et la translation entre nuages de points en ajoutant une information de distance des plans à l'origine. Cependant, ces méthodes ne peuvent pas prendre en compte un chevauchement partiel et ne peuvent donc pas être utilisées dans le contexte de notre travail.

L'algorithme ICP a rapidement surpassé les méthodes précédentes. Par la suite, de nombreux travaux ont eu pour but d'améliorer la robustesse de cet algorithme. La plupart d'entre eux sont présentés dans [6]. Certaines méthodes se concentrent sur la métrique de la fonction de coût. En effet, les correspondances trouvées entre points ne peuvent être exactes à cause de la différence des échantillonnages et du bruit introduit par les capteurs. Les variantes ICP *point à plan* [7] et ICP *généralisé* [8] ont été proposées pour remédier à ce problème. D'autres méthodes améliorent le processus de minimisation pour le rendre plus robuste aux minima locaux et pour accélérer l'optimisation. Dans [9] par exemple, l'algorithme de *Levenberg-Marquardt* est utilisé pour trouver l'optimum de la fonction. Récemment, l'algorithme Go-ICP [10] a été introduit pour régler le problème de non convexité. Il permet une alternance entre une phase ICP et un processus fondé sur une approche *branch and bound* pour réaliser une recherche dans l'espace des transformations. Il est prouvé que cet algorithme converge vers la solution optimale [10]. Cependant, ce dernier peut se révéler très coûteux en temps de calcul à cause des recherches du plus proche voisin répétées, impliquées par ICP et de la recherche dans l'espace des transformations. En pratique, cet algorithme reste difficile à utiliser sur des données réelles de scènes d'intérieur de plusieurs dizaines de milliers de points.

Holz et al. ont créé une chaîne de traitement en intégrant ICP afin d'assurer la convergence de l'algorithme [11]. Premièrement, un alignement approximatif est déterminé, puis, ICP (ou une de ses variantes) est exécuté pour ajuster le résultat. Pour trouver un alignement approximatif, des points clés sont tout d'abord sélectionnés [12], puis, leur descripteur local est calculé grâce à leur voisinage. Les descripteurs locaux les plus utilisés sont FPFH [13], SPIN[14], et SHOT [15]. Dans l'étape suivante, les points de la source et de la cible sont mis en correspondance par similarité de leur descripteur. Enfin, un alignement optimal est déduit. L'erreur principale vient des fausses correspondances attribuées dans ce processus. Différentes techniques ont été implémentées pour gérer ces incohérences. Certaines permettent de rejeter des mauvaises associations avec un *pipeline* de différents critères [11]. Zhou et al. sont allés plus loin avec la méthode FGR (Fast Global Registration) [16] en introduisant une nouvelle fonction de coût qui permet d'attribuer des poids aux correspondances de manière itérative pour leur donner plus ou moins d'importance dans la minimisation suivant leur cohérence. Cette méthode est un recalage global qui ne nécessite pas d'étape d'ajustement. Sa rapidité est remarquable et la majorité des résultats atteint une bonne précision. Néanmoins, toutes les méthodes fondées sur des descripteurs nécessitent que la scène étudiée ait des détails permettant de différencier les points localement. Elles sont moins adaptées à des scènes d'intérieur dans lesquelles beaucoup de points ont le même type de voisinage, notamment sur des plans. Dans certains cas, la convergence peut ne pas être atteinte si les correspondances ont un trop fort taux d'erreur.

Un autre moyen d'obtenir un recalage initial est d'utiliser RANSAC (RANdom SAmple Consensus). La méthode a tout d'abord été introduite dans [17]. Le principe est d'attribuer des correspondances aléatoires entre des ensembles de trois points provenant de chacun des nuages. Une transformation est déduite pour chaque correspondance et une évaluation permet de déterminer la meilleure. Super4PCS [18] a été proposé pour accélérer le traitement et le rendre plus robuste en utilisant des ensembles de quatre points et en leur appliquant différents filtres. La méthode a une convergence optimale. Cependant, le temps de calcul peut se révéler très important lorsque des données de plusieurs milliers de points sont traitées car le nombre de correspondances à tester pour obtenir une précision suffisante devient très élevé. Cette méthode ne peut donc pas être utilisée en pratique dans le contexte de notre travail.

Par ailleurs, les plans présents dans un environnement intérieur ont déjà été utilisés pour réaliser un recalage dans [19]. Cette méthode est inspirée de la NDT (Normal Distribution Transform) [20] qui est utilisée pour mettre en correspondance des densités locales de points entre les nuages. Des correspondances sont attribuées entre les plans extraits de la cible et ceux extraits de la source sur des critères géométriques. Puis la probabilité de localisation des points sur des plans après transformation est maximisée. Cette méthode établit des correspondances entre plans et non entre points ce qui se révèle plus efficace. Cependant, elle repose sur l'algorithme qui extrait les plans des nuages de points qui peut s'avérer très long dans un environnement complexe.

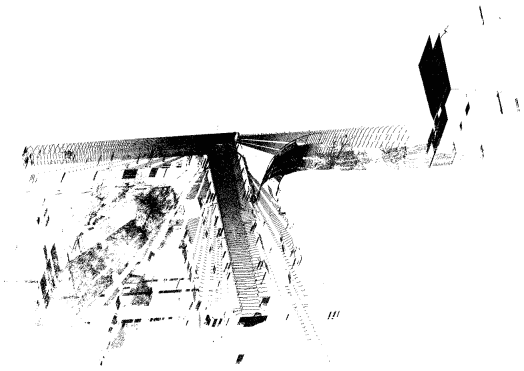
Au vue des difficultés rencontrées pour assurer un bon fonctionnement des méthodes citées précédemment, nous souhaitons introduire une nouvelle méthode inspirée de considérations géométriques, au processus non itératif et suffisament rapide pour pouvoir envisager une implémentation temps réel. La méthode proposée, dénommée Structured Scene Features based Registration (SSFR) dans cet article, est décrite dans la prochaine section.

3 Méthode proposée

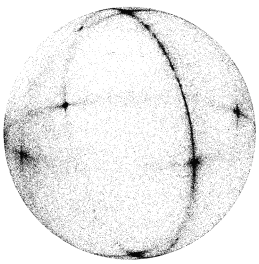
Cette méthode de recalage est adaptée à la reconstruction d'environnements structurés et est basée sur des caractéristiques géométriques de ce type de scène. Les trois murs principaux non parallèles du nuage cible doivent avoir des équivalents dans le nuage source. Cet algorithme permet d'obtenir successivement la rotation et la translation de la transfomation recherchée en deux processus distincts.

3.1 Recherche de la rotation

La première étape consiste à projeter les deux nuages de points sur la sphère Gaussienne i.e. , on représente les normales en chaque point comme un point sur la sphère de rayon unité. Un exemple de projection est donné Figure 1. Dans cette représentation, l'orientation des plans est mise en valeur par des régions plus denses de points. La propriété principale de l'image Gaussienne est que la rotation a le même effet sur la projection que sur le nuage de points initial [5]. On va donc chercher à faire correspondre l'image gaussienne de la source et l'image gaussienne de la cible pour déduire la rotation entre la source et la cible. Pour détecter les clusters de points dans l'image gaussienne, on applique un filtre par densité. On sélectionne les six zones de plus forte densité avec leur voisinage. Puis, l'algorithme du *mean shift* [21] est utilisé avec un noyau d'*Epanechnikov* pour détecter les modes des clusters extraits précédemment. Les modes obtenus sont les normales des plans principaux du nuage et sont nommés $\vec{n_i}$ avec $i = 1, ..., N$ avec N le nombre de normales.



(a) Nuage de points à projeter



(b) Projection

Figure 1 – *Exemple de projection sur la sphère gaussienne d'un scan d'intérieur, DS2 scan 1 (cf section 4)*

Un exemple de résultat de ce traitement est donné Figure 2.

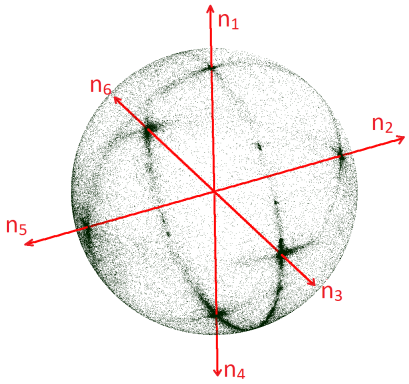


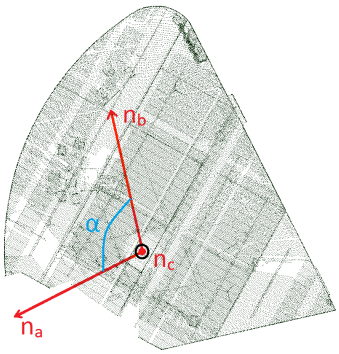
Figure 2 – *Normales des plans extraites de l'image gaussienne.*

Une fois que les normales des murs principaux sont obtenues, toutes les paires de normales $(\vec{n_i}, \vec{n_j})_C$, $(\vec{n_k}, \vec{n_l})_S$ avec $i \neq j$ et $k \neq l$ sont formées respectivement dans la cible (indice C) et dans la source (indice S) . Le but est alors de lier une paire extraite de la source à la paire correspondante extraite de la cible. Toutes les correspondances sont d'abord établies puis filtrées. Pour ce faire, on ne garde que les correspondances pour lesquelles les angles $(\vec{n_i}, \widehat{\vec{n_j}})_C$ et $(\vec{n_k}, \widehat{\vec{n_l}})_S$ sont similaires. Pour trouver la meilleure combinaison, toutes les correspondances restantes sont testées et, après calcul de la translation, les résultats des différents recalages sont comparés comme expliqué dans la section 3.3.

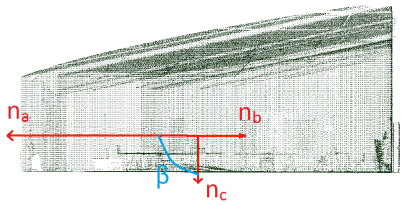
3.2 Recherche de la translation

Premièrement, trois normales $\vec{n_a}$, $\vec{n_b}$, $\vec{n_c}$, sont sélectionnées parmi les modes calculés à l'étape précédente (décrite section 3.1) et définissent des axes de translation a, b, c. Les axes a et b correspondent aux normales $\vec{n_a}$ et $\vec{n_b}$ utilisées pour réaliser la rotation. L'axe c est déduit après avoir aligné toutes les normales principales par la rotation. Il correspond à la normale de la cible la mieux alignée avec une de la source. Un exemple de sélection d'axes est donné Figure 3. Les nuages de points sont alors projetés sur chacun des axes et des histogrammes sont construits. Le maximum de la corrélation entre les histogrammes d'un même axe correspond alors au déplacement à réaliser sur l'axe étudié pour aligner les murs.

Si les murs sont perpendiculaires, les déplacements obtenus peuvent être appliqués directement sur chacun des axes a, b, c. Cependant, si les murs ne sont pas perpendiculaires, les déplacements sont corrélés les uns aux autres. On peut alors définir un nouveau repère d'axes orthogonaux menant à 3 translations indépendantes. Ces axes sont nommés x, y, z, et sont représentés par les vecteurs $\vec{n_x}$, $\vec{n_y}$, $\vec{n_z}$. Il sont définis comme suit : $\vec{n_x}$ et $\vec{n_a}$ sont superposés ; $\vec{n_z}$ est le produit vectoriel de $\vec{n_a}$ et $\vec{n_b}$; $\vec{n_y}$ est le produit vectoriel de $\vec{n_z}$ et $\vec{n_a}$.



(a) Vue du dessous



(b) Vue de côté

Figure 3 – *Axes sélectionnés par la méthode (en rouge) dans le scan d'une pièce.*

Les déplacements sur les axes a, b et c sont nommés Δa, Δb et Δc. Les déplacements à calculer sur les axes x, y et z sont nommés Δx, Δy and Δz. Une illustration 2D d'une configuration avec deux murs est donnée Figure 4. Sur cette figure simplifiée, on ne considère aucun déplacement selon l'axe z. Les axes de translation initiaux sont donc a et b. Les équations (1) et (2) permettent de calculer Δx et Δy avec α l'angle non signé entre les axes x et b. α appartient à l'intervalle $[0, \pi]$.

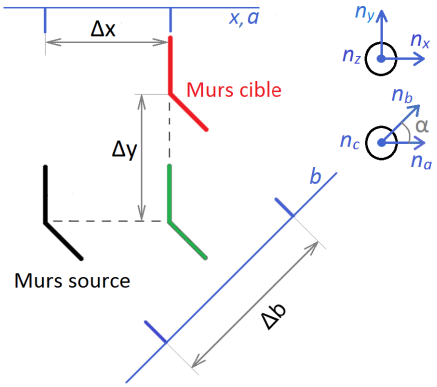


Figure 4 – *Représentation schématique vue du dessus de la translation de deux murs depuis leur position dans la source (en noir) jusqu'à leur position dans la cible (en rouge). En vert, les murs sont représentés après translation sur l'axe x.*

L'équivalence pour le calcul du déplacement selon l'axe z est donnée par l'équation (3) avec β l'angle non signé entre les axes x et c et γ l'angle non signé entre y et c. β et γ appartiennent à l'intervalle $[0, \pi]$.

$$\Delta x = \Delta a \tag{1}$$

$$\Delta y = \frac{\Delta b - \Delta x * cos(\alpha)}{sin(\alpha)} \tag{2}$$

$$\Delta z = \frac{\frac{\Delta c - \Delta x * cos(\beta)}{sin(\beta)} - \Delta y * cos(\gamma)}{sin(\gamma)} \tag{3}$$

3.3 Sélection de la transformation

Comme décrit section 3.1, toutes les possibilités de rotation pour aligner les paires de normales sont testées. Les translations correspondantes sont calculées puis les transformations sont comparées pour sélectionner la meilleure. Pour cela, nous utilisons la valeur LCP (Largest Common Pointset) [18] qui quantifie le chevauchement des nuages de points. Cette valeur correspond au pourcentage de points de la source ayant un voisin dans la cible. Le rayon du voisinage est défini comme la résolution du nuage de points. La résolution est ici définie comme la distance moyenne entre les points d'un nuage.

4 Validation

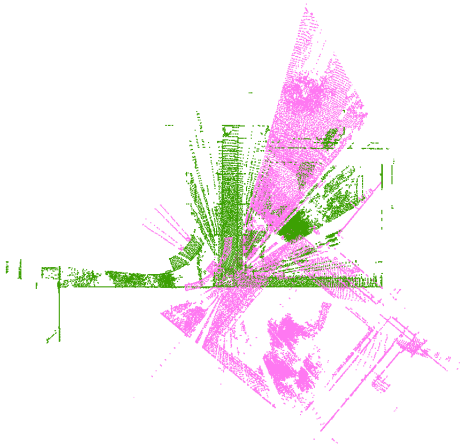
Pour évaluer notre méthode, nous utilisons deux ensembles de données de scans. Le premier est appelé "apartment" et est disponible en ligne sur le site de l'ASL (Autonomous system Lab) [22]. Il est référencée ici comme "DS1". Nous avons constitué le deuxième ensemble, référencé "DS2", à l'Institut Pascal à Clermont-Ferrand afin de pouvoir travailler sur un environnement intérieur complexe avec une vérité terrain. La complexité de ce jeu de données provient 1) de ses motifs tels que des ouvertures, des arbres ou des murs courbes, 2) du nombre de points traités, 3) de sa structure car les scans ont été acquis sur deux étages avec de grands déplacements entre chaque acquisition. Les principales caractéristiques des deux ensembles de données sont présentées dans le tableau 1.

Tableau 1 – *Caractéristiques des jeux de données avec le nombre moyen de points par scan, le nombre de scans, la distance maximale entre les extrémités des nuages de points, la résolution moyenne des nuages, la qualification des déplacements entre chaque acquisition de scan.*

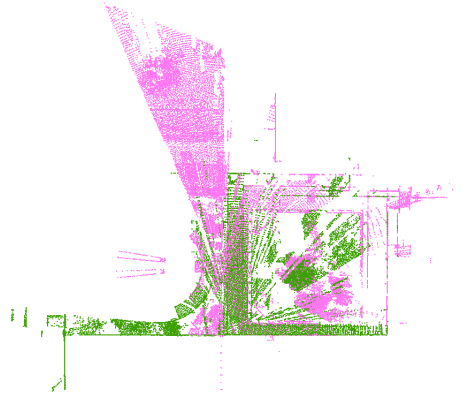
	DS1	DS2
Capteur	Hokuyo UTM-30LX	Leica P20
Nombre de points	365 000	9×10^6
Nombre de scans	45	6
Taille max (m)	11	70
Résolution (cm)	0.61	0.35
Déplacements	Faible	Fort

Des transformations issues d'une vérité terrain sont disponibles pour chacune d'elles. Dans DS2, elles proviennent

d'un recalage réalisé avec des cibles physiques. La figure 5 présente un exemple de recalage avec notre méthode pour des nuages de points extraits de DS2.



(a) Avant recalage



(b) Après recalage

Figure 5 – *Exemple de résultat obtenu : recalage du scan 2 (rose) sur le scan 1 (vert) dans DS2*

Pour valider la méthode, nous avons sélectionné trois critères : 1) l'erreur RMS (*Root Mean Square*) entre source et cible. Cette erreur est calculée à partir de la vérité terrain et dépend de la résolution des nuages de points étudiés. Les correspondances de référence entre source et cible sont définies grâce à la vérité terrain. Des paires sont formées entre les points de la source (après transformation par la référence) et les points de la cible si ces points sont assez proches. Puis la distance entre les points d'une même paire est évaluée après recalage par chacune des méthodes. 2) le taux de réussite de recalage sur un ensemble de données. On suppose que le recalage est réussi quand l'erreur RMS est inférieure à 10 cm. 3) le temps de calcul de l'algorithme (processeur 8 cœurs Intel Xeon-E5620, 2.4GHz).

Notre algorithme SSFR est comparé aux méthodes existantes sur les deux ensembles de données. Les résultats sont disponibles dans le tableau 2.

Tableau 2 – *Evaluation des méthodes sur DS1 et DS2. ICP1 : ICP point-to-point, ICP2 : ICP point-to-plane, NDT : Normal Distribution Transform, FGR : Fast Global Registration, Leica : méthode semi-manuelle fondée sur ICP, SSFR : notre méthode, VT : Vérité Terrain. RMSE : Root Mean Square Error ; S : taux de réussite ; T : time ; nd : pas de données ; la croix indique que le taux de réussite n'est pas suffisant.*

	DS1			DS2		
	RMSE (cm)	S (%)	T (s)	RMSE (cm)	S (%)	T (s)
ICP1	2.0	54	7.4	x	0	x
ICP2	1.7	82	3.2	x	0	x
NDT	2.4	61	5.2	x	0	x
FGR	2.3	100	3.9	x	0	x
Leica	nd	nd	nd	2.7	100	nd
SSFR	1.8	100	2.4	2.9	100	24.0
VT	1.4	100	nd	2.66	100	nd

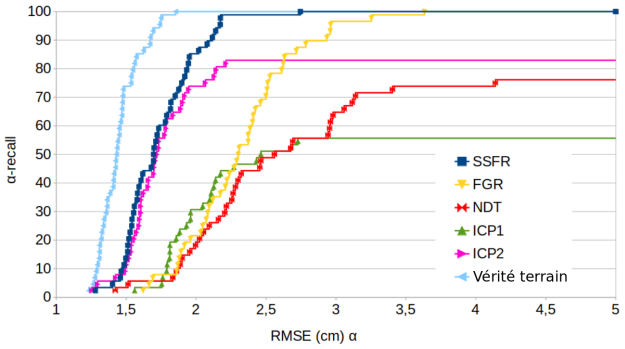


Figure 6 – *Evaluation sur DS1. α-recall est le pourcentage de scans pour lesquels une méthode donnée mène à un recalage avec une RMSE $< \alpha$. SSFR est notre méthode. Comparaison avec algorithmes ICP point-to-point (ICP1), ICP point-to-plane (ICP2), Fast Global Registration (FGR) et Normal Distribution Transform (NDT).*

Les nuages de points ont été sous-échantillonnés de manière uniforme afin d'obtenir une résolution de, respectivement, 3.6 cm et 6.1 cm, (ce qui limite le nombre de points à quelques dizaines de milliers) et de comparer avec différentes méthodes existantes : ICP point-to-point, ICP point-to-plane, NDT et FGR. Les méthodes Go-ICP et Super4PCS n'ont pas été inclues dans l'évaluation car leur temps de calcul est trop long pour des données aussi volumineuses que des scans de bâtiments. Les algorithmes ICP et NDT ont été testés en utilisant la librairie PCL et FGR, Go-ICP et Super4PCS ont été évalués avec le code disponible en ligne. Chaque scan a été recalé sur son prédecesseur dans l'ordre d'acquisition. Notre méthode est la seule qui fonctionne sur les deux ensembles de données. Les méthodes fondées sur ICP ont un taux de réussite faible dès lors que les déplacements sont trop grands. On remarque

que sur DS1, l'algorithme proposé dans cet article atteint une précision meilleure que les autres algorithmes testés (1.8 cm) en un temps plus court (2.4 s). La figure 6 détaille les résultats de tous les recalages avec les différents algorithmes étudiés sur DS1 et confirme les bonnes performances de notre algorithme SSFR.

5 Conclusions et perspectives

Nous avons présenté un nouvel algorithme de recalage global de scènes d'intérieur avec chevauchement partiel. Il fonctionne sans recalage initial et peut être utilisé sur des scènes géométriques avec peu de détails car il s'affranchit d'une recherche de descripteurs locaux. Son fonctionnement non itératif permet d'obtenir un résultat fiable pour une durée bornée. Il présente une bonne précision pour des environnements complexes et sa rapidité est compatible avec une implémentation en temps réel. A l'avenir, nous souhaitons enrichir et mettre à disposition le jeu de données réalisé pour cet article. La méthode va être évaluée sur de nouveaux ensembles de données pour assurer sa bonne adaptabilité aux différents contextes. De plus, un travail de modélisation de la scène sera réalisé en utilisant l'extraction des plans proposée dans ce travail afin de mettre en évidence des primitives de plus haut niveau.

Références

[1] Sebastian Ochmann, Richard Vock, Raoul Wessel, et Reinhard Klein. Automatic reconstruction of parametric building models from indoor point clouds. *Computers and Graphics (Pergamon)*, 54 :94–103, 2016.

[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, et J. J. Leonard. Past, present, and future of simultaneous localization and mapping : Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6) :1309–1332, 2016.

[3] Paul Besl et Neil McKay. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2) :239–256, 1992.

[4] Philippe Brou. Using the Gaussian Image to Find the Orientation of Objects. *The International Journal of Robotics Research*, 3(4) :89–125, 1984.

[5] Sang Bum Kang et Katsushi Ikeuchi. Determining 3-D object pose using the complex extended Gaussian image. Dans *Computer Vision and Pattern Recognition (CVPR)*, pages 580–585, 1991.

[6] François Pomerleau, Francis Colas, et Roland Siegwart. A Review of Point Cloud Registration Algorithms for Mobile Robotics. *Foundations and Trends in Robotics*, 4(1) :1–104, 2015.

[7] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2) :119–152, 1994.

[8] A Segal, D Haehnel, et S Thrun. Generalized-ICP. *Robotics : Science and Systems*, 5 :168–176, 2009.

[9] Andrew W Fitzgibbon. Robust registration of 2d and 3d point sets. 21 :1145–1153, 04 2002.

[10] Jiaolong Yang, Hongdong Li, et Yunde Jia. Go-ICP : Solving 3D registration efficiently and globally optimally. Dans *International Conference on Computer Vision (ICCV)*, pages 1457–1464, 2013.

[11] D. Holz, A. E. Ichim, F. Tombari, R. B. Rusu, et S. Behnke. Registration with the point cloud library : A modular framework for aligning in 3-d. *IEEE Robotics Automation Magazine*, 22(4) :110–124, 12 2015.

[12] Silvio Filipe et Luís A. Alexandre. A comparative evaluation of 3d keypoint detectors in a rgb-d object dataset. Dans *International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 476–483, 1 2014.

[13] Radu Bogdan Rusu, Nico Blodow, et Michael Beetz. Fast Point Feature Histograms (FPFH) for 3D registration. *IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009.

[14] Andrew E. Johnson et Martial Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21(5) :433–449, 1999.

[15] Federico Tombari, Samuele Salti, et Luigi Di Stefano. Unique signatures of histograms for local surface description. Dans *European Conference on Computer Vision (ECCV)*, pages 356–369, 2010.

[16] Qian-Yi Zhou, Jaesik Park, et Vladlen Koltun. *Fast Global Registration*, pages 766–782. 2016.

[17] Martin A Fischler et Robert C Bolles. Paradigm for Model. *Communications of the ACM*, 24(6) :381–395, 1981.

[18] Nicolas Mellado, Dror Aiger, et Niloy J. Mitra. Super 4PCS fast global pointcloud registration via smart indexing. *Computer Graphics Forum*, 33(5) :205–215, 2014.

[19] Kaustubh Pathak et Andreas Birk. Fast Registration Based on Noisy Planes with Unknown Correspondences for 3D Mapping. *IEEE Transactions on Robotics*, 26(3) :424–441, 2010.

[20] Martin Magnusson. *The Three-Dimensional Normal-Distributions Transform — an Efficient Representation for Registration, Surface Analysis, and Loop Detection*. Thèse de doctorat, Orebro University, 2009.

[21] Cheng Yizong. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 17(8) :790–799, 8 1995.

[22] François Pomerleau, M. Liu, Francis Colas, et Roland Siegwart. Challenging data sets for point cloud registration algorithms. *The International Journal of Robotics Research*, 31(14) :1705–1711, 12 2012.

Peut-on estimer un modèle d'illumination locale à partir d'une tache spéculaire ?

Souheil Hadj Said[1,2] Mohamed Tamaazousti[1] Adrien Bartoli[2]

[1] CEA, LIST, Point Courrier 173, F-91191 Gif-sur-Yvette, France
[2] Institut Pascal - UMR 6602 - CNRS/UCA/SIGMA, Clermont-Ferrand, France

souheil.hadjsaid@cea.fr ; mohamed.tamaazousti@cea.fr ; adrien.bartoli@gmail.com

Résumé

Dans les applications de réalité augmentée (RA) et diminuée (RD), la représentation de l'illumination dans la scène est cruciale. Cette tâche peut être assurée par les modèles d'illumination locale à condition que l'on connaisse tous les paramètres de la scène. Les dernières avancées en vision par ordinateur permettent d'estimer de manière précise et robuste la pose de la caméra et la géométrie de la scène. Pourtant, même en connaissant ces dernières, il reste difficile d'estimer les autres paramètres nécessaires aux modèles d'illumination locale : la position et l'intensité de la source lumineuse et la réflectance et la rugosité de la surface. Certains travaux proposent des méthodes pour estimer ces paramètres à partir d'une ou plusieurs images. Cependant, les limites de ces méthodes sont encore à définir afin de pouvoir étudier leur compatibilité potentielle avec des applications telles que la RA et la RD, qui nécessitent une certaine précision et robustesse. Dans cet article, nous mettons en avant la notion de tache spéculaire, qui est un élément clé de ce problème inverse, et proposons d'étudier sur des données de synthèse la possibilité de retrouver les paramètres manquants à partir d'une seule tache spéculaire. Notre but est d'avancer méthodologiquement pour cerner empiriquement les conditions optimales d'inversion de ces modèles.

Mots clefs

Modèle d'illumination locale, réflectance, rugosité, Réalité Augmentée, Réalité Diminuée, tache spéculaire, source lumineuse, problème inverse.

1 Introduction

Ces dernières années, les progrès de la reconstruction 3D et de la localisation dans le contexte de scènes statiques ont atteint un fort degré de maturité. Cela a permis de développer un large panel d'applications dans les domaines de la réalité augmentée (RA), la réalité diminuée (RD) et des effets spéciaux en général [19]. Dans ce genre d'application, la représentation de l'illumination de la scène est cruciale. Dans l'idéal, l'illumination doit être représentée par les modèles d'illumination globale qui nécessitent la résolution de l'équation de rendu [6]. Jachnik *et al.* [5] proposent une méthode pour résoudre ce problème et montrent l'intérêt de pouvoir synthétiser les reflets de la scène pour améliorer le rendu en RA. Cependant, leur méthode ne fonctionne que pour des surfaces planes et est coûteuse en terme de temps de traitement. De plus, la méthode ne gère pas l'illumination pour des points de vue autres que ceux utilisés pour la reconstruction du modèle. Plus récemment, les travaux de [3, 10] ont montré l'intérêt du rendu des spécularités pour ces applications. Ces contributions proposent des approches géométriques pour la prédiction de spécularité afin d'améliorer le rendu en temps réel. Ils montrent que leurs approches sont bien adaptées pour répondre aux contraintes des applications de RA et de RD. Cependant, ces travaux n'abordent pas le rendu des ombres. De manière intermédiaire, l'illumination de la scène peut être approximée par l'estimation de modèles d'illumination locale. En effet, une telle reconstruction permettrait par rapport aux approches précédemment mentionnés de répondre aux contraintes de temps de traitement tout en permettant une souplesse dans le rendu. Elle permettrait entre autre de synthétiser des ombres et des spécularités à partir de nouveaux points de vue, d'estimer les propriétés des matériaux, de rajouter de nouvelles sources de lumière et de gérer tout type de matériau. En supposant la pose de la caméra et la géométrie de la scène connues, comme dans [3, 5, 10], il reste difficile d'estimer les paramètres des modèles d'illumination locale : la position et l'intensité de la source lumineuse et la réflectance et la rugosité de la surface. Nous appelons *inversion de l'illumination locale* l'estimation de ces paramètres.

Certains travaux ont étudié le problème d'inversion de l'illumination locale à partir d'une ou plusieurs images et de la géométrie de la scène. Cependant, selon leurs résultats, l'unicité et le conditionnement de la solution restent ambigus. En effet, certains travaux [2, 4] proposent de reconstruire ces paramètres à partir d'une seule image alors que d'autres travaux [9, 13, 18] mentionnent la nécessité de le faire en multi-vues. Il est par conséquent nécessaire de savoir délimiter le cadre d'utilisation de ces méthodes pour mieux évaluer leurs compatibilités potentielles avec les contraintes de précision et de robustesse des applications de RA et de RD. Nous observons dans notre étude

que les spécularités sont très informatives quant à l'interaction lumière-matière. Ainsi, afin de mieux quantifier les données nécessaires à l'inversion de l'illumination locale nous proposons d'introduire les trois notions suivantes :

Définition 1 *tache spéculaire. Une tache spéculaire est une zone connexe de l'image où l'intensité observée est majoritairement issue de la réflexion directe d'une source de lumière. Elle est caractérisée par un seul maxima d'intensité.*

Définition 2 *Mono-tache. Une approche pour l'inversion de l'illumination locale est mono-tache lorsqu'elle utilise une seule tache spéculaire.*

Définition 3 *Multi-tache. Une approche pour l'inversion de l'illumination locale est multi-taches lorsqu'elle utilise plusieurs taches spéculaires, sur une ou plusieurs images.*

Dans le but de mieux cerner les conditions optimales pour l'inversion de ces modèles nous proposons dans un premier temps de répondre à la question *Peut-on estimer les paramètres associés à la source lumineuse et aux propriétés de la surface avec une approche mono-tache ?* Cela est motivé par le fait que d'un point de vue méthodologique, il n'est pas pertinent de catégoriser ces méthodes d'estimation comme étant mono-vue ou multi-vues. D'après notre étude, ce problème inverse dépend du nombre des taches spéculaires dans la scène. Pourtant, même sur de multiples images il pourrait y avoir aucune tache spéculaire. En effet, une récente étude de Morgand *et al.* [11] a montré que l'on peut associer à chaque spécularité une caméra virtuelle. Par conséquent, chaque tache spéculaire peut être associée à un point de vue virtuel distinct de ceux des autres taches spéculaires, même si ces taches sont issues d'une même image. De manière complémentaire nous nous intéressons dans cette étude à savoir s'il est nécessaire de passer par une séparation des composantes diffuses et spéculaires pour inverser l'illumination locale. Nous comparons donc deux approches. La première consiste à considérer la zone de l'image associée à une tache spéculaire et à estimer tous les paramètres du modèle. La deuxième approche sépare les deux composantes diffuse et spéculaire, estime les paramètres liées à la composante spéculaire en premier lieu puis utilise les valeurs estimés comme une initialisation et déduit tous les paramètres (des composantes spéculaire et diffuse). Trois modèles d'illumination différents sont utilisés pour évaluer chaque approche.

2 Etat de l'art

La plupart des méthodes de la littérature utilisent plusieurs images pour inverser l'illumination locale. Celles-ci peuvent alors être classées comme des méthodes multi-taches. Mercier *et al.* [9] proposent un environnement complet pour reconstruire un objet à partir d'un ensemble d'images. Cette approche utilise une version modifiée du modèle de Phong pour retrouver les paramètres de la source lumineuse et la réflectance de la surface. Xu *et al.* [18] utilisent deux images acquises par une caméra stéréo. La différence d'intensité entre les deux images est reproduite par le modèle de Blinn-Phong [1] pour estimer la position de la lumière et la réflectance de l'objet. D'autres méthodes proposent de résoudre le problème en utilisant une seule image. Par exemple, Hara *et al.* [4] ont utilisé un polariseur pour séparer les composantes diffuse et spéculaire des images. Ils estiment ensuite à partir de la composante spéculaire la position de la source de lumière en utilisant une version simplifiée du modèle d'illumination locale Torrance-Sparrow [16]. Boivin *et al.* [2] ont développé un algorithme qui permet d'estimer les réflectances de plusieurs matériaux dans la scène à partir d'une seule image. Ils ont estimé le modèle d'illumination de Ward [17]. Bien que ces deux dernières méthodes utilisent une seule image, nous ne pouvons pas les classer comme des approches mono-taches car il peut y avoir plusieurs spécularités dans une image.

3 Les modèles d'illumination

3.1 Modèle de base

Les modèles d'illumination sont utilisés en synthèse d'image pour simuler l'illumination dans les scènes 3D. En général, un modèle d'illumination locale est la somme de trois composantes : ambiante, diffuse et spéculaire. Phong [14] exprime un modèle d'illumination locale, pour chaque point 3D $\mathbf{P}$ de la surface (voir figure 1), sous la forme suivante :

$$I(\mathbf{P}) = k_a(\mathbf{P})i_a + \sum_{j=1}^{r} \mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_n(\mathbf{P})k_d(\mathbf{P})i_j + \sum_{j=1}^{r} \phi_s(\mathbf{P}, k_s(\mathbf{P}), m, i_j). \quad (1)$$

$k_a(\mathbf{P})i_a$ représente la composante ambiante qui est une approximation de l'éclairage indirecte (les inter-réflexions) en chaque point $\mathbf{P}$. $\mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_j(\mathbf{P})k_d(\mathbf{P})i_j$ et $\phi_s(\mathbf{P}, k_s(\mathbf{P}), m, i_j)$ représentent, respectivement, la contribution d'une source lumineuse S_j à la composante diffuse et à la composante spéculaire, j étant l'indice de la source, i_j l'intensité de la source et r le nombre total de sources lumineuses dans la scène. Nous comparons trois modèles utilisés dans la littérature dans le contexte de l'inversion de l'illumination locale : Le modèle de Blinn-Phong [1] auquel on réfère par BP, une version simplifiée du modèle Torrance-Sparrow [16, 12] auquel on réfère par TS et le modèle de Ward [17] pour les surfaces isotropes auquel on réfère par WI. La différence entre ces modèles réside uniquement dans la composante spéculaire, qui est

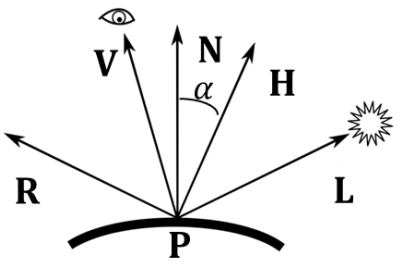


Figure 1 – *Représentation pour un point 3D* $\mathbf{P}$ *des vecteurs définissant la composante spéculaire pour les modèles d'illumination locale.*

exprimée ainsi :

$$\begin{aligned}
&\text{BP} \Rightarrow \phi_s(\mathbf{P}) = k_s(\mathbf{P}) i_j \left(\mathbf{H}(\mathbf{P}) \cdot \mathbf{N}(\mathbf{P})\right)^m. \\
&\text{TS} \Rightarrow \phi_s(\mathbf{P}) = k_s(\mathbf{P}) i_j \frac{1}{\mathbf{N}(\mathbf{P}) \cdot \mathbf{V}(\mathbf{P})} \exp\left(-\frac{\alpha^2}{2m^2}\right). \\
&\text{WI} \Rightarrow \phi_s(\mathbf{P}) = \frac{k_s(\mathbf{P}) i}{4m^2 \sqrt{(\mathbf{N}(\mathbf{P}) \cdot \mathbf{V}(\mathbf{P}))(\mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_n(\mathbf{P}))}} \\
&\qquad \exp\left(-\frac{\tan(\alpha)^2}{2m^2}\right).
\end{aligned} \tag{2}$$

$\mathbf{V}(\mathbf{P})$ désigne la direction de l'angle de vue au point $\mathbf{P}$. $\mathbf{N}(\mathbf{P})$ est la normale au point $\mathbf{P}$. $\mathbf{L}_n(\mathbf{P})$ est la direction de la source lumineuse $\mathbf{S}_n$ au point $\mathbf{P}$. $\mathbf{H}(\mathbf{P}) = \frac{\mathbf{L}_n(\mathbf{P}) + \mathbf{V}(\mathbf{P})}{\|\mathbf{L}_n(\mathbf{P}) + \mathbf{V}(\mathbf{P})\|}$ représente le vecteur mi-chemin entre $\mathbf{L}_n(\mathbf{P})$ et $\mathbf{V}(\mathbf{P})$ (aussi appelé Halfway vector). α est l'angle entre $\mathbf{H}(\mathbf{P})$ et $\mathbf{N}(\mathbf{P})$ (voir figure 1). i_n est l'intensité de la source lumineuse S_n.

3.2 Hypothèses

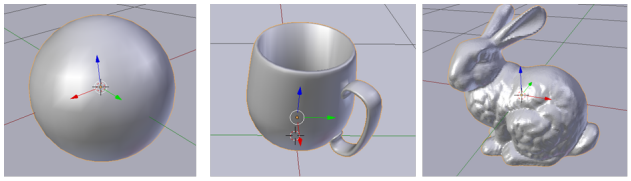
Afin de minimiser la complexité du problème, nous supposons que :

- Les reflets spéculaires observés sont causés exclusivement par la source lumineuse S_1.
- L'objet a une albédo constante.

3.3 Modèles simplifiés

La première hypothèse permet de réduire la somme sur la composante spéculaire à un seul terme lié à une seule source lumineuse. Par contre, la somme sur la composante diffuse reste la même. La deuxième hypothèse implique que les réflectances sont constantes (k_a, k_d et k_s sont indépendants de $\mathbf{P}$). Ainsi, le modèle simplifié s'écrit :

$$\begin{aligned}
I(\mathbf{P}) =& k_a i_a + \sum_{n=2}^{N} \mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_n(\mathbf{P}) k_d i_n \\
&+ k_d \mathbf{L}_1(\mathbf{P}) \cdot \mathbf{N}(\mathbf{P}) i_1 + \phi_s(\mathbf{P}, k_s, m, i_1).
\end{aligned} \tag{3}$$



(a) Objet convexe et lisse (a) Objet non convexe et lisse (a) Objet non convexe et rugueux

Figure 2 – *Les objets 3D utilisés dans la base d'images.*

En posant $C = k_a i_a + \sum_{n=2}^{N} \mathbf{N}(\mathbf{P}) \cdot \mathbf{L}_n(\mathbf{P}) k_d i_n$, la formule devient :

$$I(\mathbf{P}) = C + \mathbf{L}_1(\mathbf{P}) \cdot \mathbf{N}(\mathbf{P}) k_d i_1 + \phi_s(\mathbf{P}, k_s, m, i_1). \tag{4}$$

Notons que la constante C combine la composante ambiante et une partie de la composante diffuse. En posant cette formulation, nous nous proposons d'estimer les paramètres des composantes diffuse et spéculaire liées à la source lumineuse S_1. La composante spéculaire liée aux autres sources est supposée négligeable (hypothèse 1) et la composante diffuse correspondante est constante. Pour simplifier les notations, nous posons $K_d = k_d i_1$ et $K_s = k_s i_1$. Ainsi, nous nous proposons d'estimer les paramètres :

- $K_d \in [0,1]^3$: coefficient de la réflectance diffuse.
- $K_s \in [0,1]^3$: coefficient de la réflectance spéculaire.
- $\mathbf{S}_1 \in \mathbb{R}^3$: position de la source lumineuse S_1.
- $m \in \mathbb{R}_+$: rugosité de la surface de l'objet.
- $C \in [0,1]^3$: constante.

Notons que les paramètres K_d, K_s et C sont en trois dimensions car on les estime pour les canaux couleur RVB.

4 Approches

Nous définissons $\mathbf{\Omega}$ comme l'ensemble des points de la surface représentant une tache spéculaire. $I(\mathbf{P})$ représente l'intensité du pixel $\mathbf{p}$ projeté du point $\mathbf{P}$ dans l'image.

4.1 Approche 1 : sans séparation

Nous définissons la distance photométrique entre deux pixels représentants le même point 3D $\mathbf{P}$ par : $\|I_r - I_e\| = \sqrt{(I_r^r - I_e^r)^2 + (I_r^v - I_e^v)^2 + (I_r^b - I_e^b)^2}$, avec I_e l'intensité de l'image prédite par le modèle d'illumination au point $\mathbf{P}$ avec les paramètres à estimer et I_r l'intensité observée. L'intensité d'un pixel est exprimé dans l'espace RVB. Pour retrouver les différents paramètres en une passe, nous utilisons une minimisation au sens des moindres carrés sur la fonction de coût photométrique suivante :

$$c_{\text{photo}} = \sqrt{\frac{1}{3|\Omega|} \sum_{\mathbf{P} \in \Omega} \|I_r(\mathbf{P}) - I_e(\mathbf{P}, \mathbf{S}_0^*, K_d^*, K_s^*, C^*, m^*)\|^2}, \tag{5}$$

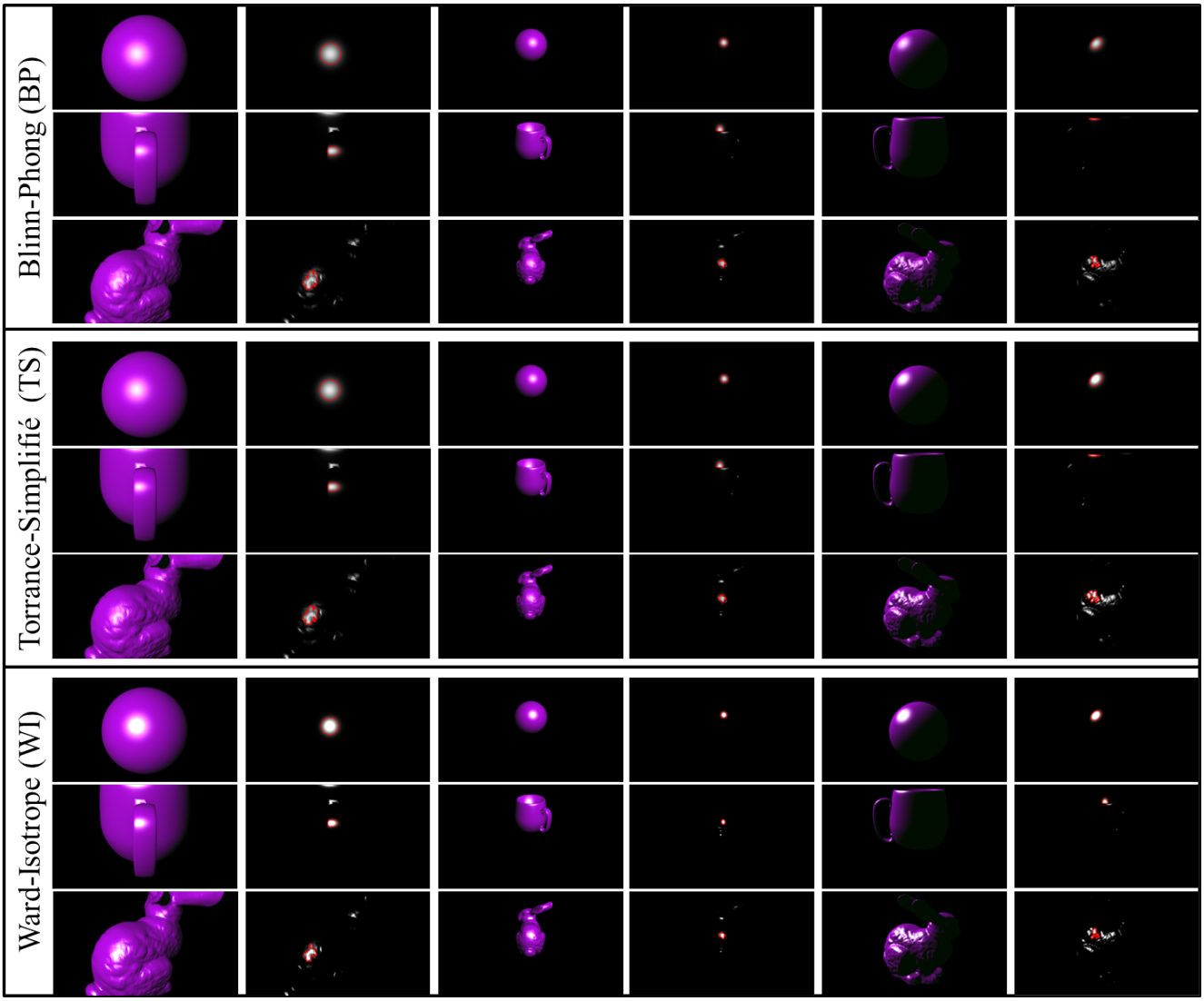


Figure 3 – *Un échantillon de la base d'images utilisés dans nos tests. Ces images générées par les 3 modèles d'illumination locales sont utilisés commen données d'entrée dans notre algorithme. Nous considérons au total 81 images (3 modèles d'illumination* $\times$ *3 angles de vues* $\times$ *3 distances à la caméra* $\times$ *3 objets 3D). A la droite de chaque image, est représenté l'image spéculaire correspondante. Les taches spéculaires utilisées pour l'estimation sont délimitées en rouge sur les images de la composante spéculaire. Un seuil d'intensité égal à 0.3 est utilisé pour délimiter cette zone (l'intensité de l'image est comprise entre 0 et 1). Seule la zone la plus grande dans l'image est considérée.*

Dans nos expériences, nous utilisons la méthode d'optimisation de Levenberg-Marquardt. Bien que cette optimisation nécessite une méthode d'initialisation, notre intérêt dans cette étude réside dans l'évaluation du bassin de convergence autour des vraies valeurs. Ainsi, pour chaque image, nous lançons 10 tests, chacun avec une initialisation différente. Nous augmentons l'amplitude de l'erreur d'initialisation par rapport à la vérité terrain avec le même pas entre deux tests.

4.2 Approche 2 : avec séparation

Notre deuxième approche estime les paramètres du modèle en deux temps, ce qui réduit la complexité du problème.

Etape 1 : séparation des composantes diffuse et spéculaire. La séparation des composantes diffuse et spéculaire est une étape de pré-traitement clé dans cette approche. Plusieurs méthodes [7, 8, 15] proposent de générer à partir d'une image deux images une diffuse et une spéculaire, mais leur précision n'a pas été quantifiée. Dans ce premier stade de l'étude, on génère directement l'image spéculaire par le modèle d'illumination considéré. Ceci nous permet de négliger le bruit introduit par cette première étape. D'après les modèles d'illumination locale, l'image spéculaire est définit en chaque pixel comme : mbox$I_s(\mathbf{P}) = \phi_s(\mathbf{P}, k_s, m, i_1)$. Les images spéculaires synthétisées sont représentées à la figure 3.

L'étape 2 : ajustement sur l'image spéculaire. Sur l'image spéculaire, la deuxième étape consiste à minimiser au sens des moindres carrés le coût suivant :

$$c_{\text{specu}} = \sqrt{\frac{1}{3|\Omega|}\sum_{\mathbf{P}\in\Omega}\|I_{sr}(\mathbf{P}) - I_{se}(\mathbf{P}, \mathbf{S}_0^*, K_s^*, m^*)\|^2}, \tag{6}$$

pour estimer les paramètres $\mathbf{S}_0, K_s$ et m. I_{sr} et I_{se} sont, respectivement, l'intensité de la composante spéculaire observée et prédite.

Etape 3 : ajustement sans séparation. Dans la troisième étape, nous utilisons les paramètres estimés à l'étape 2 comme des valeurs d'initialisation et nous raffinons le coût exprimé dans (5) pour retrouver la totalité des paramètres recherchés. En d'autre termes, l'approche 2 peut être vue comme une initialisation de l'approche 1.

5 Expériences

5.1 Base d'images

Dans le premier stade de cette étude, nous considérons une base d'images de synthèse générées par les modèles d'illumination locale considérés. Au total, nous réalisons nos expérience sur 81 images. Nous considérons 3 types d'objets illustrés par la figure 2. Pour chaque objet, 9 poses de caméra sont utilisées. Ainsi, nous utilisons 27 images par modèle dans nos expériences. Un échantillon de ces images est montré en figure 3. Comme nous testons une approche mono-tache, nous sélectionnons pour chaque image une tache spéculaire. On segmente cette zone, désignée par Ω, en fixant un seuil d'intensité $I_{\text{seuil}} = 0.3$, sachant qu'un canal de couleur est saturé pour une intensité égale à 1. Seule la zone connexe la plus grande est retenue comme étant une tache spéculaire.

5.2 L'erreur d'estimation

L'erreur d'estimation est fixée comme la différence normalisée entre les valeurs vérité terrain des différents paramètres et les valeurs issues de l'algorithme d'estimation :

$$\begin{aligned} e_{\text{g}} = &\frac{1}{T_{\mathbf{S}}}\|\mathbf{S}_0 - \mathbf{S}_0{}^*\| + \frac{1}{T_K}\|K_d - K_d{}^*\| + \frac{1}{T_K}\|C - C^*\| \\ &+ \frac{1}{T_K}\|K_s - K_s{}^*\| + \frac{1}{T_m}\|m - m^*\| \end{aligned} \tag{7}$$

où $T_{\mathbf{S}}$, T_K et T_m sont des poids permettant de normaliser l'erreur car les paramètres ont des ordres de grandeur différents. Leurs valeurs sont fixées par un test préliminaire expliqué en section 5.1.

5.3 L'erreur résiduelle

Le résidu issu de l'algorithme d'optimisation est de nature photométrique. Nous utilisons cette erreur résiduelle pour évaluer la capacité de la méthode à recréer l'image observée. Cette erreur est exprimée par l'équation (5).

5.4 Détermination des poids

Pour déterminer les poids $T_{\mathbf{S}}$, T_K et T_m, nous lançons l'algorithme avec l'approche 1 sur les images de la base en initialisant avec les paramètres de la vérité terrain. Le poids $T_{\mathbf{S}}$ est ensuite déterminé par :

$$T_{\mathbf{S}} = \mu_{\mathbf{S}} + \alpha_{\mathbf{S}}\sigma_{\mathbf{S}}, \tag{8}$$

où $\mu_{\mathbf{S}} = \frac{1}{j}\sum_{i=1}^{j}\|\mathbf{S}_0(i) - \mathbf{S}_0{}^*(i)\|$ est la moyenne quadratique sur l'erreur de la position de la source lumineuse sur j essais, $\sigma_{\mathbf{S}}$ l'écart-type et $\alpha_{\mathbf{S}}$ un coefficient choisi tel que 95% des essais satisfassent la condition suivante :

$$\mu_{\mathbf{S}} - \alpha_{\mathbf{S}}\sigma_{\mathbf{S}} \leq \|\mathbf{S}_0(i) - \mathbf{S}_0(i)^*\| \leq \mu_{\mathbf{S}} + \alpha_{\mathbf{S}}\sigma_{\mathbf{S}}. \tag{9}$$

Cette condition nous permet d'exclure les valeurs aberrantes. La même méthode est utilisée pour T_K et T_m. Les valeurs issus de ce test dépendent des ordres de grandeur des paramètres estimés. C'est pourquoi on les utilise pour normaliser l'erreur d'estimation décrite dans l'équation 7. Les poids nous renseignent également sur la taille du bassin de convergence.

6 Résultats

6.1 Poids

Dans le tableau 1 nous présentons les poids par modèle d'illumination : Ces poids ont été calculés par le test expliqué dans la section 5.4.

Seuils	BP	TS	WI
$T_{\mathbf{S}}$	0.207	3.620	0.600
T_K	0.154	0.160	0.180
T_m	0.557	0.004	0.004

Tableau 1 – *Les poids par modèle d'illumination.*

6.2 Comparaisons

Pour chaque image de la base de donnée, nous lançons 10 essais avec des initialisations de plus en plus éloignées de la vérité terrain. Dans la figure 4, nous comparons entre les trois modèles associés à à l'approche 1 puis à l'approche 2 avec le même échelle. Dans la figure 5, nous comparons entre les deux approches pour chaque modèle séparément. Sur la colonne à droite pour les deux figures, nous reportons l'évolution des moyennes de l'erreur d'estimation et sur la colonne à gauche, nous reportons l'évolution de l'erreur résiduelle par pixel. D'après ces courbes, il est clair que les deux approches 1 et 2 échouent pour tous les modèles compte tenu des valeurs élevées des erreurs. Notons que pour Blinn-Phong, l'erreur d'estimation de l'approche 1 est beaucoup plus petite que celle de l'approche 2, mais, reste trop loin de la vérité terrain (celà peut être bien observé sur la figure 4). Ainsi, nous pouvons conclure qu'une

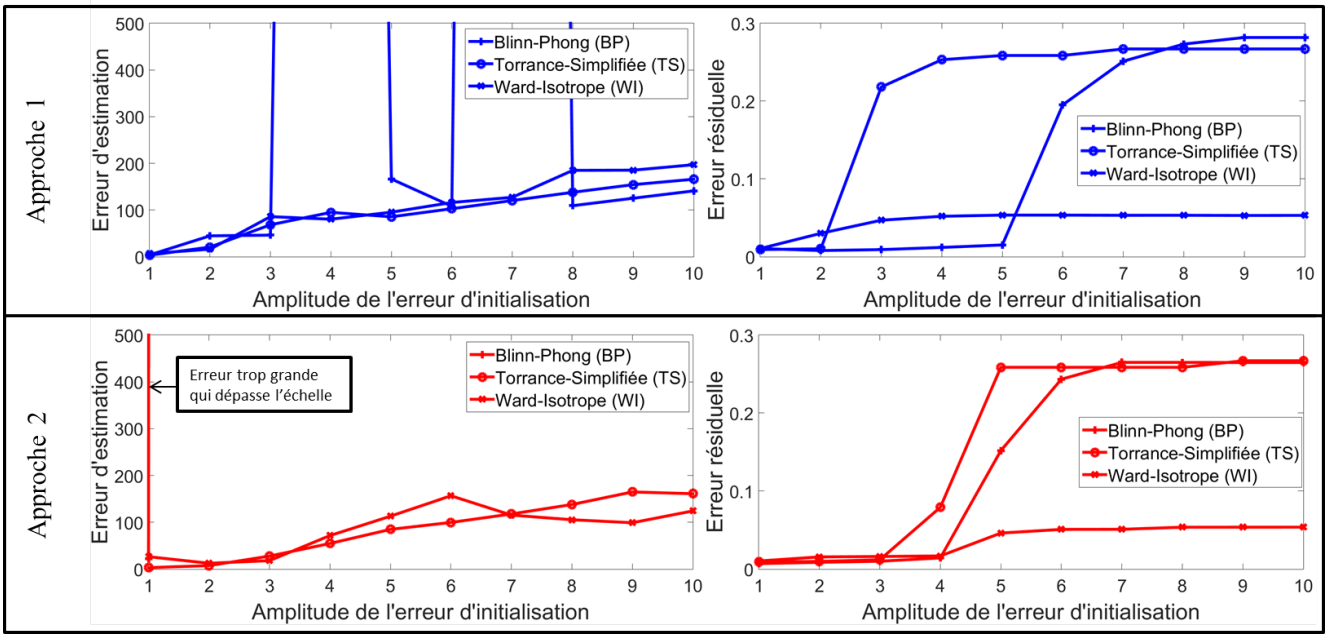


Figure 4 – *Comparaison entre les trois modèles d'illumination. Les erreurs d'estimation sont affichés sur les courbes à droite. Les erreurs résiduelles par pixel sont affichés dur les courbes à gauche.*

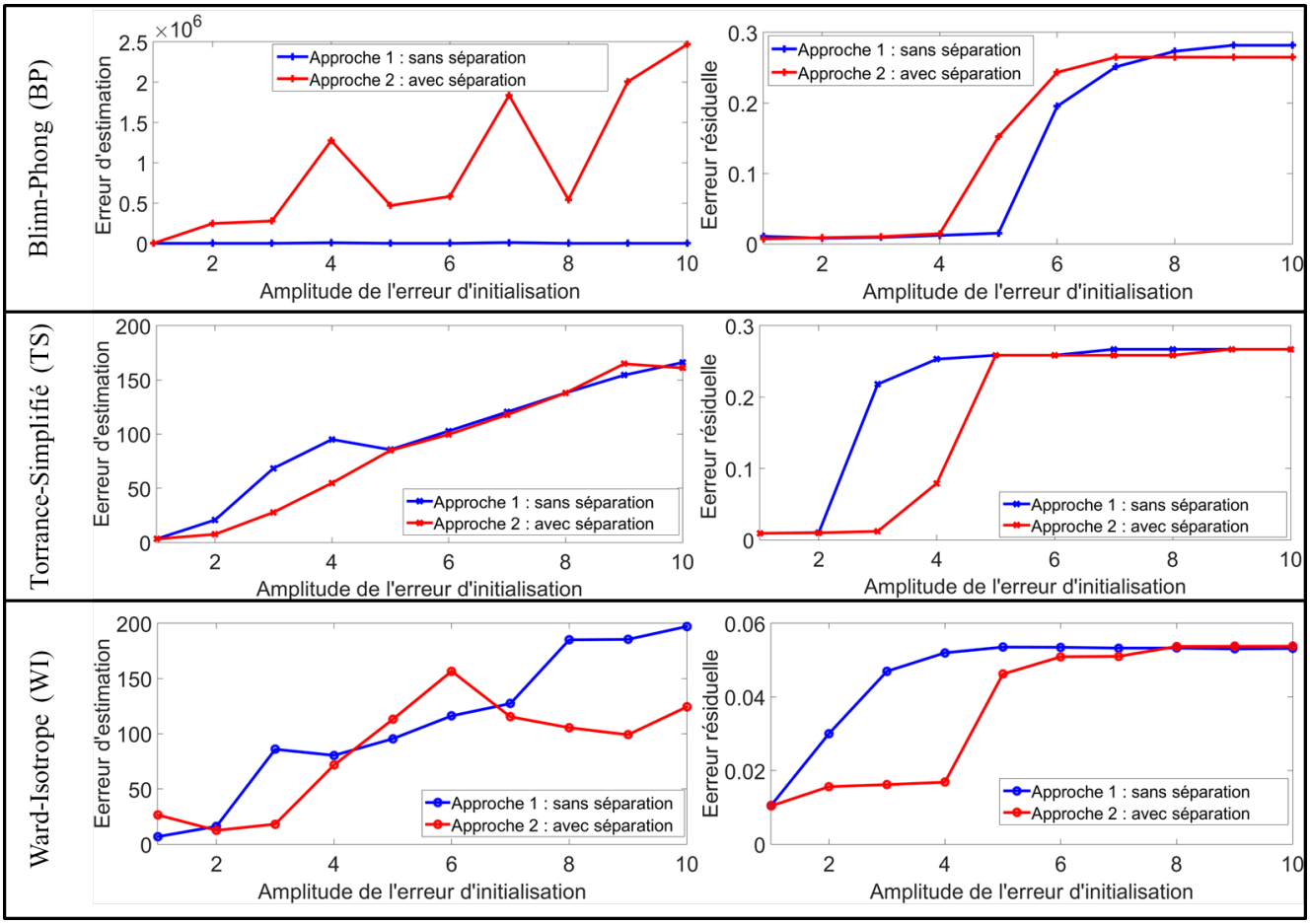


Figure 5 – *Comparaison entre les deux approches testées. Les erreurs d'estimation sont affichés sur les courbes à droite. Les erreurs résiduelle par pixel sont affichés dur les courbes à gauche.*

approche mono-tache ne permet pas de résoudre le problème avec les trois modèles d'illumination testés. Néanmoins, sur les courbes de l'erreur résiduelle, nous observons que le modèle WI est le plus stable parmi les trois modèles testés mais l'erreur d'estimation reste élevée avec les deux approches.

7 Conclusion

Dans cette étude, nous adressons le problème d'inversion d'illumination locale à partir d'une unique tache spéculaire. Nous vérifions sur des images de synthèse sous diverses conditions et avec trois modèles d'illumination que les approches mono-taches sont incapables de résoudre ce problème. Nous en concluons aussi que pour des images réelles qui sont des cas plus difficiles, les approches monotaches ne résolvent pas le problème d'inversion de l'illumination locale. Ceci permet d'orienter notre recherche vers une solution optimale et de comprendre toute incertitude sur les approches de l'état de l'art qui eux proposent de résoudre ce problème à partir d'une seule image mais de plusieurs taches spéculaires.

8 Perspectives

Les résultats nous permettent d'orienter mieux notre étude. Ainsi, dans le futur, nous étudierons une approche multitaches. Comme les erreurs sont élevées pour les approches 1 et 2, nous essayerons de d'évaluer si la séparation des deux composantes diffuses et spéculaires est indispensable. Une base d'images réelles avec vérité terrain sera testée. Finalement, il serait intéressant de proposer une approche analytique pour résoudre ce problème.

References

[1] J. F. Blinn. Models of light reflection for computer synthesized pictures. In *ACM SIGGRAPH Computer Graphics*, volume 11, pages 192–198, 1977.

[2] S. Boivin and A. Gagalowicz. Image-based rendering of diffuse, specular and glossy surfaces from a single image. In *ACM SIGGRAPH Computer graphics*, pages 107–116, 2001.

[3] S. Hadj Said, M. Tamaazousti, and A. Bartoli. Image-based models for specularity propagation in diminished reality. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[4] K. Hara, K. Nishino, et al. Light source position and reflectance estimation from a single view without the distant illumination assumption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):493–505, 2005.

[5] J. Jachnik, R. A. Newcombe, and A. J. Davison. Real-time surface light-field capture for augmentation of planar specular surfaces. In *IEEE ISMAR*, pages 91–97, 2012.

[6] J. T. Kajiya. The rendering equation. In *ACM SIGGRAPH Computer Graphics*, volume 20, pages 143–150, 1986.

[7] H. Kim, H. Jin, S. Hadap, and I. Kweon. Specular reflection separation using dark channel prior. In *IEEE CVPR*, pages 1460–1467, 2013.

[8] Y. Liu, Z. Yuan, N. Zheng, and Y. Wu. Saturation-preserving specular reflection separation. In *IEEE CVPR*, pages 3725–3733, 2015.

[9] B. Mercier, D. Meneveaux, and A. Fournier. A framework for automatically recovering object shape, reflectance and light sources from calibrated images. *International Journal of Computer Vision*, 73(1):77–93, 2007.

[10] A. Morgand, M. Tamaazousti, and A. Bartoli. A geometric model for specularity prediction on planar surfaces with multiple light sources. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[11] A. Morgand, M. Tamaazousti, and A. Bartoli. A multiple-view geometric model of specularities on non-planar shapes with application to dynamic retexturing. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[12] S. K. Nayar, K. Ikeuchi, and T. Kanade. Surface reflection: physical and geometrical perspectives. Technical report, 1989.

[13] K. Nishino, Z. Zhang, and K. Ikeuchi. Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis. In *IEEE ICCV*, volume 1, pages 599–606, 2001.

[14] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[15] R. T. Tan and K. Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):178–193, 2005.

[16] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America*, 57(9):1105–1112, 1967.

[17] G. J. Ward. Measuring and modeling anisotropic reflection. *ACM SIGGRAPH Computer Graphics*, 26(2):265–272, 1992.

[18] S. Xu and A. M. Wallace. Recovering surface reflectance and multiple light locations and intensities from image data. *Pattern Recognition Letters*, 29(11):1639–1647, 2008.

[19] E. Zhang, M. F. Cohen, and B. Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics*, 35(6):174, 2016.

Analyse de trajectoires sur des variétés de matrices pour la reconnaissance des expressions faciales *

A. Kacem[1] M. Daoudi[1] B. Ben Amor[1]

[1] IMT Lille Douai,
Univ. Lille CNRS, UMR 9189 - CRIStAL
Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

Résumé

Dans ce travail nous nous intéressons au problème de reconnaissance des expressions faciales en se basant sur une analyse invariante aux variabilités temporelles des trajectoires de matrices vivant sur des variétés Riemannienes bien définies. Nous considérons l'évolution temporelle des matrices X des marqueurs de visages pour construire une trajectoire sur une Grassmannienne $\mathscr{G}(n,2)$. Une autre trajectoire est introduite sur une variété moins connue, $\mathscr{S}^+(n,2)$ ou encore le cône des matrices symétriques semi-définie positives de rang fixe 2. Cette dernière est obtenue en considérant l'évolution temporelle des matrices de Gram XX^t obtenues à partir des matrices originales des marqueurs de visages notées X. Nous avons développé des outils géométriques pour aligner, comparer et calculer des moyennes statisitiques d'un ensemble de trajectoires sur les deux variétés Riemanniennes considérées. L'approche proposée, testée sur la base de données CK+, a donné des résultats de reconnaissance compétitifs par rapport à ceux de la littérature tandis qu'elle ne requiert pas des techniques d'apprentissage automatique.

Mots clefs

Expressions faciales, Matrices de Gram, Géométrie Riemannienne, Alignement temporel

1 Introduction

La comparaison de séquences faciales est un problème fondamental en reconnaissance des expressions faciales. En plus des propriétés d'invariance telles que l'invariance aux transformations géométriques nécessaires aux algorithmes d'analyse des expressions faciales, il faut s'assurer que ces méthodes sont robustes aux variabilités temporelles des séquences faciales. En effet, des séquences non alignées produisent des erreurs dans le calcul des métriques lors de la comparaison des séquences. Par conséquent, les quantités statistiques, telles que la moyenne des séquences temporelles, ne sont pas pertinentes pour la tâche de classification car leurs valeurs sont erronées. La méthode la plus utilisée pour résoudre le probléme de l'alignement temporelle des caratéristiques temporelles représentant les séquences est la programmation dynamique (DTW). Bien que cette approche et d'autres variantes aient été utilisés pour l'alignement temporelle, ces méthodes utilisent des métriques Euclidiennes, et n'exploitent pas la non-linéarité et la dynamique de l'évolution des caractéristiques extraites à partir des séquences. Récemment, la géométrie Riemannienne a trouvé un large panel d'applications en vision par ordinateur telles que la reconnaissance des piétons [1] et l'analyse des séquences vidéos [2]. Elle offre un cadre unifié pour calculer des métriques dans des variétés non linéaires, et permet aussi de faire des calculs statistiques dans ces variétés. Dans cet article, nous cherchons une représentation suffisamment discriminative pour classer une séquence donnée à l'une des classes d'expressions faciales invariantes à la variabilité temporelle et aux mouvement rigides des marqueurs du visage. Nous proposons d'embarquer les séquences temporelles des configurations de marqueurs du visage dans la variété de Grassmann et dans la variété de matrices symétriques semi-définies positives de rang fixe. Ainsi, les séquences sont considérées comme des trajectoires paramétrées par le temps sur ces deux variétés. La figure 1 résume l'approche proposée.
Les principales contributions de ce travail sont :

- Une nouvelle représentation géométrique de la dynamique des séquences faciales par des trajectoires paramétrées par le temps sur la variété Riemannienne des matrices symétriques semi-définie positives de rang fixe.
- Des outils statistiques pour la classification des trajectoires sur ces variétés.
- Une analyse invariante aux variabilités temporelles en utilisant le DTW (Dynamic Time Warping) appliqué à ces trajectoires.

Le reste du papier est organisé comme suit : Dans la section 2 nous présentons brièvement la géométrie des variétés étudiées. Dans la section 3 nous décrivons notre approche proposée. Les résultats expérimentaux et les discussions sont fournis dans la section 4. Finalement, nous concluons et présentons quelques perspectives dans la section 5.

*Une version étendue de ce papier a été présentée lors des journées ORASIA 2017

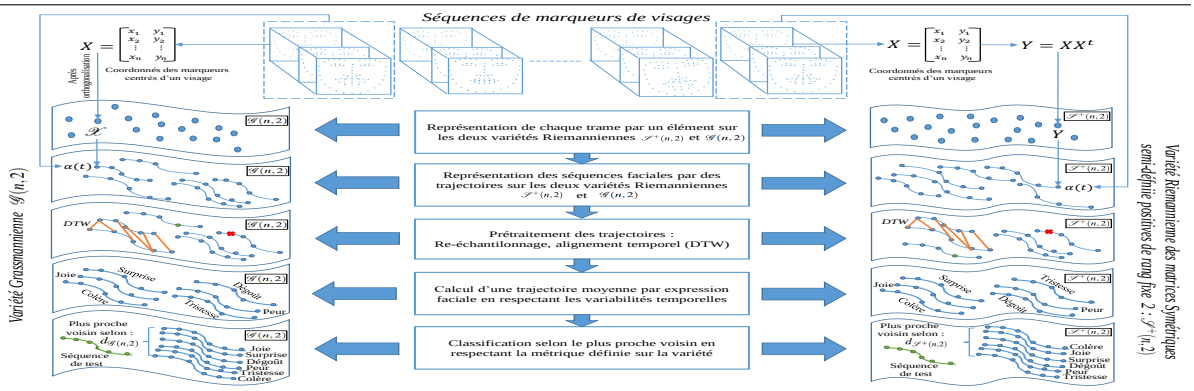


Figure 1 – *Vue d'ensemble de l'approche proposée*

2 Fondements théoriques

Dans cette section, nous présentons brièvement les bases théoriques des variétés Grassmannienne, la variété des matrices symétriques définies positives et la variété des matrices symétriques semi-définies positives de rang fixe.

2.1 Variété Grassmannienne

Soit $\mathcal{G}(n,k)$ l'ensemble des sous-espaces vectoriels de dimension k de $\mathbb{R}^n$ (e.g. $\mathcal{G}(3,2)$ est l'ensemble des plans de dimension 2 de $\mathbb{R}^3$) où $n > k$. Un sous-espace $\mathcal{X}$ de $\mathcal{G}(n,k)$, est représenté par une matrice X de taille $n \times k$ dont les vecteurs colonnes forment une base orthogonale de ce sous-espace. On dit que la matrice X engendre le sous-espace $\mathcal{X}$.

L'ensemble des matrices de taille $n \times k$ avec des vecteurs colonnes orthogonaux forme une variété de Stiefel $\mathcal{V}(n,k)$. Les éléments de la Grassmannienne $\mathcal{G}(n,k)$ sont des classes d'équivalence des éléments de $\mathcal{V}(n,k)$ de façon que deux éléments sont équivalents si leurs bases orthogonales engendrent le même sous-espace [3]. Plus précisément, soit une matrice X qui représente un élément de $\mathcal{V}(n,k)$, si on effectue une multiplication à droite de X par une matrice orthogonale O de taille $k \times k$ on obtient une nouvelle matrice X_1. D'une part, X_1 représente un autre élément différent de celui de X dans $\mathcal{V}(n,k)$, d'autre part X_1 représente le même élément que X dans $\mathcal{G}(n,k)$. Ainsi, on peut noter l'invariance de $\mathcal{G}(n,k)$ aux transformations orthogonales de $\mathbb{R}^k$.

Une distance géodésique de la variété Grassmannienne est donnée par la norme l_2 du vecteur formé par les angles principaux entre deux sous-espaces. Etant donnés deux sous-espaces $\mathcal{X}_1, \mathcal{X}_2 \in \mathcal{G}(n,k)$ engendrés, respectivement, par X_1 et X_2, la distance géodésique $d_{\mathcal{G}}$ entre $\mathcal{X}_1$ et $\mathcal{X}_2$ est définie par :

$$d^2_{\mathcal{G}}(\mathcal{X}_1, \mathcal{X}_2) = \sum_i \theta_i^2 \tag{1}$$

Où, $\theta_i = \cos^{-1}\left(\max_{u_i \in \mathcal{X}_1} \max_{v_i \in \mathcal{X}_2} \langle u_i^t, v_i \rangle\right)$, u et v sont les vecteurs des bases qui engendrent, respectivement, $\mathcal{X}_1$ et $\mathcal{X}_2$, $\langle .,. \rangle$ désigne le produit scalaire dans $\mathbb{R}^n$.

2.2 Variété Riemannienne des matrices symétriques définies positives

Soit $\mathcal{S}^{++}(k)$ l'ensemble des matrices symétriques définies positives de taille $k \times k$. Une matrice symétrique à coefficients réels est dite définie positive, si et seulement si, pour tout x non nul $\in \mathbb{R}^n$, $x^{\mathrm{T}}Mx > 0$. L'espace de ces matrices n'est pas un espace vectoriel, si on multiplie une matrice appartenant à $\mathcal{S}^{++}(k)$ par un scalaire négatif, on obtient une matrice qui n'est plus définie positive et donc n'appartient plus à $\mathcal{S}^{++}(k)$. En équipant $\mathcal{S}^{++}(k)$ d'une métrique Riemannienne, plusieurs travaux l'ont utilisé pour étudier la variété Riemannienne des matrices de covariance [4, 5, 6]. Plusieurs métriques ont été proposées pour $\mathcal{S}^{++}(k)$, les plus utilisées sont : (1) *"affine-invariant metric"* [7], (2) *"log-euclidean metric"* [8]. Dans ce travail, nous considérons uniquement la métrique *"affine-invariant metric"* qui est une vraie métrique Riemannienne. Etant données deux matrices S_1 et $S_2 \in \mathcal{S}^{++}(k)$, la distance géodésique est donnée par ,

$$d^2_{\mathcal{S}^{++}}(S_1, S_2) = \|\log(S_1^{-1/2} S_2 S_1^{-1/2})\|_F \tag{2}$$

Où $\|.\|_F$ désigne la norme *Frobenius*.

2.3 Variété Riemannienne des matrices symétriques semi-définies positives de rang fixe

Soit $\mathcal{S}^+(n,k)$ l'ensemble des matrices symétriques semi-définies positives de taille $n \times n$ et de rang $k < n$. Une matrice symétrique à coefficients réels est dite semi-définie positive, si et seulement si, pour tout x non nul $\in \mathbb{R}^n$, $x^{\mathrm{T}}Mx \geq 0$. De même que $\mathcal{S}^{++}(k)$, $\mathcal{S}^+(n,k)$ n'est pas un espace vectoriel. Bonnabel *et al.* [9] ont introduit une métrique Riemannienne pour $\mathcal{S}^+(n,k)$ qui est presque égale à la somme des métriques dans la Grassmannienne $\mathcal{G}(n,k)$ et la variété des matrices symétriques définies positives $\mathcal{S}^{++}(k)$. Pour aboutir à ces conclusions, les auteurs de [9] ont procédé par les factorisations matricielles suivantes :

$$Y = XX^t = (UR)(UR)^t = UR^2U^t \tag{3}$$

Où $Y \in \mathscr{S}^+(n,k)$, X est une matrice de taille $n \times k$ et de rang k, $U \in \mathscr{V}(n,k)$ et $R \in \mathscr{S}^{++}(k)$.
La première factorisation est effectuée par une décomposition de *Cholesky* et la deuxième par une décomposition polaire [10]. Suite à la première factorisation, on peut remarquer l'invariance de ces matrices $Y \in \mathscr{S}^+(n,k)$ par rapport aux transformations orthogonales appliquées à la matrice X.

Démonstration 1. *Soit $O \in \mathscr{O}(k)$ une matrice orthogonale, X_1 et X_2 deux matrices de taille $n \times k$ et de rang k et $Y_1, Y_2 \in \mathscr{S}^+(n,k)$ tel que $X_1 = X_2 O$, $Y_1 = X_1 X_1^t$ et $Y_2 = X_2 X_2^t$ alors,*

$$Y_1 = X_1 X_1^t = (X_2 O)(X_2 O)^t = X_2 O O^t X_2^t = X_2 X_2^t = Y_2$$

Avec la deuxième factorisation, une matrice $Y \in \mathscr{S}^+(n,k)$ peut être représentée par un couple $(U, R^2) \in \mathscr{V}(n,k) \times \mathscr{S}^{++}(k)$. Bonnabel *et al.* [9] ont identifié une représentation de $\mathscr{S}^+(n,k)$ par un espace quotient où les éléments de $\mathscr{S}^+(n,k)$ sont des classes d'équivalence du groupe orthogonal $\mathscr{O}(k)$:

$$\mathscr{S}^+(n,k) \approx \mathscr{V}(n,k) \times \mathscr{S}^{++}(k) / \mathscr{O}(k) \tag{4}$$

En se basant sur cette représentation, les auteurs ont introduit une quasi-géodésique reliant deux éléments $Y_1, Y_2 \in \mathscr{S}^+(n,k)$,

$$\gamma_{Y_1 \to Y_2} : [0,1] \to \mathscr{S}^+(n,k)$$

$$\gamma_{Y_1 \to Y_2}(t) = U(t) R^2(t) U^t(t) \tag{5}$$

Où $U(t)$ est une géodésique sur la Grassmannienne $\mathscr{G}(n,k)$ et $R^2(t)$ est une géodésique sur $\mathscr{S}^{++}(k)$. La longueur de cette quasi-géodésique représente une mesure de similarité entre $Y_1 = U_1 R_1^2 U_1^t$ et $Y_2 = U_2 R_2^2 U_2^t$,

$$d^2_{\mathscr{S}^+}(Y_1, Y_2) = d^2_{\mathscr{G}(n,k)}(U_1, U_2) + \lambda d^2_{\mathscr{S}^{++}(k)}(R_1^2, R_2^2) \tag{6}$$

Où $\lambda > 0$ est un paramètre qui contrôle la contribution des distances de $\mathscr{S}^{++}(k)$ et $\mathscr{G}(n,k)$. Il est à préciser que $d_{\mathscr{S}^+(n,k)}$ n'est pas une distance car elle ne vérifie pas l'inégalité triangulaire et que des valeurs faibles de λ sont recommandées [9].

2.4 Moyenne intrinsèque

Etant donné L échantillons $S_1, S_2, .., S_L$ sur une variété Riemannienne $\mathscr{M}$, nous nous intéressons au calcul d'une moyenne intrinsèque $\tilde{S}$ qui vit sur cette variété $\mathscr{M}$. En effet, une moyenne Euclidienne calculée à partir de ces matrices n'appartiendra pas à cette variété à cause de la non-linéarité de cet espace. Pour remédier à ce problème, nous utilisons la métrique $d_{\mathscr{M}}$ définie sur $\mathscr{M}$ pour le calcul d'une moyenne intrinsèque $\tilde{S} \in \mathscr{M}$ selon :

$$\tilde{S} = \arg\min_{S \in \mathscr{M}} \sum_{i=1}^{L} d_{\mathscr{M}}(S, S_i)^2, \tag{7}$$

qui minimise l'erreur quadratique moyenne en utilisant une métrique appropriée $d_{\mathscr{M}}$ [11]. Des algorithmes itératifs ont été proposé pour résoudre ce problème pour $\mathscr{G}(n,k)$ [3] et $\mathscr{S}^{++}(k)$ [12].
Dans [9], les auteurs ont introduit une moyenne intrinsèque $\tilde{Y} \in \mathscr{S}^+(n,k)$ de deux éléments $Y_1 = U_1 R_1^2 U_1^t$ et $Y_2 = U_2 R_2^2 U_2^t \in \mathscr{S}^+(n,k)$:

$$\tilde{Y} = \tilde{U} \tilde{R}^2 \tilde{U}^t \tag{8}$$

Où $\tilde{U}$ est une moyenne intrinsèque de U_1 et U_2 dans $\mathscr{G}(n,k)$ et $\tilde{R}^2$ est une moyenne intrinsèque de R_1^2 et R_2^2 dans $\mathscr{S}^{++}(k)$.
Pour calculer une moyenne intrinsèque d'un ensemble d'éléments sur $\mathscr{S}^+(n,k)$, nous proposons de calculer une moyenne intrinsèque des matrices obtenues après la factorisation sur $\mathscr{G}(n,k)$ selon la méthode décrite dans [3] et une autre moyenne sur $\mathscr{S}^{++}(k)$ selon [11]. La moyenne de cet ensemble d'échantillons sur $\mathscr{S}^+(n,k)$ est obtenue en combinant les deux moyennes selon l'équation (8).

3 Représentation des séquences faciales par des trajectoires sur des variétés Riemanniennes

Pour analyser les séquences faciales, nous nous sommes intéressés uniquement aux marqueurs des visages et leur évolution temporelle. Dans ce contexte, plusieurs travaux ont représenté l'évolution temporelle des marqueurs par des trajectoires paramétrées par le temps sur des variétés Riemanniennes bien définies : variété Grassmannienne [13, 14], espace de formes de Kendall [15], etc. Dans cette section, nous introduisons une nouvelle approche pour représenter les marqueurs des visages et leur évolution temporelle par des trajectoires de matrices de Gram qui vivent sur la Variété Riemannienne des matrices symétriques semi-définies positives de rang fixe 2. Nous commençons par aligner ces trajectoires dans le temps pour avoir une distance significative entre elles. Ensuite, nous présentons quelques outils permettant de les analyser.

3.1 Trajectoires de marqueurs de visages

La représentation par des marqueurs de visages est une technique très utilisée pour modéliser approximativement la géométrie des visages. Dans un premier temps, en s'inspirant de [13], nous avons modélisé la géométrie du visage par une matrice $X = [(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]^t$, de taille $n \times 2$ où n représente le nombre de marqueurs pour un visage. Après des procédures d'orthogonalisation et de centrage, ces matrices vivent sur une Grassmannienne $\mathscr{G}(n,2)$ que nous avons présentée dans la section 2.1.
En se basant sur cette représentation, nous proposons dans notre travail de représenter chaque trame par une matrice $Y = XX^t$ de taille $n \times n$ appelée matrice de Gram. D'une part, ces matrices sont très riches car elles décrivent la corrélation entre les coordonnées des marqueurs du visage et donc, les petits déplacements des marqueurs seront plus prononcés dans ces matrices. D'autre part, elles permettent de représenter et analyser les matrices des mar-

queurs X sans passer par une orthogonalisation qui engendre une perte d'information. Les matrices de Gram sont symétriques semi-définies positives et de rang égal à celui de la matrice primitive ,i.e. $rang(Y) = rang(X)$. Dans notre cas, on peut remarquer que $rang(X) = 2$, et donc ces matrices de Gram ont un rang fixe égal à 2 et vivent sur la variété Riemannienne $\mathscr{S}^+(n,2)$ des matrices symétriques semi-définies positives de rang 2 qui a été présentée dans la section 2.3. Il est à préciser que nous n'effectuons pas une décomposition de *Cholesky* comme indiqué dans l'équation (3). En effet, les matrices primitives des marqueurs X vérifient les conditions de la première factorisation (X est une matrice de taille $n \times 2$ et de rang 2) et sont utilisées pour la deuxième factorisation.
En considérant la succession de ces matrices, l'analyse des séquences faciales revient à analyser les trajectoires de ces matrices sur les variétés sur lesquelles elles vivent. Une trajectoire sur une variété Riemannienne $\mathscr{M}$ qui généralise $\mathscr{S}^+(n,2)$ et $\mathscr{G}(n,2)$ peut être considérée comme une courbe $\alpha : I \to \mathscr{M}$, où I désigne le domaine temporel.
Un autre avantage de ces représentations, est l'invariance des marqueurs aux rotations de $\mathbb{R}^2$ due à l'invariance aux transformations orthogonales mentionnée dans les sections 2.1 et 2.3. Ainsi, l'alignement spatial des marqueurs de visage n'est pas nécessaire avant l'analyse de ces trajectoires.
Veuillez noter que nous utilisons une notation $(\mathscr{M}, d_{\mathscr{M}})$ pour généraliser $(\mathscr{G}(n,2), d_{\mathscr{G}(n,2)})$ et $(\mathscr{S}^+(n,2), d_{\mathscr{S}^+(n,2)})$.

3.2 Re-échantillonnage adaptatif des trajectoires

Un outil important dans notre approche consiste à augmenter ou réduire le nombre d'échantillons dans une trajectoire.
D'une part, l'augmentation des échantillons implique une plus haute résolution temporelle des séquences qui garantie une meilleur performance dans l'alignement temporel ou le calcul de la moyenne des séquences temporelles. Pour augmenter le nombre d'échantillons d'une trajectoire, nous cherchons les formes de visages qui ont des distances maximales par rapport à l'échantillon précédent et nous générons un nouvel échantillon entre eux, défini par la moyenne géométrique de l'échantillon et son précédent.
D'autre part, pour réduire les échantillons, nous supprimons de la séquence les échantillons ayant les distances minimales par rapport à l'échantillon précédent. En d'autres termes, nous éliminons les échantillons les moins importants représentant des formes de visages similaires à celles qui les précédent. La réduction des échantillons nous permet de réduire les temps de calcul.

3.3 Alignement temporel des trajectoires

Pour pouvoir comparer et analyser les séquences faciales, nous devons tenir compte des variabilités temporelles qui peuvent survenir. L'alignement temporel des séquences faciales, qui sont représentées par des trajectoires sur des variétés Riemanniennes consiste à aligner dans le temps ces trajectoires.
Etant donnée une variété Riemannienne $\mathscr{M}$ équipée d'une métrique $d_{\mathscr{M}}$ et deux trajectoires sur $\mathscr{M}$: $\alpha_1(t)$, $\alpha_2(t)$: $I \to \mathscr{M}$ où I désigne le domaine temporel, le problème de l'alignement des deux trajectoires α_1 et α_2 revient à trouver la fonction de re-paramétrisation optimale $\gamma^\star$ appliquée sur l'une des trajectoires et minimisant la distance entre elles selon,

$$\gamma^\star = \arg\min_{\gamma \in \Gamma} \int_I d_{\mathscr{M}}(\alpha_1(t), \alpha_2(\gamma(t)))\,\mathrm{d}t \tag{9}$$

Où Γ désigne l'ensemble des fonctions croissantes $\gamma : I \to I$. La méthode la plus utilisée pour résoudre ce problème d'optimisation est l'algorithme de *Dynamic Time Warping (DTW)*. L'adaptation du DTW pour des séquences de matrices vivant sur des variétés Riemanniennes peut être fait en considérant la métrique appropriée au lieu de la distance Euclidienne.

3.4 Calcul d'une trajectoire moyenne

Pour générer des modèles pour les différentes expressions faciales, nous proposons de calculer une séquence faciale moyenne par expression faciale tout en respectant les variabilités temporelles qui peuvent survenir. Les séquences faciales étant représentées par des trajectoires sur une variété Riemannienne $\mathscr{M}$, le problème du calcul de cette moyenne revient à calculer une trajectoire moyenne intrinsèque à la variété $\mathscr{M}$ tout en respectant les variabilités temporelles. Pour cela, nous procédons comme suit : La trajectoire moyenne est initialisée aléatoirement à l'une des trajectoires en entrée. Ensuite, nous alignons dans le temps ces trajectoires à la trajectoire moyenne en utilisant le DTW introduit dans la section 3.3. Une nouvelle trajectoire moyenne est obtenue en calculant une moyenne intrinsèque *(élément par élément)* de toutes les trajectoires comme présenté dans la section 2.4. Finalement, nous itérons ces étapes jusqu'à convergence atteinte lorsque la distance $d_{\mathscr{M}}$ entre la nouvelle trajectoire moyenne et la trajectoire moyenne courante est inférieure à un certain petit seuil ε. Nous pouvons noter qu'une étape de re-échantillonnage des trajectoires, comme mentionné dans la section section 3.2, est nécessaire pour avoir un nombre fixe d'échantillons pour toutes les trajectoires.

4 Résultats expérimentaux

Pour illustrer l'efficacité de l'approche proposée et comparer les deux métriques étudiées dans ce travail (trajectoires sur $\mathscr{G}(n,2)$ et $\mathscr{S}^+(n,2)$), nous avons utilisé la base de données CK+.

4.1 Corpus de test

La base de données CK+ (The Cohn-Kanade Extended Facial Expression) [16] a été développé pour l'analyse et la synthèse des expressions faciales. Elle contient 123 sujets et 593 séquences d'images frontales. Parmi ces sujets, 118

ont été annotés par les sept expressions universelles (la colère, le mépris, le dégoût, la peur, la joie, la tristesse et la surprise). Les visages sont annotées avec 68 marqueurs en deux dimensions que nous utilisons dans notre approche.

4.2 Alignement des trajectoires et calcul d'une trajectoire moyenne

Pour montrer l'importance de l'alignement temporel avant la comparaison des séquences comme indiqué dans la section section 3.3, nous avons sélectionné deux séquences de la base CK+ de deux sujets effectuant la même expression faciale (la joie) avec des variabilités temporelles (voir la partie gauche de la figure 2). Les résultats de l'alignement temporel de ces séquences sur $\mathscr{G}(n,2)$ et $\mathscr{S}^+(n,2)$ sont donnés par la partie droite de la figure 2. Les distances à la première trame sont utilisées comme indiqué pour illustrer les phases temporelles avant et après l'alignement temporel. On remarque une différence peu significative entre les deux courbes entre la trame 5 et la trame 15 qui peut être vue dans la partie gauche de la figure 2 où, l'alignement dans $\mathscr{S}^+(n,2)$ est visuellement plus performant que dans $\mathscr{G}(n,2)$.

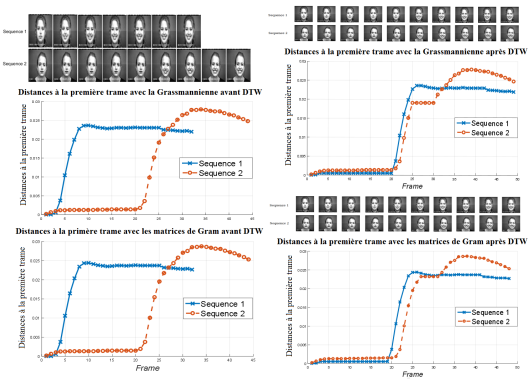


Figure 2 – *Gauche : Distances à la première trame avant l'alignement temporel ; Droite : Après l'alignement temporel.*

Pour montrer l'efficacité de la méthode proposée du calcul de la moyenne spatio-temporelle des séquences, nous avons calculé une trajectoire moyenne dans $\mathscr{G}(n,2)$ et $\mathscr{S}^+(n,2)$ de toutes les séquences de la classe "joie" (69 séquences) comme décrit dans la section 3.4. La figure 3 montre 5 trames de chaque trajectoire moyenne obtenue. On remarque visuellement que la trajectoire moyenne dans $\mathscr{S}^+(n,2)$ [1] montre des déformations plus accentuées que celle dans $\mathscr{G}(n,2)$. Veuillez noter que nous avons choisi une rotation arbitraire pour visualiser les marqueurs puisque notre approche est invariante aux rotations de $\mathbb{R}^2$.

1. Nous avons appliqué une *décomposition de Cholesky* sur les matrices de Gram pour visualiser la configuration des marqueurs correspondante.

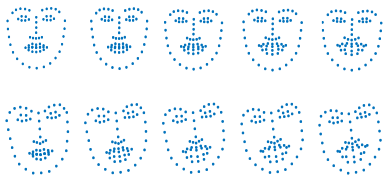


Figure 3 – *Trajectoires moyennes (haut) sur $\mathscr{G}(n,2)$ et (bas) sur $\mathscr{S}^+(n,2)$, calculées à partir de 69 sujets effectuant l'expression joie de la base CK+.*

4.3 Reconnaissance des expressions faciales

Afin de valider notre représentation et évaluer la robustesse des outils statistiques introduits, nous nous sommes intéressés à la reconnaissance des expressions faciales dans la base CK+. A cet effet, nous avons généré une trajectoire moyenne pour chacune des 6 classes (colère, dégoût, peur, joie, tristesse et surprise). Ensuite, nous avons aligné toutes les séquences (trajectoires) de test à ces trajectoires moyennes comme indiqué dans la section 3.3. Finalement, nous avons utilisé la méthode du plus proche voisin pour prédire la classe de la séquence de test. En suivant le protocole expérimental le plus utilisé pour cette base de données [17, 18, 13, 19], nous avons effectué une validation croisée *(leave-one-subject-out)*. En utilisant des trajectoires sur $\mathscr{S}^+(n,2)$ nous avons obtenu un taux de reconnaissance moyen de 87.7% dépassant celui des trajectoires dans $\mathscr{G}(n,2)$ de 1.8% (et donc 5 séquences sont reconnues en plus). Cette amélioration est illustrée plus en détails dans les matrices de confusion (figure 4) où *la joie*, *la colère*, *la surprise* et *le dégoût* sont un peu plus reconnus dans $\mathscr{S}^+(n,2)$.

Impact du paramètre λ – Nous rappelons que la métrique définie sur $\mathscr{S}^+(n,2)$ dans l'équation (6) fait intervenir un paramètre $\lambda > 0$ qui contrôle la contribution des métriques sur la Grassmannienne $\mathscr{G}(n,2)$ et $\mathscr{S}^{++}(2)$. Bonnabel *et al.* [9] recommandent d'utiliser des petites valeurs pour ce paramètre. Des résultats de reconnaissance pour différentes valeurs de λ sont illustrés dans la figure 4. Une meilleure performance est obtenue pour une valeur de $\lambda = 0.01$ qui est maintenue pour le reste des expérimentations.

Différences entre $\mathscr{G}(n,2)$ and $\mathscr{S}^+(n,2)$ – Comme indiqué dans l'équation (6), la mesure de similarité (qui n'est pas une distance) contient deux termes. Le premier terme se résume à la métrique de la Grassmannienne qui décrit la forme construite par les marqueurs, et le deuxième encode la corrélation entre ces marqueurs et se résume à la métrique sur $\mathscr{S}^{++}(2)$. Le paramètre λ discuté dans la section précédente permet de contrôler la contribution de ces deux termes. En considérant uniquement le premier terme ($\lambda = 0$), l'étude sur $\mathscr{S}^+(n,2)$ se résume à celle sur $\mathscr{G}(n,2)$. En ajoutant le deuxième terme, le taux de reconnaissance passe de 86.08% à 87.87% pour une valeur de $\lambda = 0.01$.

Impact de la résolution des trajectoires – Un autre paramètre important dans notre approche est le nombre

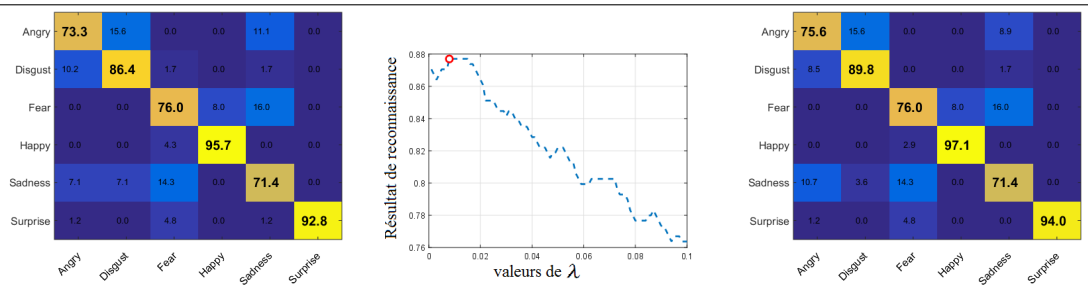


Figure 4 – *De gauche à droite - Matrice de confusion de la reconnaissance des trajectoires sur $\mathscr{G}(n,2)$; Taux de reconnaissance pour différentes valeurs de λ ; Matrice de confusion de la reconnaissance des trajectoires sur $\mathscr{S}^+(n,2)$.*

d'échantillons à considérer dans une trajectoire. En effet, les trajectoires qui représentent les séquences doivent contenir le même nombre d'échantillons (trames) pour pouvoir être comparées. A cet effet, nous avons utilisé la méthode de re-échantillonnage adaptatif décrite dans la section 3.2. Le meilleur taux de reconnaissance est obtenu pour une valeur de 50 échantillons.

Impact de l'alignement temporel (DTW) – Pour quantifier la contribution de l'alignement temporel des trajectoires, nous avons réalisé les mêmes expériences sans appliquer le DTW. La performance de notre approche avec des trajectoires sur $\mathscr{G}(n,2)$ diminue de 86.08% à 78.96% et de 87.87% à 79.05% avec des trajectoires sur $\mathscr{S}^+(n,2)$.

Complexité de calcul – Un avantage de notre méthode est son efficacité en termes de temps d'exécution. Contrairement à [19], notre approche ne nécessite aucune normalisation pour une analyse invariante aux rotations de $\mathbb{R}^2$. Dans le tableau 1, nous présentons le temps d'exécution [2] nécessaire (en millisecondes) pour comparer deux trajectoires arbitraires avec et sans alignement temporel (DTW). Nous rappelons que nous utilisons 68 marqueurs pour chaque visage et que le nombre d'échantillons considérés pour chaque trajectoire est égal à 50.

Tableau 1 – *Temps d'exécution de la comparaison de deux trajectoires sur $\mathscr{G}(n,2)$ et $\mathscr{S}^+(n,2)$.*

Trajectoires sur	**Sans DTW**	**Avec DTW**
$\mathscr{G}(n,2)$	1.055	**78.985**
$\mathscr{S}^+(n,2)$	10.118	**148.645**

Étude comparative avec l'état de l'art – Dans la littérature plusieurs travaux ont choisi la base de données CK+ pour évaluer leurs méthodes. Dans notre étude, nous nous comparons uniquement aux travaux basés sur une représentation par les marqueurs du visage sans considérer l'information donnée par l'image couleur. Notre approche donne des résultats comparables à ceux de l'état de l'art tandis qu'elle ne requiert aucune technique d'apprentissage automatique. Comme le montre le tableau 2, le taux de reconnaissance que nous avons obtenu dépasse celui de toutes les méthodes existantes à l'exception de [19]. En particulier, tandis que Taheri *et al.* [13] représentent les séquences par des trajectoires sur $\mathscr{G}(n,2)$ et calculent les vecteurs vélocités pour les utiliser dans un classifieur SVM, notre approche se base sur une métrique appropriée. De plus, le calcul de ces vecteurs vélocités est coûteux en termes de temps d'exécution.

Tableau 2 – *Étude comparative de la méthode proposée avec les méthodes existantes (basées sur les marqueurs de visages) sur la base CK+*

Méthode	**RR (%)**
Taheri *et al.* [13]	85.8
Wang *et al.* [17]	86.3
Li *et al.* [20]	87.43
Ghimire et al. [19]	97.35
Traj. sur $\mathscr{G}(n,2)$	**86.08**
Traj. sur $\mathscr{S}^+(n,2)$	**87.87**

Dans [19], les auteurs obtiennent un taux de reconnaissance de 97.35% en utilisant un classifieur SVM sur des descripteurs géométriques (distances et angles entre les marqueurs) boostés par AdaBoost. Cette représentation rend l'approche sensible aux systèmes de détection et suivi des marqueurs. De plus, cette approche requiert, (1) une normalisation géométrique pour chaque trame de la séquence, (2) un nombre de trames fixe pour toutes les séquences qui est obtenu par une interpolation linéaire tandis qu'une interpolation non-linéaire par les géodésiques est utilisée dans notre approche.

5 Conclusions

Dans ce travail, nous avons proposé une nouvelle approche géométrique pour modéliser la dynamique des séquences faciales. Les matrices de Gram relatives aux matrices des coordonnées des marqueurs des visages ont été utilisé pour représenter chaque trame d'une séquence. L'évolution temporelle des marqueurs est étudiée en considérant des trajectoires paramétrées par le temps sur la variété Riemannienne des matrices symétriques semi-définie positives de rang fixe. L'adaptation du DTW pour des trajectoires sur ces variétés a résolu le problème de variabilités temporelles

2. Notre programme Matlab est exécuté sur un ordinateur avec un processeur 2.8 GHZ CPU.

des séquences. Nous avons aussi introduit une trajectoire moyenne qui tient en compte des variabilités temporelles. En utilisant un simple classifieur basé sur l'algorithme du *"plus proche voisin"* par rapport à ces modèles, les résultats de reconnaissance obtenus sont comparables par rapport à ceux de la littérature.
Combiner notre approche de représentation avec des techniques avancées d'apprentissage automatique sur ces trajectoires pourrait augmenter nos résultats de reconnaissance. Une extension de ce travail pourrait aussi inclure l'utilisation d'autres informations sur les séquences faciales telles que la texture ou l'image 3D.

Références

[1] Oncel Tuzel, Fatih Porikli, et Peter Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(10) :1713–1727, Octobre 2008.

[2] Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, et Stefano Berretti. A grassmann framework for 4d facial shape analysis. *Pattern Recognition*, 3(3) :349–365, 2016.

[3] Pavan Turaga, Ashok Veeraraghavan, Anuj Srivastava, et Rama Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11) :2273–2286, 2011.

[4] Oncel Tuzel, Fatih Porikli, et Peter Meer. Region covariance : A fast descriptor for detection and classification. Dans *European conference on computer vision*, pages 589–600. Springer, 2006.

[5] Chunfeng Yuan, Weiming Hu, Xi Li, Stephen Maybank, et Guan Luo. Human action recognition under log-euclidean riemannian metric. Dans *Asian Conference on Computer Vision*, pages 343–353. Springer, 2009.

[6] Ruiping Wang, Huimin Guo, Larry S Davis, et Qionghai Dai. Covariance discriminative learning : A natural and efficient approach to image set classification. Dans *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2496–2503. IEEE, 2012.

[7] Xavier Pennec, Pierre Fillard, et Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1) :41–66, 2006.

[8] Vincent Arsigny, Pierre Fillard, Xavier Pennec, et Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2) :411–421, 2006.

[9] Silvere Bonnabel et Rodolphe Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3) :1055–1070, 2009.

[10] Nicholas J Higham. Computing the polar decomposition-with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4) :1160–1174, 1986.

[11] Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5) :509–541, 1977.

[12] Dario A Bini et Bruno Iannazzo. Computing the karcher mean of symmetric positive definite matrices. *Linear Algebra and its Applications*, 438(4) :1700–1710, 2013.

[13] Sima Taheri, Pavan Turaga, et Rama Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. Dans *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 306–313. IEEE, 2011.

[14] Taleb Alashkar, Boulbaba Ben Amor, Stefano Berretti, et Mohamed Daoudi. Analyzing trajectories on grassmann manifold for early emotion detection from depth videos. Dans *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–6. IEEE, 2015.

[15] Boulbaba Ben Amor, Jingyong Su, et Anuj Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE transactions on pattern analysis and machine intelligence*, 38(1) :1–13, 2016.

[16] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, et Iain Matthews. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. Dans *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[17] Ziheng Wang, Shangfei Wang, et Qiang Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3422–3429, 2013.

[18] Mengyi Liu, Ruiping Wang, Shiguang Shan, et Xilin Chen. Learning prototypes and similes on grassmann manifold for spontaneous expression recognition. *Computer Vision and Image Understanding*, 147 :95–101, 2016.

[19] Deepak Ghimire et Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6) :7714–7734, 2013.

[20] Yongqiang Li, Shangfei Wang, Yongping Zhao, et Qiang Ji. Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on Image Processing*, 22(7) :2559–2573, 2013.

Sélection non supervisée de variables par algorithme génétique pour une histologie spectrale optimale: application aux images Raman de carcinomes cutanés

Abbas Rammal[1, 2], Nathalie Mainreck[1, 2], Olivier Piot[1, 2], Cyril Gobinet[1, 2]

[1] Université de Reims Champagne-Ardenne, Equipe MéDIAN - Biophotonique et Technologies pour la Santé, UFR de Pharmacie, 51 rue Cognacq-Jay, 51096 Reims Cedex, France

[2] CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC), Reims, France

[1,2] abbas.rammal@univ-reims.fr, nathalie.mainreck@univ-reims.fr, olivier.piot@univ-reims.fr, cyril.gobinet@univ-reims.fr,

Résumé

L'imagerie spectrale Raman est une technique d'analyse de la composition biomoléculaire d'un échantillon biologique basée sur l'interaction lumière - matière. Cette technique nous permet d'obtenir des images spectrales qui sont des cubes de données 3D composé de plusieurs centaines de variables. L'objectif de ce papier est de développer une méthode permettant de sélectionner des variables spectrales permettant de discriminer des structures histopathologiques des tissus biologiques étudiés. Nous proposons un algorithme réalisant simultanément une classification non supervisé et une sélection de variable. Cette nouvelle méthode est basée sur un algorithme génétique couplée à un indice de validité. Cette méthode appliquée sur des images spectrales Raman de carcinomes cutanés permet: de trouver, de manière non-supervisée, les variables spectrales permettant de discriminer au mieux les classes tissulaires, de sélectionner des variables spectrales qui peuvent être corrélés des groupes chimiques fonctionnels spécifiques.

Mots clefs

Traitement de données 3D, Sélection de variable, classification non supervisé, Algorithme Génétique, Indice de validité, Alternative Silhouette Width Criterion, Imagerie spectrale Raman, Discrimination des structures histopathologiques.

1 Introduction

L'imagerie spectrale Raman (ISR) a récemment émergé comme un outil analytique particulièrement performant pour la caractérisation chimique des tissus biologiques avec des résolutions spatiales inférieures au µm [1]. Cette approche non destructive délivre des signatures spectrales contenant une information riche qui s'apparente à une véritable empreinte moléculaire de l'échantillon étudié.
Les images acquises sont mathématiquement définies comme des cubes de données composés de deux dimensions spatiales et d'une dimension spectrale (nombres d'onde). Chaque pixel d'une image représente un spectre Raman. Cependant, toutes les informations spectrales ne sont pas pertinentes pour la construction de biomarqueurs ou pour l'application de modèles de régression ou de classification. L'identification de ces nombres d'onde intéressants peut conduire à un meilleur traitement et une meilleure interprétation des données. Habituellement, cette étape est réalisée en utilisant des méthodes de sélection de variables supervisées [2]. Cependant, l'un des défis actuels dans l'application de l'imagerie Raman en histopathologie spectrale est l'identification non supervisée de marqueurs spectraux spécifiques des structures histopathologiques des tissus biologiques étudiés.

Pour atteindre cet objectif, nous avons développé un nouvel algorithme réalisant simultanément une classification non-supervisée et une sélection de variable. Cette nouvelle méthode est basée sur un algorithme génétique (AG) couplé à un indice de validité (Alternative Silhouette Width Criterion ; ASWC) [3]. Nous avons ensuite appliqué notre algorithme sur des images spectrales Raman acquises sur des coupes tissulaires de carcinomes cutanés (CBC) nodulaires pour discriminer les structures histopathologiques du CBC nodulaire. Puis, nous avons comparé les résultats obtenus par AG à ceux d'une image analysée sur la coupe colorée à l'hématoxyline-éosine.

2 Prétraitement des spectres Raman

Les méthodes de prétraitements appliquées sur les spectres Raman sont des méthodes classiques largement utilisées dans la littérature.

2.1 Estimation de la ligne de base (LB) par la méthode de Lieber

Cette méthode itérative modélise la ligne de base par un polynôme P d'ordre d [4]. A chaque itération, les parties du spectre situées au-dessus du polynôme estimé par moindres carrées sont rognées de façon à éliminer successivement les pics Raman et ne conserver que la ligne de base. Une fois estimé, le polynôme P est soustrait du spectre brut x, pour donner le spectre corrigé :

$$\mathbf{x}_{LB} = \mathbf{x} - P$$

2.2 Normalisation par standard normal variate (SNV)

La normalisation (SNV) est l'une des méthodes de prétraitement les plus utilisées en spectroscopie infrarouge pour corriger des problèmes de dispersion de données spectrales. Le spectre corrigé par SNV est donné par la relation suivante :

$$\mathbf{x}_{SNV} = (\mathbf{x}_{LB} - \overline{\mathbf{x}_{LB}})/\sigma(\mathbf{x}_{LB})$$

où $\overline{\mathbf{x}_{LB}}$ est la valeur moyenne du spectre $\mathbf{x}_{LB}$ et $\sigma(\mathbf{x}_{LB})$ son écart type [4].

3 Sélection de variable

3.1 Algorithme Génétique

Les algorithmes génétiques sont des méthodes stochastiques basées sur une analogie avec des systèmes biologiques. Ils reposent sur un codage des variables en structures chromosomiques et prennent modèle sur les principes de l'évolution naturelle pour déterminer une solution optimale.
L'idée est de générer des populations de N solutions, chaque solution de la population étant représentée sous la forme d'un « chromosome ». Chaque chromosome est lui-même formé d'un nombre restreint, noté par la suite « L », de nombres d'ondes (bandes spectrales) sélectionnés et positionnés comme des « gènes » dans le chromosome. A chaque étape, l'algorithme évalue les chromosomes à travers une fonction fitness. Il conserve les chromosomes ayant les meilleures valeurs de fitness pour la génération suivante. Il combine également les meilleurs chromosomes dans l'étape de croisement, puis il fait subir des mutations aux chromosomes restants [5]. La Figure 1 montre les étapes de l'AG.

L'AG peut être appliquée sur la matrice de données formée de J spectres Raman $\mathbf{X} = \{\mathbf{x}_j(\mathbf{y})\}_{j=1}^{J} = \{\mathbf{x}_1(\mathbf{y}), \dots \mathbf{x}_j(\mathbf{y}), \dots \mathbf{x}_J(\mathbf{y})\}$, $\mathbf{x}_j(\mathbf{y}) \in \mathrm{IR}^N$.

Chaque spectre est enregistré sur le vecteur de nombre d'ondes $\mathbf{y} \in \mathrm{R}^N$. Ces spectres appartiennent à un ensemble de classes C = $\{c_1, \dots c_k, \dots, c_K\}$ avec K < J, où K désigne le nombre de classes.

Les étapes de notre algorithme sont les suivantes :

a) **Initialisation de la population** : Les N chromosomes sont générés aléatoirement pour former une population initiale P(0) = $\{\underline{z}_i = [z_{i1}, \dots, z_{il}, \dots, z_{iL}] \in \mathbb{R}^L\}$ telle que chaque gène z_{il} est un variable (nombre d'onde) choisi aléatoirement dans le vecteur $\mathbf{y}$. Chaque $\underline{z}_i$ est donc un vecteur formé de L (taille de chromosome) nombres d'ondes aléatoirement sélectionnés dans le vecteur $\mathbf{y}$

b) **Évaluation** : la performance de chaque chromosome est évaluée dans la population initiale suivant la valeur de leur fonction fitness. Cette fonction est la mesure de qualité du chromosome. Comme nous essayons de classer les spectres dans K classes inconnues et de façon non supervisée, nous avons testé différentes fonctions fitness en s'appuyant sur des indices de validité qui visent à avoir les clusters les plus compacts et les plus séparés. Ces indices sont des fonctions statistiques bien connues et largement utilisées pour évaluer et mesurer la qualité des classes obtenues, comme : Davies Bouldin (DB), Calinski-Harabasz (CH), Xie Beni (XB), Silhouette (SIL), Alternative Silhouette Width Criterion (ASWC) [3, 6]. Nous avons trouvé que l'indice ASWC est le plus adapté pour la sélection de variables par AG appliquée à l'imagerie Raman en histopathologie spectrale. Cet indice de validité a été appliquée après l'application de méthode de classification non supervisée K means.

Alternative Silhouette Width Criterion (ASWC):
Ce critère mesure la similarité de chaque spectre par rapport aux autres spectres de son cluster en comparaison avec les spectres dans les autres clusters.

$$F(\underline{z}_i) = ASWC(\underline{z}_i) = \frac{1}{J}\sum_{j=1}^{J} s_j = \frac{1}{J}\sum_{j=1}^{J} \frac{b_j}{a_j + \varepsilon}$$

où

$$a_j = \frac{1}{card(C_k) - 1} \sum_{j'=1}^{card(C_k)} \|\mathbf{x}_j^{C_k}(\underline{z}_i) - \mathbf{x}_{j'}^{C_k}(\underline{z}_i)\|^2$$

est la distance moyenne du spectre $\mathbf{x}_j^{C_k}$ par rapport à tous les autres spectres appartenant à la même classe C_k

$$b_j = \min_{k', k \neq k'} \frac{1}{card(C_{k'})} \sum_{j'=1}^{card(C_{k'})} \|\mathbf{x}_j^{C_k}(\underline{z}_i) - \mathbf{x}_{j'}^{C_{k'}}(\underline{z}_i)\|^2$$

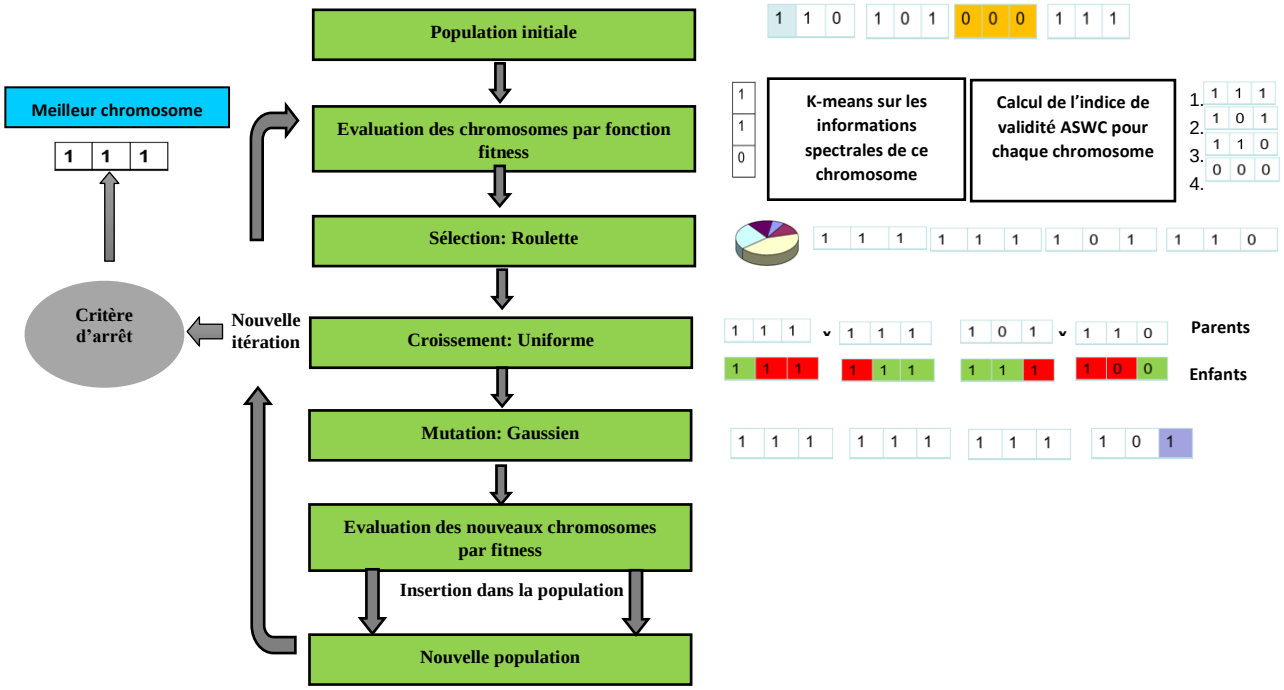


Figure 1 : *Représentation synoptique de la méthodologie de l'AG proposée*

est la distance moyenne du spectre $\mathbf{x}_j^{C_k}(\underline{z}_i)$ appartenant à la classe C_k par rapport à tous les autres spectres $\mathbf{x}_j^{C_{k'}}(\underline{z}_i)$ appartenant à une autre classe $C_{k'}$.

c) **Sélection** : Le rôle de l'opérateur de sélection est de filtrer la population de manière à conserver les chromosomes possédant de « bonnes » caractéristiques génétiques. Cette étape est utilisée pour choisir les nouveaux parents (ceux qui ont les meilleurs résultats de fitness). Nous avons choisi la sélection de type « roulette » car cette méthode a comme avantage de ne pas avoir de biais d'estimation, et une dispersion minimale.

d) **Croisement**: Cette étape est utilisée pour recombiner les meilleurs chromosomes pour obtenir des enfants potentiellement supérieurs. Nous avons choisi la méthode de croisement uniforme qui a donné de bons résultats dans la majorité des cas [7].

e) **Mutation** : L'opérateur de mutation a pour rôle d'assurer la diversité des solutions pour sortir des minima locaux. Elle consiste à modifier un ou plusieurs gènes d'un individu sélectionné par l'étape de sélection. Nous avons choisi l'opérateur de mutation Gaussien car il produit les meilleurs résultats pour la plupart des fonctions fitness [8].

f) On répète les étapes (b) à (e) jusqu'à ce que le nombre d'itérations maximal soit atteint.

3.2 Paramètres de l'algorithme génétique

Pour choisir la taille des chromosomes L, il n'existe pas de méthode clairement définie. Nous sommes donc obligés de déterminer de façon itérative les valeurs optimales de ces paramètres. Pour cela nous itérons pour différentes valeurs de tailles des chromosomes (L=10 jusqu'à 100. Nous choisissons ensuite la valeur de L optimale à partir de meilleur valeur moyen de fitness obtenu [9] :

$$L_{\text{optimale}} = \min_L \left\{ \overline{F(\underline{z}_i)} \right\}, i = 1 \dots L$$

4 Application

4.1 Échantillons et paramètres

Nous disposons d'images spectrales Raman acquises sur des coupes tissulaires de carcinomes cutanés (CBC) nodulaires contenant plusieurs nodules. La figure 2 met en évidence les structures existantes au sein des tissus composés de nodules tumoraux et de derme. Les spectres de chaque image ont été corrigés numériquement de la ligne de base par un polynôme d'ordre 5 et ont ensuite été normalisées par SNV. Nous avons choisi la gamme spectrale informative de 500 à 1800 cm^{-1} appelée "fingerprint" qui renseigne sur les vibrations principales des groupes chimiques spécifiques de l'échantillon biologique.

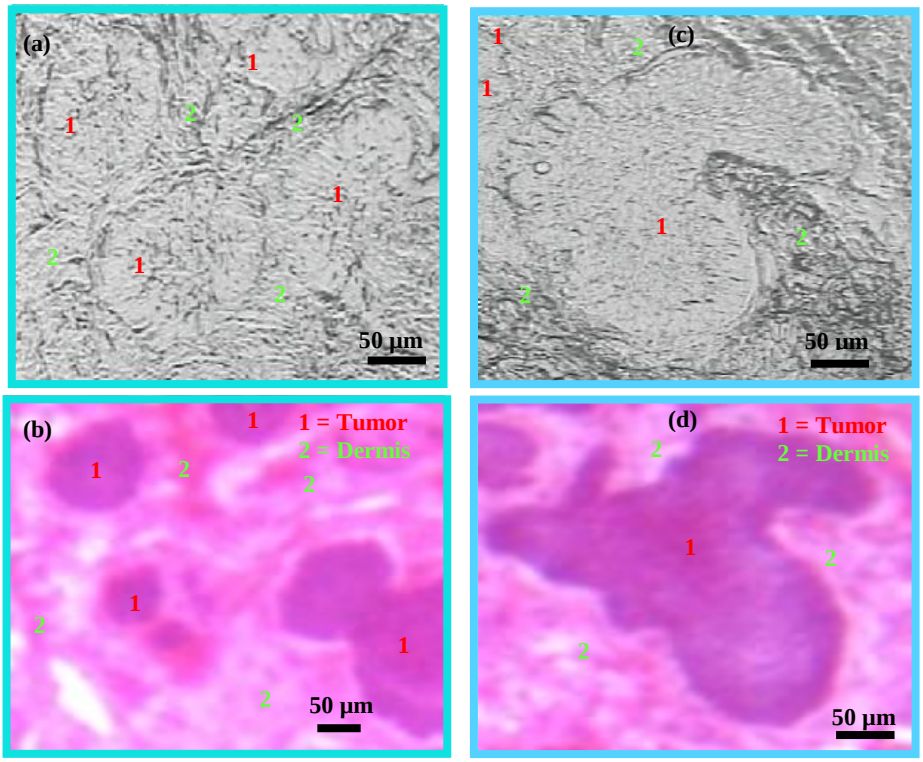


Figure 2 : *(a) (c) Zone analysée par imagerie Raman sur les deux coupes différentes non colorées ; (b) (d) Même zone sur des coupes adjacentes colorées à l'hématoxyline-éosine*

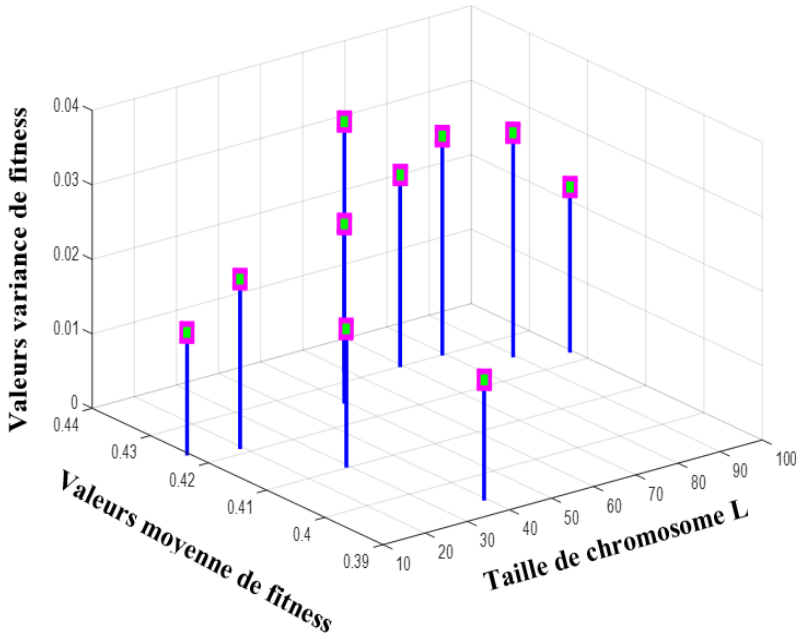


Figure 3 : *Valeurs moyennes et variances de la fonction fitness de l'AG pour différentes tailles de chromosomes L*

4.2 Résultats

La Figure 3 montre les valeurs moyennes et les variances de la fonction fitness (ASCW) pour différentes valeurs de L d'AG appliquée sur la première coupe tissulaire de carcinomes cutanés (CBC). D'après cette figure, nous avons trouvé que la taille de chromosomes L = 40 donne la plus petite valeur de moyenne de la fonction fitness. De plus, nous avons obtenu des résultats similaires appliqués sur la deuxième coupe tissulaire (CBC).

Les Figures 4 (a) et 5 (a) montrent les résultats obtenus après l'application de notre méthodologie d'AG des images spectrales Raman contenant plusieurs nodules de CBC.

La classification effectuée par notre méthodologie d'AG permet de bien identifier les nodules (Tumor) isolés présents sur ces images, et de retrouver les structures très hétérogènes du derme (Dermis).

Notre méthodologie d'AG avec la fonction fitness ASWC permet l'identification de nombres d'ondes (Figure 4 (a) et Figure 5 (a) associés à des groupes chimiques fonctionnels spécifiques de la structure des échantillons étudiés, et donc mener à la définition de marqueurs spectraux, notamment tumoraux.

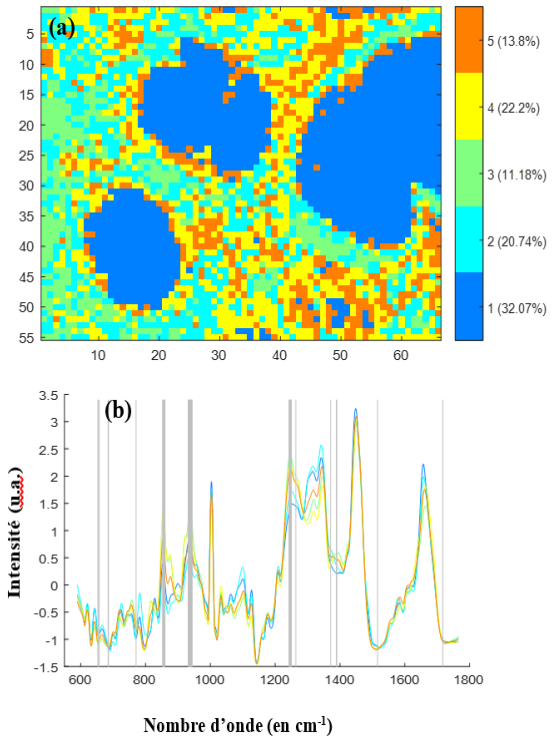


Figure 4 : *Image pseudo-colorée obtenue après l'application AG en sélectionnant 40 variables spectrales sur la première coupe tissulaire de carcinomes cutanés, (b) les 5 centroïdes et les 40 variables sélectionnées par l'AG*

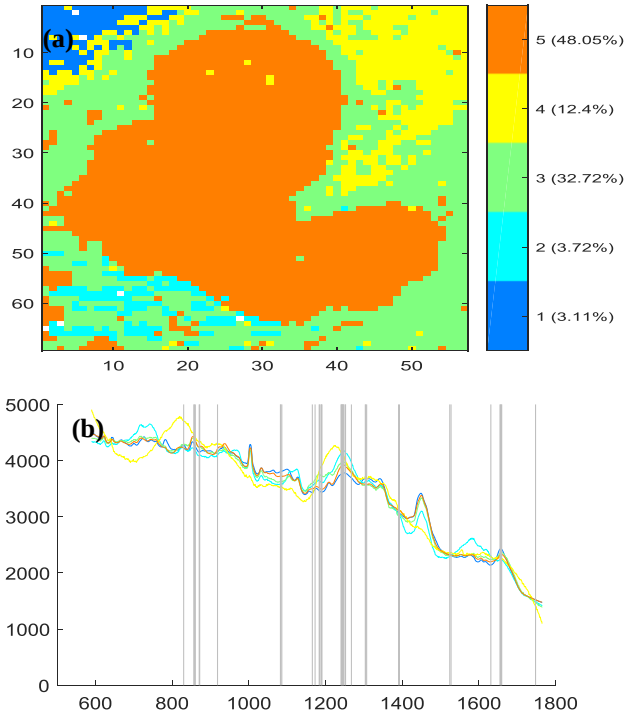


Figure 5 : (a) *Image pseudo-colorée obtenue après l'application AG en sélectionnant 40 variables spectrales sur la deuxième coupe tissulaire de carcinomes cutanés, (b) les 5 centroïdes et les 40 variables sélectionnées par l'AG*

5 Conclusion

Notre méthodologie d'AG, en choisissant l'indice ASWC comme fonction fitness et le choix de paramètres adaptés, est capable de diviser les données spectrales en classes tissulaires et de trouver, de manière non-supervisée, les variables spectrales (nombres d'onde) permettant de discriminer au mieux ces classes tissulaires. Ces variables spectrales (nombres d'onde) intéressantes peuvent être corrélées à des groupes chimiques fonctionnels spécifiques des échantillons étudiés, et donc mener à la définition de marqueurs spectraux, notamment tumoraux

Références

[1] K. Kong, C. Kendall, N. Stone, I. Notingher. Raman spectroscopy for medical diagnostics-From in-vitro biofluid assays to in-vivo cancer detection. *Advanced Drug Delivery Reviews*, 89: 121–134, 2015.

[2] Y. Saeys, I. Inza, P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics Rev.*, 23: 2507-2517, 2007.

[3] L. Vendramin, R. J. G. B. Campello, E. R. Hruschka. On the Comparison of Relative Clustering Validity Criteria, In: 2009 *SIAM International Conference on Data Mining, Sparks*, pages 733-744, Nevada, 2009.
[4] C. A. Lieber, A. Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied spectroscopy*, 57: 1363-1367, 2003.
[5] J. Holland. Adaptation in natural and artificial systems. University of Michigan Press, 1975.
[6] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M.Prez, I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46: 243-256, 2013.
[7] S. Picek. Comparison of a Crossover Operator in Binary-coded Genetic Algorithms, *WSEAS Trans. on Computers*, 9: 1064-1073, 2010.
[8] R. Hinterding. Gaussian Mutation and Self-adaption for Numeric Genetic Algorithms. *IEEE*, 384-389, 1995.
[9] S. Ullah, T. A. Groen, M. Schlerf, A. K. Skidmore, W. Nieuwenhuis, C. Vaiphasa. Using a Genetic Algorithm as an Optimal Band Selector in the mid and Thermal Infrared (2.5–14 μm) to Discriminate Vegetation Species, *Sensors*, 12: 8755-8769, 2012

Filtre multi-sorties pour la réduction de bruit appliqué aux images médicales ultrasons

Meriem OUTTAS[1,2], Lu ZHANG[1], Olivier DEFORGES[1], Amina SERIR[2] Wassim HAMIDOUCHE[1]

[1]Institut d'Electronique et de Télécommunications de Rennes
INSA Rennes, 20 avenue des buttes de Coësmes, 35043 Rennes Cédex, FRANCE

[2]Laboratoire Traitement d'Image et Rayonnement
USTHB, BP 32 EL ALIA, BAB EZZOUAR ALGER, ALGERIE, 16111

mouttas@insa-rennes.fr, lu.ge@insa-rennes.fr, odeforge@insa-rennes.fr
aserir@usthb.dz, wassim.hamidouche@insa-rennes.fr

Résumé – L'échographie est la modalité d'imagerie la plus répandue car la moins nocive. Cependant, la présence de bruit de « speckle » dans l'imagerie médicale ultrasonographique reste un inconvénient majeur. La principale limite des méthodes de réduction de bruits proposées est qu'une réduction efficace du bruit s'effectue au détriment de la qualité (lissage excessif, effet de flou, un aspect artificiel, ...). Dans cet article, nous proposons un nouveau filtre Multi-Sorties basé sur la Décomposition Multiplicative Multirésolutions (MOF-MMD). Cette méthode multi-échelles, particulièrement efficace dans le cas de bruit multiplicatif, améliore distinctement trois sorties: bords, texture, et image globale. Le filtre multi-sorties vise à offrir aux praticiens des images améliorées en fonction des besoins de diagnostic. Les différentes structures, textures et bordures sont filtrées selon l'image contour obtenue par des opérateurs de morphologie mathématique. Les images résultantes du filtrage sont évaluées et validées par comparaison avec deux méthodes de l'état de l'art.

Abstract – Ultrasonographic examination, either as visual inspection or quantitative analysis, is less effective than other medical imaging systems due to speckle noise. In this paper, a new Multi-Output Filter based on a Multiplicative Multiresolution Decomposition (MOF-MMD) is proposed. This multiscale based method, particularly efficient in the case of multiplicative noise, enhances distinctively three outputs: edges, texture and the global image. Ultrasound is the most widespread imaging modality. However, it suffers from a major disadvantage as it is corrupted by speckle noise. The state-of-the-art speckle reduction methods often offers an effective speckle reduction at the expense of the quality (oversmoothig, blurring effect, artificial appearance...). In this paper, a new Multi-Output Filter based on a Multiplicative Multiresolution Decomposition (MOF-MMD) is proposed. This multiscale based method, particularly efficient in the case of multiplicative noise, enhances distinctively three outputs: edges, texture and the global image. The multi-output filter aims at offering an enhanced images according to the features desired by radiologists. The different structures, textures and edges are filtered according to the contour image obtained by morphological operators.The resulting images are evaluated and validated by comparison with two state-of-the-art methods.

1 Introduction

Depuis plusieurs décennies, l'échographie est utilisée pour visualiser en temps réel les différentes structures internes du corps humain. Cette technique non invasive permet d'explorer le cœur, les organes digestifs (foie, rate, pancréas, vésicule biliaire), urinaires (vessie, reins) et génitaux (prostate et testicules, ovaires et utérus). L'échographie permet aussi de guider avec précision le praticien lors d'interventions chirurgicales, on parle alors d'interventions écho guidées,

L'examen ultrasonographique est réalisé de deux façons : une évaluation visuelle qualitative basée sur l'interprétation et l'expérience du clinicien, et une analyse quantitative par mesures de grandeurs ou de biomarqueurs qui aident au diagnostic. En cardiologie la mesure du « Strain » et du « Strain Rate » est un outil de quantification de la fonction et de la contractilité myocardique [1]. En gastro-entérologie, la texture du parenchyme hépatique est une caractéristique subjective pour la détection de la cirrhose du foie [2]. Cependant, cette modalité d'imagerie présente un inconvénient : la qualité de l'image est dégradée du fait de la présence du bruit « speckle ». Ce bruit donne un aspect granuleux à l'image échographique. Il existe beaucoup de recherches visant à réduire le speckle et à améliorer la qualité de l'image échographique. Nombre de ces techniques, efficaces en termes de débruitage, présentent une limitation majeure en terme de qualité : lissage excessif de texture ou encore pertes de détails importants au diagnostic [7]. Dans cet article, nous proposons un filtre multi-sorties basé sur la Décomposition Multiplicative Multiresolution (MOF-MMD), en se basant sur le fait que l'interprétation d'une échographie s'effectue suivant : l'évaluation de l'aspect général, l'examen des contours et l'analyse de la texture. Dans la suite nous détaillerons la méthodologie mise en œuvre, présenterons les expérimantations et les résultats, et terminerons par une conclusion.

2 Méthodologie

Dans cette section, nous détaillons les différentes étapes de notre méthode.

2.1 La Décomposition Multiplicative Multirésolution (MMD)

La décomposition Multiplicative Multirésolution (MMD) est une décomposition multi-échelle non linéaire [3]. La MMD est basée sur l'utilisation de bancs de filtres non linéaires multiplicatifs avec un sous échantillonnage critique et une reconstruction parfaite. Cette méthode a l'avantage de reconstruire parfaitement le signal 2D "l'image" à partir d'une décomposition multiplicative. Cette dernière caractéristique fait de la MMD une méthode adaptée à l'analyse des images distordues par du bruit multiplicatif. Les figures 1 et 2 illustrent la Décomposition Multiplicative Multirésolution (MMD) d'une image $I(n,m)$.

La MMD suppose l'analyse et synthèse des bancs de filtres en termes de systèmes à quatre entrées-sorties avec des taux d'entrée et de sortie égaux. La structure voulue est obtenue en réalisant la décomposition polyphase de l'image. Les quatre composantes polyphases x_{11}, x_{12}, x_{21} et x_{22} de l'image d'entrée $I\left(n,m\right)$ sont définis par :

$$x_{ij}\left(n,m\right) = I\left(2\left(n-1\right)+i, 2\left(m-1\right)+j\right) \quad i,j \in \{1,2\} \tag{1}$$

Ce système requiert deux paires de filtres d'analyses et de synthèses $\left(\{h_{i,j}\}, D\right)$ et $\left(\{f_{i,j}\}, R\right)$, respectivement. La réponse impulsionnelle des filtres d'analyses $\{h_{i,j}\}$ et de synthèses $\{f_{i,j}\}$ doit satisfaire les conditions suivantes :

$$f_{i,j}\left(k,l\right) = {h_{i,j}}^{-1}\left(k,l\right), \quad i,j \in \{1,2\} \tag{2}$$

$h_{12} = \alpha h_{11}, h_{21} = \nu h_{11}, h_{22} = \gamma h_{11}$

Où α, ν et γ sont des scalaires positifs.

Ainsi, les filtres polyphases linéaires $h_{i,j}$ et $f_{i,j}$ sont définis par :

$$h_{i,j}\left(k,l\right) = h\left(2\left(k-1\right)+i, 2\left(l-1\right)+j\right) \quad i,j \in \{1,2\} \tag{3}$$

$$f_{i,j}\left(k,l\right) = f\left(2\left(k-1\right)+i, 2\left(l-1\right)+j\right) \quad i,j \in \{1,2\} \tag{4}$$

Les filtres non linéaires d'analyse et de synthèse D et R, illustrés dans les figures 1 et 2, sont définis par leurs sorties y_{2v}, y_{2h} et y_2d comme suit :

$$y_{2v} = \begin{cases} \beta\frac{x_{12}}{x_{11}}, & x_{11} \geq x_{12} \\ \beta\left(2-\frac{x_{11}}{x_{12}}\right), & x_{11} < x_{12} \\ \alpha, & x_{11} = x_{12} = 0 \end{cases} \tag{5}$$

$$y_{2h} = \begin{cases} \beta\frac{x_{21}}{x_{11}}, & x_{11} \geq x_{21} \\ \beta\left(2-\frac{x_{11}}{x_{21}}\right), & x_{11} < x_{21} \\ \nu, & x_{11} = x_{12} = 0 \end{cases} \tag{6}$$

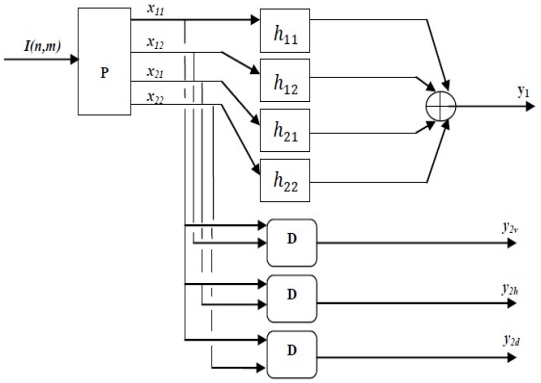


FIGURE 1 – Schéma d'Analyse de la décomposition multiplicative multirésolution MMD

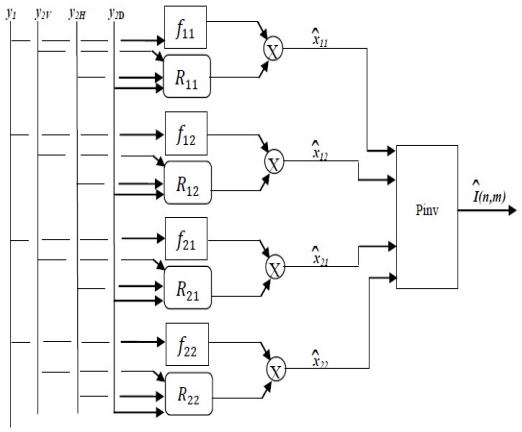


FIGURE 2 – Schéma de synthèse de la décomposition multiplicative multirésolution MMD

$$y_{2d} = \begin{cases} \beta\frac{x_{22}}{x_{11}}, & x_{11} \geq x_{22} \\ \beta\left(2-\frac{x_{11}}{x_{22}}\right), & x_{11} < x_{22} \\ \gamma, & x_{11} = x_{22} = 0 \end{cases} \tag{7}$$

où β est un scalaire positif.

Les réponses des filtres de synthèses non linéaires r_{ij}, permettant d'assurer une reconstruction parfaite du signal, sont exprimées comme suit :

$$r_{11}\left(y_{2h}, y_{2v}, y_{2d}\right) = \frac{1}{1+\alpha\frac{x_{12}}{x_{11}}+\mu\frac{x_{21}}{x_{11}}+\gamma\frac{x_{22}}{x_{11}}} \tag{8}$$

$$r_{12}\left(y_{2h}, y_{2v}, y_{2d}\right) = \alpha\frac{x_{12}}{x_{11}} r_{11}\left(y_{2h}, y_{2v}, y_{2d}\right) \tag{9}$$

$$r_{21}\left(y_{2h}, y_{2v}, y_{2d}\right) = \mu\frac{x_{21}}{x_{11}} r_{11}\left(y_{2h}, y_{2v}, y_{2d}\right) \tag{10}$$

$$r_{22}\left(y_{2h}, y_{2v}, y_{2d}\right) = \gamma\frac{x_{22}}{x_{11}} r_{11}\left(y_{2h}, y_{2v}, y_{2d}\right) \tag{11}$$

Pour $\beta = 0.5$, y_{2h}, y_{2v}, y_{2d} varient sur un intervale $[0;1]$. Il est à noter que les valeurs proches de β correspondent à des

régions homogènes de l'image à l'inverse des valeurs loins de β correspondent à des détails plus contrastés.

Pour réaliser une représentation multirésolution, plusieurs décompositions en sous bandes sont mises en cascade. Ainsi, la sous bande y_1 (figure 1) est re-décomposée. A la résolution J, le signal original est représenté par S défini par :

$$S = \left({y_1}^{(j)}, \left({y_{2h}}^{(j)}, {y_{2v}}^{(j)}, {y_{2d}}^{(j)} \right) \right)_{2 \leq j \leq J} \quad (12)$$

Inversement, l'approximation du signal reconstruit à la résolution $j = 1$ est obtenue en utilisant la synthèse multirésolution en sous bandes et la représentation par l'ensemble des signaux S. La méthode entièrement détaillée est présentée dans [3].

Dans la suite nous noterons D_C tous les détails y_{2H}, y_{2V}, y_{2D}.

2.2 Pseudo-segmentation de l'image par opérateurs morphologiques

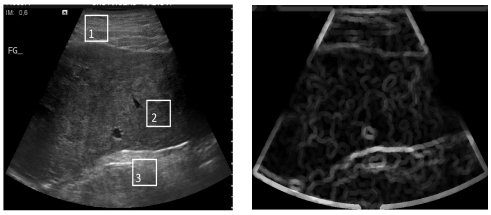
Le but de ce travail étant de réduire le bruit speckle dans les images échographiques tout en préservant les détails et les contours, il est important de délimiter les structures importantes constituants l'image. Pour ce faire, nous proposons une pseudo-segmentation des différentes régions qui composent l'image basée sur des opérateurs de morphologie mathématique. Les transformations morphologiques requièrt l'utilisation d'un élément structurant, caractérisé par sa forme et sa taille. Etant donné que les images médicales contiennent principalement des structures de forme ovale, l'élément structurant C choisi est de forme circulaire, de rayon égale à huit pixels fixé de façon expérimentale.

Soit I l'image originale et C l'élément structurant. La pseudo-segmentation est obtenue comme suit :

$$I_{(oc)} = (I \circ C) \bullet C \quad (13)$$

$$S = (I_{(oc)} \oplus C) - (I_{(oc)} \ominus C) \quad (14)$$

où $\oplus$, $\ominus$, $\circ$ et $\bullet$ sont respectivement les opérateurs de dilatation, d'erosion, d'ouverture et de fermeture. S est l'image contour obtenue représentée Figure. 3(b).



(a) Image originale (b) Image contour

FIGURE 3 – Pseudo-segmentation de l'image

2.3 Sélection de W et calcul de C_n

Le filtre multi-échelle proposé effectue un seuillage de chaque coefficient D_C et à toutes les échelles J. Le seuil T est calculé à partir du niveau de bruit C_n, qui lui est calculé à partir de la variance et la moyenne d'une fenêtre locale W d'intensité constante sélectionnée par l'utilisateur (Figure.3(a)) $C_n^2 = \frac{var(W)}{Moyenne(W)}$.

Afin d'assurer la sélection du meilleur C_n et par conséquent le meilleur seuil T, un nombre N de la fenêtre W sont sélectionnés dans ce travail. La Figure. 3(a) montre 3 fenêtres sélectionnées ($N = 3$), pour le calcul de C_n et T.

2.4 Seuillage : filtre Multi-Sorties

Comme indiqué section 2.1, les coefficients de la MMD sont proches de $\beta = 0.5$ dans les régions homogènes. Cette propriété contribue à filtrer le bruit présent dans l'image tout en préservant les détails structurels et en évitant l'effet de flou. Afin d'obtenir le filtrage multi-sorties désiré, le seuillage est effectué en fonction de l'intensité des pixels de l'image contour S obtenue (cf.section2.2). Le pixel considéré est filtré selon qu'il appartienne à un contour ou à une texture.

2.4.1 Sortie 1 du filtre - rehaussement des contours

La première sortie de notre filtre donne une image dont les contours, principalement, ont été rehaussés. Cette image permet une mesure plus précise des lésions et des distances. Pour chaque pixel de l'image contour S, un seuillage est appliqué comme suit :

$$\hat{D}_C^{(j)} = \begin{cases} \beta \quad si \quad \beta - \beta_t \leq D_C^{(j)} \leq \beta + \beta_t \\ D_C^{(j)} \times \nu - (\gamma \times \alpha \times S) \quad si \quad D_C^{(j)} + \tau \leq \beta \\ D_C^{(j)} \times \nu + (\gamma \times \alpha \times S) \quad si \quad D_C^{(j)} - \tau \geq \beta \\ \beta \quad sinon \end{cases} \quad (15)$$

où $D_C^{(j)}$ et $\hat{D}_C^{(j)}$ représentent respectivement les coefficients MMD à la résolution j de l'image originale bruitée I, et leurs correspondants seuillés. β_t est un seuil qui vise à réduire le bruit dans les pixels les plus homogènes et est fixé à 0.0016. $\tau = (T \times \nu \times \alpha \times S)$, avec $T = C_n \times j/J$ est le seuil calculé à partir du niveau de bruit pour chaque résolution j avec J le nombre total de résolutions ($J = 4$). α est choisi à 0.25 . ν et γ sont donnés par $\nu = \frac{1}{\sqrt{1+{C_n}^2}}$ et $\gamma = 1 - \frac{1}{\sqrt{1+{C_n}^2}}$. Notons que le seuillage est proportionnel à l'image contour *i.e.* plus le contour est important, plus le rehaussement l'est également.

2.4.2 Sortie 2 du filtre - rehaussement de la texture

La seconde sortie du filtre donne une image dont la texture est rehaussée, le bruit est réduit tout en préservant le motif de texture ainsi que sa granularité. Soit $\bar{S}$ l'image des pixels qui n'appartiennent pas au contour S. La seconde sortie résulte du seuillage suivant :

$$\hat{D}_C^{(j)} = \begin{cases} \beta \quad si \quad \beta - \beta_t \leq D_C^{(j)} \leq \beta + \beta_t \\ D_C^{(j)} \times \nu - \left(\gamma \times \bar{S}\right) \quad si \quad D_C^{(j)} + \xi \leq \beta \\ D_C^{(j)} \times \nu + \left(\gamma \times \bar{S}\right) \quad si \quad D_C^{(j)} - \xi \geq \beta \\ \beta \quad sinon \end{cases} \quad (16)$$

avec $\xi = \left(T \times \nu \times \bar{S}\right)$,

2.4.3 Sortie 3 du filtre- Rehaussement globale de l'image

La troisième sortie du filtre donne une image avec un rehaussement global. Elle offre un compromis entre les deux sorties introduites précédemment, et est basée sur l'image complément de l'image contour Co_S. Les coefficients MMD sont seuillés selon le schéma de l'image Co_S, comme suit :

$$\hat{D}_C^{(j)} = \begin{cases} \beta \quad si \quad \beta - \beta_t \leq D_C^{(j)} \leq \beta + \beta_t \\ D_C^{(j)} \times \nu - (\gamma \times Co_S) \quad si \quad D_C^{(j)} - (\gamma \times Co_S) \leq \beta \\ D_C^{(j)} \times \nu + (\gamma \times Co_S) \quad si \quad D_C^{(j)} + (\gamma \times Co_S) \geq \beta \\ \beta \quad sinon \end{cases} \tag{17}$$

2.5 Reconstruction MMD et sélection des sorties basée sur la qualité

Les trois résultats obtenus sont reconstruits par recomposition MMD en trois images rehaussées $\hat{I}_1, \hat{I}_2, \hat{I}_3$.

Dans le cas de $N \times C_n$, nous obtiendrons $N \times \left\{\hat{I}_1, \hat{I}_2, \hat{I}_3\right\}$. Les trois images finales sont choisies en fonction du critère de mesure de la qualité NIQE et NIQE-K. Ces métriques sont présentées dans la sous-section 3-3.2. Pour l'image représentée dans la Figure. 3(a), la première fenêtre est sélectionnée pour le calcul du niveau de bruit C_n.

Les étapes du filtre multi-sorties sont résumées comme suit :
• Décomposer l'image original(x,y) en MMD
• Effectuer la pseudo-segmentation de l'image original(x,y)
• Sélectionner N fenêtre W par l'utilisateur
• Calculer C_n pour les N fenêtre
• Seuiller les coefficients MMD suivant les C_n calcmulés : Contours $\hat{I}_1$ - Texture $\hat{I}_2$ - Global $\hat{I}_3$
• Reconstruire les images en MMD $N \times \left\{\hat{I}_1, \hat{I}_2, \hat{I}_3\right\}$
• Selectionner les images $\left\{\hat{I}_1, \hat{I}_2, \hat{I}_3\right\}$ en se basant sur les métriques de qualité NIQE et NIQE-K

3 Expérimentation et résultats

Dans cette section nous comparons le filtre proposé avec deux méthodes récentes et efficaces de réduction de speckle : Filtre Bayesian Optimisé à moyenne-non-local avec sélection de blocs(OBNLM) [5] et Filtre de diffusion anisotrope avec mémoire basée sur les statistiques de Speckle(ADMSS) [6]. Cette comparaison est effectuée en termes de capacité de réduction de speckle et en terme d'amélioration de la qualité.

3.1 Images expérimentales

Afin d'évaluer notre méthode de filtrage, nous avons utilisé 21 images d'échographies abdominales de foies sains et malades, à partir d'une base de données rétrospective du CHU d'Angers.

3.2 Métriques d'évaluation

L'évaluation de la capacité de réduction de speckle peut être quantifiée par le calcul du SSNR (speckle's signal-to-noise ratio) [10]. Par ailleurs, la qualité des images filtrées est évaluée en utilisant trois métriques de qualité sans référence. Dédiées aux images naturelles, la NIQE et la BIQES ont été utilisées pour évaluer des images médicales dans [7] et [11]. Par ailleurs, nous avons proposé dans [11] une métrique baptisée NIQE-K, inspirée des métriques pour images naturelles, elle est mieux adaptée aux images médicales.

3.3 Résultats et discussion

Les « boxplot » représentées en Figure. 4 offrent une comparaison objective des performances des méthodes de réduction de speckle. Le SSNR (Figure. 4) indique que la méthode proposée ainsi que les deux méthodes de l'état de l'art réduisent le bruit, avec un léger avantage pour le OBNLM. Néanmoins en terme de qualité de l'image les trois métriques indiquent que le filtre OBNLM donne des images de mauvaise qualité ce qui est dû au lissage excessif. Concernant les trois sorties de la méthode proposée, la NIQE, NIQE-K et BIQES indiquent une amélioration significative de la qualité de l'image filtrée comparativement à celle obtenue par OBNLM et ADMSS. De plus, les écarts-types des boites représentant les trois sorties du filtre proposé sont inférieurs à ceux correspondant à OBNLM et ADMSS ce qui indique une moindre dispersion.

Les images filtrées obtenues par les différentes méthodes sont présentées Figure. 5. Nous pouvons constater que la sortie 1 de MOF-MMD améliore l'aspect des vaisseaux hépatique (la ligne hyper-échogène Figure. 5(d)) ; la sortie 2 du MOF-MMD fait ressortir la texture tout en réduisant le speckle (voir Figure. 5(e)). La sortie 3 du MOF-MMD offre une amélioration globale de l'image (voir Figure. 5(f)). L'image filtrée par la méthode OBNLM révèle un lissage excessif de la texture, la granularité de l'image, élément essentiel au diagnostic dans le cas du foie, est complètement supprimée. Le filtre ADMSS quant à lui lisse l'image et préserve les bords de manière satisfaisante, cependant nous pouvons constaté la suppression de détails structurels, considéré comme des régions sans intérêt par sa fonction mémoire. Le MOF-MMD réduit considérablement le speckle tout en améliorant la qualité des images traitées.

4 Conclusion

Dans cet article nous proposons une nouvelle méthode de réduction de bruit type speckle dans les images médicales échographiques. Cette approche basée sur une décomposition multi résolution (MMD) vise à fournir trois images filtrées en sorties, adaptées aux besoins des praticiens : bords accentués, texture rehaussée ou encore aspect globale amélioré. Les résultats expérimentaux montrent bien l'amélioration de la qualité globale de l'image, associée à une réduction efficace du speckle. Afin

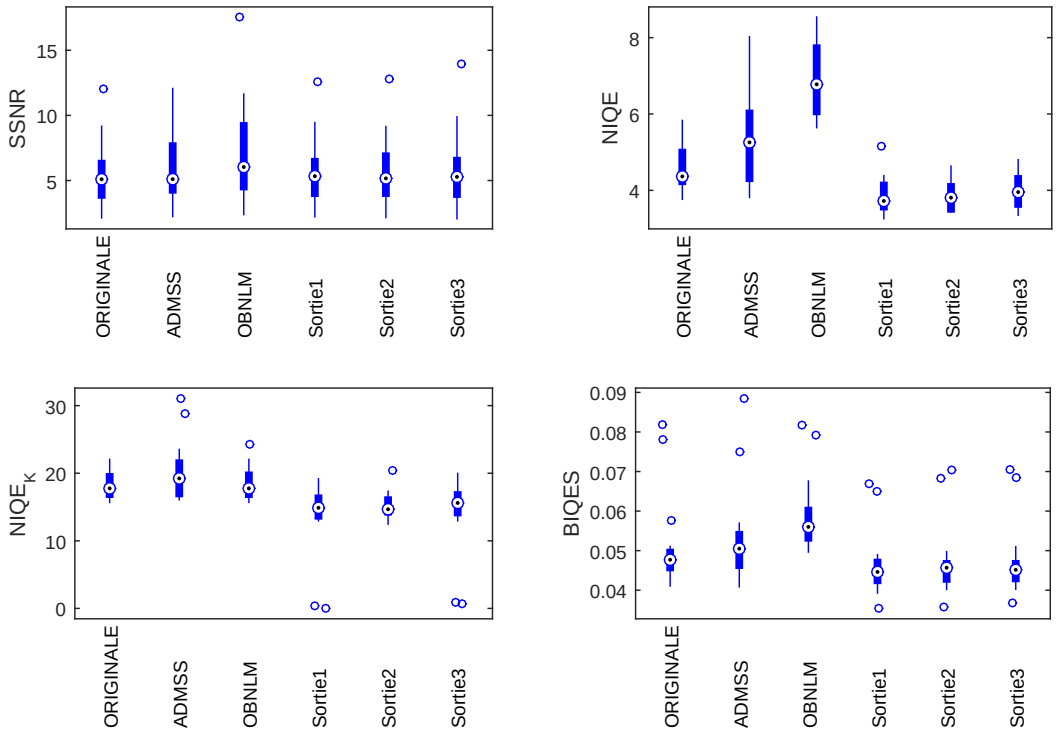


FIGURE 4 – Représentation de SSNR, NIQE, BIQES et NIQE-K en boxplot pour ADMSS, OBNLM, Sortie 1, 2 et 3 de la méthode proposée(Une valeur plus grande du SSNR indique moins de speckle. Une valeur plus grande de NIQE, BIQES et NIQE-K indique une moins bonne qualité)

de valider notre méthode des tests subjectifs d'évaluation de la qualité sont en cours avec l'aide de praticiens médicaux.

Références

[1] N. Haddour, C Meuleman, G. Dufaitre, S. Janower, E. Berthelot-Garcias, F. Douna, S. Ederhy, F. Boccara et A. Cohen, *Qu'est-ce que l' échocardiographie 2D strain*, Revues Générales-Echocardiographie, 2011.

[2] J. F. Gerstenmaier, and R. N. Gibson, *Ultrasound in chronic liver disease*, Insights into imaging, 2014.

[3] A. Serir, and A. Belouchrani, *Multiplicative multiresolution decomposition for 2D signals : application to speckle reduction in SAR images*, International Conference on Image Processing, ICIP 2004.

[4] M. Outtas, A. Serir, O. Deforges and L. Zhang, *Réduction de bruit multiplicatif dans les images ultrasons basée sur la Décomposition Multiplicative Multiresolution (MMD)*, CORESA 2016

[5] P. Coupé, P. Hellier, C. Kervrann, and C. Barillot, *Nonlocal Means-Based Speckle Filtering for Ultrasound Images*, IEEE Transactions on Image Processing, 2009

[6] G. Ramos-Llordén, G. Vegas-Sánchez-Ferrero, M. Martin-Fernandez, C. Alberola-López, and S. Aja-Fernández, *Anisotropic Diffusion Filter with Memory based on Speckle Statistics for Ultrasound Image*, IEEE Transactions on Image Processing, 2015

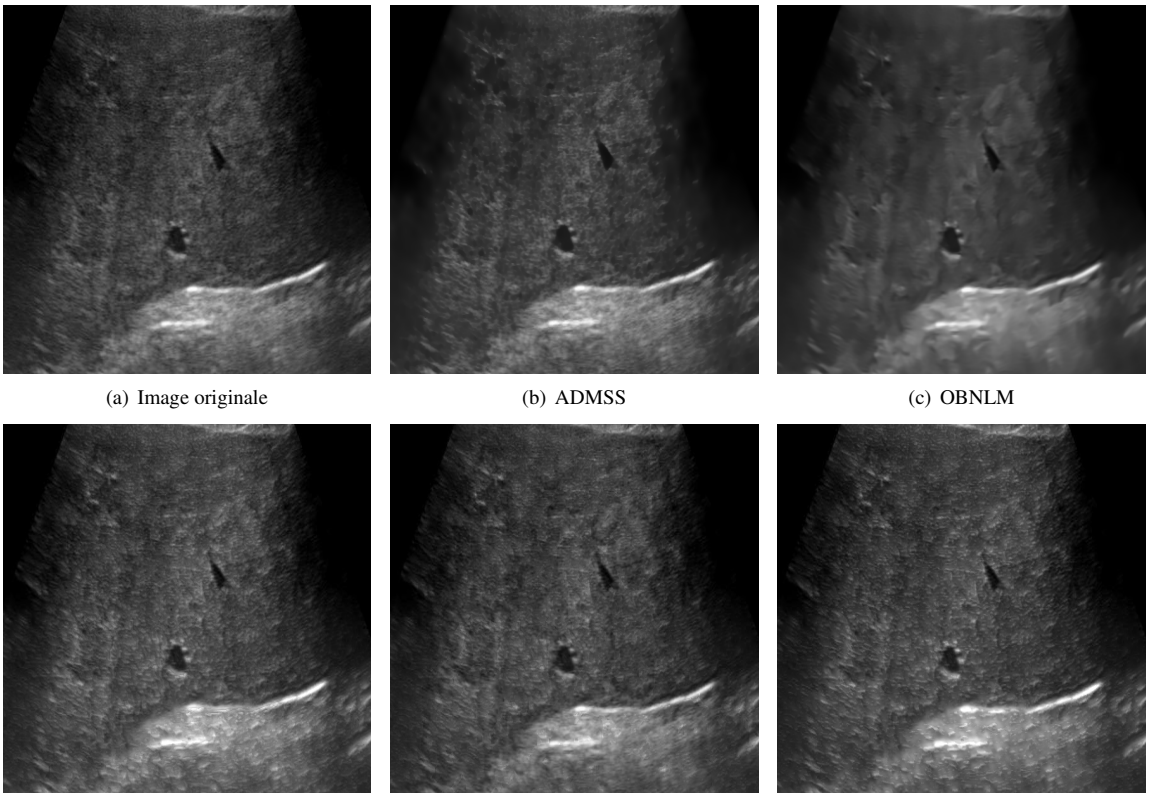
[7] J. Zhang, C. Wang, and Y. Cheng, *Comparison of Despeckle Filters for Breast Ultrasound Images*, Circuits, Systems, and Signal Processing , 2015

[8] Saha, Ashirbani and Wu, Qing Ming Jonathan *Utilizing Image Scales Towards Totally Training Free Blind Image Quality Assessment*, IEEE Transactions on Image Processing, 2015

[9] A. Mittal, R. Soundararajan, and A. C.Bovik, *Making a "Completely Blind" Image Quality Analyzer*,2013

[10] J. Kang, J.Y. Lee, and Y. Yoo, , *A New Feature-Enhanced Speckle Reduction Method Based on Multiscale Analysis for Ultrasound B-Mode Imaging*, IEEE Transactions on Biomedical Engineering, 2016

[11] M. Outtas, L. Zhang, O. Deforges, W. Hamidouche, A. Serir, and C. Cavaro-Menard, *A study on the usability of opinion-unaware no-reference natural image quality metrics in the context of medical images*,ISIVC 2016.



(a) Image originale (b) ADMSS (c) OBNLM

(d) Multi-Sortie1 Con-tours (e) Multi-Sortie2 Texture (f) Multi-Sortie3 Global

FIGURE 5 – Comparaison des méthodes de reduction de speckle

Qualité image/vidéo

Evaluation de la qualité des images stéréoscopiques basée sur les propriétés du système visuel humain

Yu FAN[1,2] Mohamed-Chaker Larabi[1] Faouzi Alaya Cheikh[2] Christine Fernandez-Maloigne[1]

[1] Laboratoire XLIM, Université de Poitiers, France

[2] Norwegian Coulour and Visual Computing Laboratory, NTNU-Gjovik, Norvège

{yu.fan, chaker.larabi}@univ-poitiers.fr

Résumé

L'un des plus grands défis pour l'évaluation de la qualité des images stéréoscopiques réside dans la façon de modéliser correctement les comportements binoculaires 3D du Système Visuel Humain (SVH). Cet article présente une métrique de qualité basée sur les propriétés visuelles monoculaires et binoculaires. Au lieu de mesurer la qualité à partir des vues de gauche et de droite séparément, la méthode proposée prédit la qualité d'une image cyclopéenne. Cette dernière est une approximation de la combinaison effectuée par le cerveau, et est obtenue en fusionnant les deux vues grâce à l'entropie locale dans le but de simuler les phénomènes de la fusion/rivalité binoculaire. Ainsi, une métrique de qualité 2D est utilisée pour estimer à la fois la qualité de l'image cyclopéenne et celle de la carte de disparité. De plus, la qualité de l'image cyclopéenne est modulée en fonction de l'importance visuelle de chaque pixel définie par le seuil de différence juste perceptible (JND). Enfin, L'estimation de la qualité d'une image stéréoscopique est obtenue en combinant les deux scores évoqués précédemment. Les résultats expérimentaux montrent que la méthode proposée surpasse les autres méthodes en termes de corrélation avec le jugement humain moyen et de complexité.

Mots clefs

Evaluation de la qualité des images stéréoscopiques, SVH, Image cyclopéenne, Fusion/rivalité binoculaire, JND.

1 Introduction

Ces dernières années, de grand efforts dans les technologies de la 3D stéréoscopique (3DS) ont été faits pour apporter une expérience visuelle 3D réaliste aux consommateurs. Cependant, l'avancée technologique apporte dans le même temps sont lot de défis à relever. Un des défis majeurs est liée à la qualité d'expérience comprenant les aspects de confort et de fatigue. Pour y parvenir, il est important de développer des métriques de qualité robustes et reproduisant le jugement humain pour un contenu 3DS. Alors que le domaine de la qualité des images 2D a beaucoup avancé ces dernières années, la recherche pour les images 3DS reste relativement marginale.

La complexité du problème de prédiction de la qualité 3DS vient du fait que la perception binoculaire résulte de la fusion effectuée par le cerveau en impliquant des caractéristiques jusque-là non prise en compte pour la qualité 2D. Bien entendu, la meilleure évaluation possible reste par le biais des expériences psychovisuelles qui malheureusement sont très fastidieuses et coûteuses. Par conséquent, les mesures objectives représentent la meilleure solution pour une intégration dans des systèmes en temps réel.

Une image stéréoscopique contient deux points de vue légèrement distants (c.-à-d. vues de gauche et de droite), dont chacun est projeté séparément sur la rétine. Lorsqu'une image 3D est observée, le Système Visuel Humain (SVH) fusionne les deux vues pour avoir une seule vue mentale (c.-à-d. l'image cyclopéenne) basée sur les propriétés de la perception binoculaire [1]. Ainsi, la qualité perceptuelle 3D dépend non seulement des qualités des vues 2D [2], mais également de l'information de profondeur [3] et des caractéristiques visuelles binoculaires [4]. L'idée est d'explorer comment ces attributs contribuent à la qualité globale 3D. Par conséquence, afin de développer une métrique 3D précise et efficace, il est important de comprendre et de tenir compte des différents processus perceptuels du SVH.

Dans cet article, nous proposons une nouvelle métrique de qualité stéréoscopique basée sur des propriétés de la vision binoculaire et combinant des qualités de l'image cyclopéenne et de la carte de disparité. La contribution majeure de cet article réside dans le développement d'une mesure de qualité 3D en modélisant les phénomènes de la Rivalité/Suppression Binoculaire (RB/SB), et la considération de la sensibilité spatiale monoculaire du SVH, ainsi que la qualité de la carte de disparité. En outre, nous fournissons une évaluation expérimentale complète pour notre méthode proposée et une comparaison détaillée avec d'autres méthodes de la littérature. Cet article est organisé comme suit : dans la section 3, nous faisons une revue brève de littérature à propos des métriques de qualité stéréoscopique. La section 4 présente et détaille la méthode proposée et nous évaluons et discutons les ses performances dans la section 5. Ce article se termine par une conclusion et des perspectives.

2 Etat de l'art

Dans cette section, nous examinons brièvement les méthodes récentes d'évaluation de la qualité d'images stéréoscopiques (EQIS). Selon le type et la quantité de l'information utilisée à partir des images stéréoscopiques, les méthodes sont divisées en trois classes [5] : (1) Méthodes basées sur la qualité des paires stéréoscopiques, (2) Méthodes basées sur la qualité de la paire stéréoscopique et sur des informations 3D, (3) Méthodes considérant des propriétés visuelles binoculaires et monoculaires.

Des méthodes EQIS de la première classe [6, 7] emploient des métriques de qualité 2D pour mesurer la qualité des vues de gauche et de droite séparément et ensuite combiner les deux qualités afin d'obtenir un score 3D. Par exemple, Campisi *et al.* [6] ont évalué la qualité 3D par quatre métriques 2D dont la SSIM [8], UQI [9], C4 [10] et une métrique à référence réduite [11]. Toutefois, considérant la combinaison des qualités de deux vues comme une qualité 3D ne correspond pas aux mécanismes de perception binoculaire du SVH [12]. Ainsi, les méthodes dans cette classe ne sont pas robustes pour des images stéréoscopiques dégradées de manière asymétrique. Ceci est principalement dû au fait que ces métriques 2D ne tiennent pas compte de l'information de profondeur, qui joue un rôle important sur la perception 3D.

Les méthodes de deuxième classe évaluent la qualité 3D à l'aide des informations de disparité/profondeur en plus des qualités de deux vues. Initialement, Benoit *et al.* [2] ont proposé une métrique 3D utilisant une référence image complète qui applique les métriques SSIM et C4 sur les vues de gauche et de droite indépendamment, et qui combine ces qualités par la suite avec l'estimation de dégradation de carte de disparité. Plus tard, You *et al.* [13] ont examiné les performances des mesures de qualité 2D utilisées dans le contexte de l'évaluation de la qualité 3D en utilisant différentes façon de combiner la qualité de la carte de disparité et la qualité des vues. Hwang et Wu [14] ont développé un modèle de prédiction de qualité 3D qui intègre les qualités des deux vues avec la qualité de profondeur et la saillance visuelle 3D. Récemment, Wang *et al.* [15] ont proposé une métrique utilisant une référence réduite, tenant compte des qualités des images de luminance et de la carte de disparité, basée sur les statistiques des images dans le domaine des contourlets. Comme la vérité terrain des cartes de profondeur/disparité n'est pas toujours disponible, les méthodes dans cette classe estiment les cartes de disparité en utilisant des algorithmes de mise en correspondance. Par conséquent, la précision de ces algorithmes peut affecter les performances de prédiction de la qualité 3D.

Les vues de gauche et de droite d'une paire stéréo peuvent être sujettes au même type/niveau de dégradation (à savoir la distorsion symétrique) ou à des types et/ou niveaux de dégradation différents (à savoir la distorsion asymétrique). Les distorsions symétriques conduisent à la Fusion Binoculaire (FB) [16], alors que les distorsions asymétriques causent la RB [17] ou la SB [18]. ces dernières ont un grand impact sur la qualité 3D perçue. Les méthodes EQIS de la première et seconde classes susmentionnées sont très utiles dans le cas d'une distorsion symétrique, mais le sont beaucoup moins pour des images stéréoscopiques avec une dégradation asymétrique. Cette dernière est de plus en plus adoptée dans les applications comme le codage des vidéos 3D. Afin d'améliorer les performances des métriques EQIS pour des distorsions asymétriques, la métrique 3D doit modéliser les énergies binoculaires d'une paire et prendre en compte la combinaison binoculaire. Les méthodes EQIS de la troisième classe considèrent les caractéristiques visuelles binoculaires comme la FB et la RB/SB en plus de la qualité 2D et l'information de disparité/profondeur.

Il est connu que le SVH est incapable de percevoir des changements de pixels en dessous d'un certain seuil visuel comme modélisé par les seuils de différence juste perceptibles (JND) en raison de leur sensibilité spatial/temporelle et les effets de masquage inhérents au SVH [19]. Certains modèles JND dédiés à la 3DS (JND-3D) prenant en compte les indices de profondeur binoculaire et monoculaire ont été proposés [20]. Par exemple, un modèle JND binoculaire (BJND), qui mesure le seuil de visibilité des distorsions asymétriques basée sur des effets de masquage, a été appliqué dans l'évaluation de la qualité 3D [4, 21].

D'autre approches mesurent la qualité 3D en exploitant l'image cyclopéenne obtenue à partir des deux vues. Par exemple, Chen *et al.* [22] ont développé une métrique qui mesure la qualité de l'image cyclopéenne construite par un modèle linéaire. Les poids de ce modèle sont obtenus par des réponses de l'amplitude du filtre Gabor, qui simule la RB. Par ailleurs, Lin et Wu [23] prédisent la qualité 3D grâce à la combinaison binoculaire et l'intégration de fréquence binoculaire. Récemment, Zhang et Chandler [5] ont présenté une métrique utilisant une référence image complète basée sur les qualités des images de gauche et de droite, de l'image cyclopéenne à l'aide de distance de luminance et de contraste des pixels.

3 Approche proposée

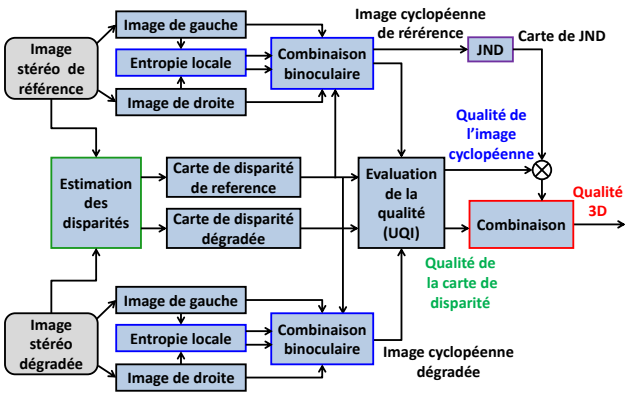


Figure 1 – *Schéma bloc de la métrique de l'EQSI proposée*

Comme mentionné précédemment, le SVH ne traite pas séparément les stimuli de gauche et de droite. Il perçoit les dégradations de l'image cyclopéenne comme des distorsion 2D, et les dégradations de l'image de profondeur comme des distorsions 3D. Inspiré par cela, notre métrique évalue la qualité 3D en combinant la qualité de l'image cyclopéenne avec celle de la carte de disparité comme le montre le diagramme de la Figure 1.

Ainsi, la première étape de notre métrique consiste à extraire les cartes de disparité des paires stéréoscopiques de référence et dégradée. Pour y parvenir, nous utilisons l'algorithme de mise en correspondance récemment proposé par Lee *et al.* [24]. Cet algorithme peut efficacement estimer des valeurs de disparité et résoudre les problèmes d'occlusion et de discontinuité. Ensuite, en fonction du modèle de combinaison linéaire décrit dans [22, 25], pour la modélisation des phénomènes de FB/RB lorsqu'un stimulus stéréo est présenté, nous générons l'image cyclopéenne synthétisée de la manière suivante :

$$I_c(i,j) = \frac{EL_l(i,j)}{EL_T(i,j)} \cdot I_l(i,j) + \frac{EL_r(i,j-d_l)}{EL_T(i,j)} \cdot I_r(i,j-d_l), \quad (1)$$

$$EL_T(i,j) = EL_l(i,j) + EL_r(i,j-d_l), \quad (2)$$

où I_l et I_r représentent respectivement les vues de gauche et de droite, EL_l et EL_r indiquent leurs cartes d'énergie locale, avec (i,j) les coordonnées du pixel. d_l représente la valeur de disparité du pixel (i,j) correspondant au déplacement horizontal de la vue de gauche à la vue de droite. L'étude présentée dans [12] a montré que la perception binoculaire est dominée par la vue la plus contrastée ou ayant les contours les plus riches. Autrement dit, la qualité perceptuelle 3D suit la qualité de la vue contenant une plus grande quantité d'informations. Par conséquent, nous proposons de modéliser l'énergie locale par l'entropie locale de l'image pour déterminer l'intensité de stimulus de chaque vue et ce en s'inspirant des travaux de Fezza et Larabi [21, 31]. Ici, l'entropie d'un pixel est calculée dans un voisinage 11×11 avec une forme spécifique autour de ce pixel, comme formalisé par l'équation suivante :

$$EL(i,j) = -\sum_{s=g_{min}}^{g_{max}} p(x_s) \times \log_2(p(x_s)), \quad (3)$$

où g_{min} et g_{max} sont respectivement les valeurs minimum et maximum des pixels voisins, et $p(x_s)$ correspond à la probabilité que la différence entre deux pixels adjacents soit égale à s. En partant des équations 1, 2 et 3, la métrique proposée simule les phénomènes de FB/RB. Plus précisément, les différentes entropies locales dans les deux vues conduisent à la RB, et la qualité 3D d'une région de l'image stéréoscopique est plus affectée par la vue contenant des entropies plus élevées. Étant donné les images cyclopéennes (I_{rc}, I_{dc}) et les cartes de disparité (Dp_r, Dp_d) des images stéréoscopiques de référence et dégradée, nous mesurons séparément la qualité cyclopéenne et de disparité en utilisant une métrique de qualité 2D. You *et al.* [13] ont montré qua la métrique UQI présente une meilleure performance pour la prédiction de la qualité 3D parmi toutes les métriques 2D testées. Par ailleurs, UQI a également offert une bonne performance pour la qualité de la carte de disparité. En fait, la métrique UQI utilisée dans l'estimation de qualité de disparité est basée sur une comparaison d'information structurelle et la disparité peut représenter ces informations à partir des images originales. Ainsi, les qualités de l'image cyclopéenne et de la carte de disparité sont calculées comme suit :

$$Q_c(i,j) = UQI(I_{rc}, I_{dc}), \;\; Q_d = UQI(Dp_r, Dp_d), \quad (4)$$

où Q_c est la carte UQI de l'image cyclopéenne testée, et Q_d indique la qualité de la carte de disparité. Afin d'améliorer la performance de métrique 3D, nous avons utilisé l'importance visuelle du pixel afin de pondérer la qualité de l'image cyclopéenne [4]. L'importance visuelle, ce qui correspond à une sensibilité spatiale monoculaire du SVH, est décrite par les seuils JND [26] de l'image cyclopéenne de référence. En conséquence, la qualité de l'image cyclopéenne pondérée par le JND Q_c^{JND} est calculée par :

$$Q_c^{JND} = \frac{\sum_{i,j}^{N} \left[\frac{1}{JND(i,j)} \times Q_c(i,j) \right]}{\sum_{i,j}^{N} \frac{1}{JND(i,j)}}, \quad (5)$$

où N est le nombre de pixels dans l'image cyclopéenne. Une valeur élevée de JND d'un pixel signifie que ce pixel peut tolérer une dégradation importante, et a donc un faible impact visuel sur la qualité perçue. Enfin, la qualité 3D Q_{3D} est déterminée par le modèle linéaire suivante :

$$Q_{3D} = \alpha \times Q_c^{JND} + (1-\alpha) \times Q_d, \quad (6)$$

où α est le paramètre permettant de régler l'importance relative des Q_c et Q_d dans la qualité 3D globale. Dans notre implémentation, nous avons choisi $\alpha = 0.95$ pour la base de données LIVE 3D IQA phase I [27] et $\alpha = 0.6$ pour la base de données LIVE 3D IQA phase II [22]. Cette variation de la pondération s'explique par le fait que la qualité de l'image de disparité influence moins la qualité 3D dans la phase I (ne contenant que des images stéréoscopiques symétriques) que dans la phase II (contenant des images stéréoscopiques dégradées à la fois symétriques et asymétriques).

4 Évaluation des performances

Dans cette section, nous évaluons les performances de la méthode proposée en comparaison à d'autres méthodes EQIS sur deux bases de données accessibles au public fournissant des notes subjectives de qualité (valeurs DMOS) à savoir, LIVE 3D IQA phase I [27] and phase II [22]. La base de données LIVE 3D phase I contient 20 paires stéréoscopiques originales et 365 paires dégradées de façon symétrique, comprenant le bruit blanc (WN), le flou gaussien (GB), la compression JPEG et JPEG 2000 (JP2K) et le fast fading (FF). La base de données LIVE 3D phase

Tableau 1 – *Performance de la métrique proposée comparée aux approches standard 2D (en italique) et 3D sur la base de données LIVE 3D IQA (phase I). TC indique le temps de calcul (en secondes) pour toutes les images. Les meilleurs résultats sont donnés en gras.*

Type de distorsion	Critères	*SSIM [8]*	*MS-SSIM [28]*	*FSIM [29]*	*VIF [30]*	*UQI [9]*	Wang [4]	Fezza [21]	Fezza [31]	Chen [22]	Proposée
WN	LCC	*0.944*	*0.952*	*0.931*	*0.930*	*0.927*	0.949	0.941	0.947	**0.955**	0.939
	SROCC	*0.939*	*0.942*	*0.929*	*0.931*	*0.926*	0.947	0.935	0.944	**0.948**	0.935
	RMSE	*5.500*	*5.070*	*6.094*	*6.103*	*6.240*	5.254	5.620	5.351	**4.963**	5.732
JPEG	LCC	*0.475*	*0.633*	*0.623*	*0.603*	*0.769*	0.473	0.274	0.706	0.527	**0.779**
	SROCC	*0.435*	*0.613*	*0.582*	*0.580*	*0.737*	0.450	0.246	0.657	0.521	**0.745**
	RMSE	*5.755*	*5.063*	*5.116*	*5.216*	*4.178*	5.762	6.289	4.632	5.557	**4.105**
JP2K	LCC	*0.858*	*0.930*	*0.908*	*0.888*	*0.944*	0.875	0.783	0.937	0.920	**0.956**
	SROCC	*0.857*	*0.892*	*0.905*	*0.902*	*0.910*	0.856	0.774	0.896	0.887	**0.916**
	RMSE	*6.663*	*4.752*	*5.424*	5.959	4.270	6.272	8.822	4.532	5.070	**3.787**
GB	LCC	*0.907*	*0.944*	*0.933*	***0.962***	*0.952*	0.893	0.908	0.934	0.943	0.957
	SROCC	*0.879*	*0.925*	*0.922*	***0.934***	*0.925*	0.871	0.867	0.909	0.924	0.921
	RMSE	*8.774*	*4.790*	*5.205*	***3.955***	*4.451*	6.512	6.058	5.173	4.813	4.196
FF	LCC	*0.670*	*0.803*	*0.815*	*0.862*	*0.879*	0.644	0.641	0.783	0.776	**0.884**
	SROCC	*0.584*	*0.722*	*0.729*	*0.804*	***0.833***	0.525	0.515	0.693	0.700	0.819
	RMSE	*9.227*	*7.405*	*7.199*	*6.306*	*5.925*	9.508	9.541	7.730	7.832	**5.915**
ALL	LCC	*0.877*	*0.856*	*0.915*	*0.925*	*0.943*	0.868	0.833	0.821	0.922	**0.944**
	SROCC	*0.877*	*0.824*	*0.928*	*0.920*	*0.937*	0.868	0.823	0.922	0.914	**0.941**
	RMSE	*7.889*	*8.472*	*6.614*	*6.230*	*5.478*	8.131	9.063	9.358	6.351	**5.430**
	TC	***30***	*51*	*903*	*688*	*41*	644	1263	1231	10435	2259

II est quant à elle composée de 8 paires stéréoscopiques originales et 360 paires dégradées symétriquement et asymétriquement correspondant aux mêmes types de distorsion que la phase I. Nous comparons la méthode proposée avec quatre métriques récentes [4, 21, 31, 22]. Nous utilisons le même algorithme de correspondance [24] afin d'estimer les disparités pour toutes les méthodes sauf [22] dans le but d'assurer une comparaison juste. En outre, nous avons également étudié les performances des métriques 2D dont SSIM, MS-SSIM [28], FSIM [29], VIF [30] et UQI. Pour ces méthodes basées seulement sur les métriques 2D, la moyenne des qualités des vues de gauche et de droite a été considérée comme la qualité 3D. Les performances de ces métriques ont été évaluées à l'aide de trois critères/mesures : Le coefficient de corrélation linéaire (LCC), La corrélation d'ordre de Spearman (SROCC) and l'erreur quadratique moyenne (RMSE). Les trois mesures ont été calculées entre les notes subjectives (DMOS) et les notes objectives après l'application d'une régression non linéaire avec une fonction logistique de cinq paramètres décrite dans [32]. Tous les tests ont été effectués en exécutant le code MATLAB sur un ordinateur portable (Inter Core i7-2630 QM processeur à 2,00 GHz, 4 Go de RAM, Windows 7).

4.1 Performances globales et partielles

Le tableau 1 présente les performances de méthodes EQIS pour LIVE 3D phase I. Nous remarquons dans ce tableau que la méthode proposée surpasse la plupart des approches 2D/3D en terme de corrélation. La méthode de Chen [22] est plus performante que celles de Fezza [21] et de Wang [4] puisqu'elle a tenue compte des phénomènes de FB/RB modélisés par l'image cyclopéenne. Cependant, en raison de l'utilisation du filtre de Gabor 2D, cette dernière est moins rapide et plus coûteuse. Toutes les méthodes basées sur les métriques 2D atteignent des performances intéressantes pour les dégradations symétriques, et les résultats des métriques UQI et de VIF sont meilleures que certaines métriques 3D à l'exception de la métrique proposée. Plus précisément, nous examinons également dans ce tableau les performances sur chaque type de distorsion. La méthode proposée donne des corrélations et des précisions de prédiction plus intéressantes que celles obtenues avec d'autres méthodes sur la plupart des types de distorsion à l'exception de WN et GB. Les performances de ces dernières sont compétitives et restent d'un niveau acceptable. Pour la distorsion WN, nous avons constaté que la méthode de Chen a obtenu les meilleures performances parce que la métrique MS-SSIM utilisée dans l'approche peut engendrer une prédiction haute pour les images stéréoscopiques bruitées. Cette observation indique que les performances de certaines méthodes dépendent fortement de celles de la métrique 2D utilisée. En général, toutes les métriques 2D ou 3D ont obtenu des résultats intéressants dans la base de données de LIVE 3D phase I.

La prédiction de qualité sur la base de données LIVE 3D phase II, contenant partiellement des dégradations asymétriques, est plus difficile que sur LIVE 3D phase I. Pour chaque méthode d'EQIS, le tableau 2 présente les performances globales et sur les sous-ensembles séparés en fonction de la nature des dégradations (symétrique et asymé-

Tableau 2 – *Performance de la métrique proposée comparée aux approches standard 2D (en italique) et 3D sur la base de données LIVE 3D IQA (phase II). AS et S dénotent respectivement les distorsions asymétriques et symétriques.*

Method	LCC			SROCC			RMSE			TC
	S	As	Total	S	As	Total	S	As	Total	Total
SSIM [8]	*0.852*	*0.767*	*0.802*	*0.826*	*0.736*	*0.793*	*6.543*	*6.510*	*6.736*	***30***
MS-SSIM [28]	*0.927*	*0.719*	*0.795*	*0.912*	*0.684*	*0.777*	*4.694*	*7.047*	*6.851*	*49*
FSIM [29]	*0.929*	*0.731*	*0.808*	*0.912*	*0.684*	*0.786*	*4.623*	*6.913*	*6.654*	*919*
VIF [30]	*0.928*	*0.777*	*0.837*	*0.916*	*0.732*	*0.819*	*4.653*	*6.383*	*6.184*	*684*
UQI [9]	***0.940***	*0.794*	*0.863*	***0.938***	*0.755*	*0.841*	***4.223***	*6.159*	*5.685*	*38*
Wang [4]	0.862	0.743	0.771	0.826	0.696	0.771	6.334	6.787	7.188	82
Fezza [21]	0.788	0.713	0.751	0.778	0.676	0.734	7.685	7.104	7.453	163
Fezza [31]	0.930	0.820	0.871	0.921	0.796	0.862	4.576	5.801	5.553	1410
Chen [22]	0.939	**0.878**	**0.909**	0.927	**0.858**	**0.904**	4.277	**4.846**	**4.700**	14089
Proposed	**0.940**	0.877	0.906	**0.938**	0.839	0.893	4.269	4.878	4.780	2157

trique) dans LIVE 3D phase II.

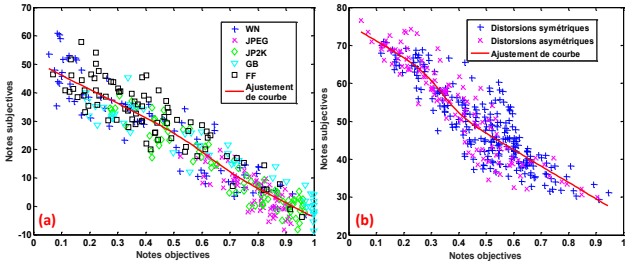


Figure 2 – *Nuage de points représentant les scores subjectifs DMOS en fonction des scores prédits par la métrique proposée pour les deux bases de données : LIVE 3D phase I (gauche), LIVE 3D phase II (droite).*

Les résultats illustrés dans le tableau 2 démontrent que la méthode proposée surpasse les autres méthodes à l'exception de la méthode de Chen [22] dans le cas de distorsions asymétriques. En fait, la méthode proposée est assez similaire à celle de Chen au regard de la performance globale, mais notre méthode est évidemment beaucoup plus rapide. De plus, la méthode proposée est plus performante dans le cas de distorsions symétriques. Ainsi, la méthode proposée engendre une précision de prédiction élevée avec une complexité de calcul faible. En revanche, la plupart des métriques 2D sont aussi efficaces que celles 3D pour les paires stéréoscopiques symétriques, mais elles donnent en général les mauvaises performances pour les distorsions asymétriques. Ceci est principalement dû au fait que les méthodes basées sur une métrique 2D évaluent la qualité 3D sans tenir compte de l'information de profondeur/disparité ni des caractéristiques de la vision binoculaire.

Ainsi les méthodes basées sur image cyclopéenne (c.-à-d., de Chen [22], [31] et notre méthode) donnent de meilleures performances que les autres. En plus de la comparaison des performances mentionnée ci-dessus, nous illustrons dans la Figure 2 les distributions de notes subjectives (DMOS) en fonction des notes objectives estimées par la méthode proposée, ainsi que la courbe de régression non linéaire.

Tableau 3 – *Performances de la métrique proposée sur la base de donnée LIVE 3D IQA (phase II) utilisant des stratégies différents.*

Stratégies	LCC	SROCC	RMSE
Sans JND	0.903	0.883	4.779
Sans EQD	0.889	0.864	5.163
Sans JND et EQD	0.887	0.867	5.177
Avec JND et EQD	**0.906**	**0.893**	**4.779**

4.2 Influence de JND et de la qualité de disparité sur performance

Dans cette section, nous montrons l'intérêt de la prise en compte du JND et de la qualité de la carte de disparité dans la méthode proposée. Nous comparons dans le Tableau 3 les performances et l'influence de chaque composante du schéma bloc de la métrique illustré dans la Figure 1. L'approche sans JND n'utilise pas la carte de JND pour pondérer la qualité de l'image cyclopéenne de référence, tandis que l'approche sans la qualité de la carte de disparité (EQD) ne considère que la qualité de l'image cyclopéenne comme qualité 3D. A partir de ces résultats, nous remarquons que la méthode proposée (c.-à-d., avec JND et EQD) donne les meilleures performances parmi toutes les approches. Toutefois, la méthode proposée surpasse légèrement l'approche sans JND en terme de LCC. En résumé, les résultats du Tableau 3 montrent que la performance de la prédiction de la qualité 3D peut être améliorée en tenant compte du JND et de la qualité de la disparité. Par ailleurs, nous avons également examiné les performances de la méthode proposée pour différents types de distorsion. Notre méthode fonctionne bien pour les distorsions GB et FF.

5 Conclusion

Nous avons présenté dans cet article une méthode d'évaluation de qualité pour des images stéréoscopiques basée sur les propriétés visuelles binoculaires et monoculaires du SVH. La méthode proposée modélise la vision binoculaire humaine en fusionnant les vues de gauche et de droite pour générer une image cyclopéenne et prendre en compte l'aspect de disparité ainsi que la sensibilité spatiale monoculaire du SVH. Les résultats expérimentaux montrent que la

méthode proposée est bien corrélée avec la perception humaine, et surpasse les autres méthodes en termes de précision de la prédiction et de complexité de calcul. La prise en compte de la modélisation du processus de rivalité binoculaire sera effectuée dans les travaux à venir afin d'améliorer les performances de cette approche.

Références

[1] Bela Julesz. Foundations of cyclopean perception. 1971.

[2] A. Benoit, P. Le Callet, P. Campisi, et R. Cousseau. Quality assessment of stereoscopic images. *EURASIP journal on image and video processing*, 2008.

[3] ZM Sazzad, R. Akhter, J. Baltes, et Y. Horita. Objective no-reference stereoscopic image quality prediction based on 2d image features and relative disparity. *Advances in Multimedia*, 2012 :8, 2012.

[4] X. Wang, S. Kwong, et Y. Zhang. Considering binocular spatial sensitivity in stereoscopic image quality assessment. Dans *VCIP*, pages 1–4. IEEE, 2011.

[5] Y. Zhang et D. M Chandler. 3d-mad : A full reference stereoscopic image quality estimator based on binocular lightness and contrast perception. *IEEE Trans. Image Process.*, 24(11) :3810–3825, 2015.

[6] P. Campisi, P. Le Callet, et E. Marini. Stereoscopic images quality assessment. Dans *Signal Processing Conference, 2007 15th European*, pages 2110–2114. IEEE, 2007.

[7] J. Yang, C. Hou, Y. Zhou, Z. Zhang, et J. Guo. Objective quality assessment method of stereo images. Dans *3DTV Conference : The True Vision-Capture, Transmission and Display of 3D Video*, pages 1–4, 2009.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, et E. P Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Trans. Im. Process.*, 13(4) :600–12, 2004.

[9] Z. Wang et A. C Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3) :81–84, 2002.

[10] M. Carnec, P. Le Callet, et D. Barba. An image quality assessment method based on perception of structural information. Dans *ICIP*, volume 3, pages III–185. IEEE, 2003.

[11] Z. Wang et E. P Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. Dans *Electronic Imaging 2005*, pages 149–159. International Society for Optics and Photonics, 2005.

[12] D. V Meegan, L. B Stelmach, et W J. Tam. Unequal weighting of monocular inputs in binocular combination : implications for the compression of stereoscopic imagery. *Journal of Experimental Psychology : Applied*, 7(2) :143, 2001.

[13] J. You, L. Xing, A. Perkis, et X. Wang. Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis. Dans *VPQM, AZ*, 2010.

[14] J.-J. Hwang et H. R. Wu. Stereo image quality assessment using visual attention and distortion predictors. *TIIS*, 5(9) :1613–1631, 2011.

[15] X. Wang, Q. Liu, R. Wang, et Z. Chen. Natural image statistics based 3d reduced reference image quality assessment in contourlet domain. *Neurocomputing*, 151 :683–691, 2015.

[16] Jeremy M Wolfe. Stereopsis and binocular rivalry. *Psychological review*, 93(3) :269, 1986.

[17] R. Blake et N. K Logothetis. Visual competition. *Nature Reviews Neuroscience*, 3(1) :13–21, 2002.

[18] J. Brascamp, H. Sohn, S.-H. Lee, et R. B.. A monocular contribution to stimulus rivalry. *PNAS*, 110(21) :8337–8344, 2013.

[19] N. Jayant, J. Johnston, et R. Safranek. Signal compression based on models of human perception. *Proceedings of the IEEE*, 81(10) :1385–1422, 1993.

[20] Y. Fan, M.-C. Larabi, F. A. Cheikh, et C. Fernandez. On the performance of 3d just noticeable difference models. Dans *ICIP*, pages 1017–1021. IEEE, 2016.

[21] S. A. Fezza, M.-C. Larabi, et K. M. Faraoun. Stereoscopic image quality metric based on local entropy and binocular just noticeable difference. Dans *ICIP*, pages 2002–6, 2014.

[22] M.-J. Chen, C.-C. Su, D-K. Kwon, L. K Cormack, et A. C Bovik. Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Processing : Image Communication*, 28(9) :1143–1155, 2013.

[23] Y.-H. Lin et J.-L. Wu. Quality assessment of stereoscopic 3d image compression by binocular integration behaviors. *IEEE Trans. Image Process.*, 23(4) :1527–1542, 2014.

[24] S. Lee, J. H. Lee, J. Lim, et Il H. Suh. Robust stereo matching using adaptive random walk with restart algorithm. *Image and Vision Computing*, 37 :1–11, 2015.

[25] W. JM Levelt. *On binocular rivalry*. Thèse de doctorat, Van Gorcum Assen, 1965.

[26] A. Liu, W. Lin, M. Paul, C. Deng, et F. Zhang. Just noticeable difference for images with decomposition model for separating edge and textured regions. *IEEE Trans. Circuits Syst. Video Technol.*, 20(11) :1648–1652, 2010.

[27] A. K. Moorthy, C.-C. Su, A. Mittal, et A. C. Bovik. Subjective evaluation of stereoscopic image quality. *Signal Processing : Image Communication*, 28(8) :870–883, 2013.

[28] Z. Wang, E. P Simoncelli, et A. C Bovik. Multiscale structural similarity for image quality assessment. Dans *Signals, Systems and Computers*, pages 1398–1402. IEEE, 2003.

[29] L. Zhang, L. Zhang, X. Mou, et D. Zhang. Fsim : a feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20(8) :2378–2386, 2011.

[30] H. R. Sheikh et A. C Bovik. Image information and visual quality. *IEEE Trans. Image Process.*, 15(2) :430–444, 2006.

[31] S. A. Fezza et M.-C. Larabi. Stereoscopic 3d image quality assessment based on cyclopean view and depth map. Dans *ICCE*, pages 335–339. IEEE, 2014.

[32] H. R. Sheikh, M. F. Sabir, et A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11) :3440–3451, 2006.

Evaluation de la qualité de vidéo médicales compressées par MPEG-4 AVC/H.264 et HEVC pour la Télémédecine

A. Chaabouni[1] J. Lambert[3] Y. Gaudeau[1,2] N. Tizon[4] D. Nicholson[4] J-M. Moureaux[1]

[1] Université de Lorraine, CRAN UMR 7039, 9 avenue de la Foret de Haye Vandoeuvre les Nancy, 54500 France

{amine.chaabouni, jean-marie.moureaux}@univ-lorraine.fr

[2] Université de Strasbourg, 30 Rue du Maire Andre Traband, Haguenau, 67500 France

{yann.gaudeau}@unistra.fr

[3] Institut Mines Telecom, 37-39 Rue Dareau, 75014, Paris France

{julien.lambert}@imt.fr

[4] VITEC, 99 Rue Pierre Semard, 92320 Châtillon France

{nicolas.tizon, didier.nicholson}@vitec.com

Résumé

Afin de répondre aux besoins de la communauté médicale en termes de stockage ou de partage des données à distance, la compression avec pertes semble, aujourd'hui, être la solution la plus appropriée pour gérer la grande quantité de ces données. Cependant, la distorsion engendrée par la compression doit être évaluée de façon à vérifier que les données compressées restent compatibles avec les usages. Dans ce contexte, la réalisation de tests subjectifs normalisés d'évaluation de la qualité par des experts médicaux est essentielle pour valider les résultats de compression. Ici, nous nous proposons de comparer les performances des deux derniers standards d'encodage vidéo, mesurant ainsi le compromis qualité/débit après compression. Les résultats nous montrent que HEVC est plus efficace que MPEG-4 AVC/H.264, offrant jusqu'à 54% d'économie de débit par rapport à AVC / H.264. En outre, nous avons montré que des vidéos médicales telles que des vidéos endoscopiques ORL pouvaient être, avantageusement, codées en résolution SD au lieu de Full HD pour les applications à faible débit. Enfin, par rapport à la perception des médecins, les métriques objectives appropriées MSE, NQM, SSIM et MSSIM valident les résultats précédents et confirment la supériorité de HEVC sur MPEG-4 AVC/H.264. Ces résultats sont très prometteurs pour les applications en télémédecine, en particulier dans un contexte à faible débit.

Mots clefs

Evaluation objective et subjective de la qualité, standard de compression MPEG-4 AVC/H.264/HEVC, résolution SD/Full HD, traitement biomédical de l'image.

1 Introduction

De nos jours, le développement de la télémédecine continue à s'accélérer, en particulier pour faire face aux problèmes de distance et de coûts connexes entre les hôpitaux locaux et les hôpitaux de référence ; mais aussi pour d'autres raisons, comme par exemple pour maintenir au maximum les patients à domicile. Grâce à la télémédecine, les médecins peuvent fournir à des soins de haute qualité grâce à des consultations à distance, à la télé-radiologie et à la surveillance à distance, par exemple. En outre, la télémédecine offre un outil efficace pour les praticiens de santé qui veulent partager leur expertise à travers des conseils médicaux et des diagnostics à distance, en particulier pour les cas difficiles. Afin de réaliser ces différents scénarios, il est nécessaire de stocker et de transmettre des données très volumineuses telles que les flux vidéo médicaux et les images sur les réseaux à large bande passante, mais aussi sur les faibles bandes passantes, en particulier dans les zones rurales. Ainsi, pour un flux endoscopique de très haute définition (FHD) original codé sur environ 2 Gbits/s, il est essentiel de pouvoir le compresser pour assurer une retransmission en temps réel tout en conservant une qualité suffisante pour une utilisation régulière par les professionnels de la santé. Si la compression sans perte a été utilisée pour les applications médicales, car elle préserve l'intégrité des données, sa faible performance (en termes de débit binaire) n'est pas adaptée à ces applications. D'autre part, de nombreux travaux [1, 2] ont montré que les images médicales (et les données vidéo) étaient tolérantes à une compression avec perte à condition que la distorsion due à la compression soit maîtrisée. Dans ce contexte, l'Association Canadienne des Radiologistes (CAR) ainsi que le Collège Américain de Radiologie (ACR) [3], recommandent l'utilisation d'une compression avec perte dans le contexte médical sous certaines conditions. La manière la plus appropriée d'évaluer la qualité des données médicales compressées par rapport à leur utilisation consiste à effectuer des tests subjectifs afin de trouver un compromis entre l'efficacité de la compression et la perception des experts de la qualité des données médicales après compression. Une pre-

mière étude [4, 5] a été menée par une partie des auteurs s'appuyant sur le projet européen Celtic Plus HIPERMED (HIgh PERformance teleMEDicine plateform [1]), traitant le problème de l'évaluation de la qualité des séquences vidéo compressées AVC. Ici, nous proposons d'étendre cette étude au cas de la nouvelle norme HEVC et à d'autres séquences vidéo. En plus du service de télémédecine de haute qualité (HIPERMED), disponible dans les hôpitaux dotés de très bonnes infrastructures de réseau, le projet européen Celtic Plus E3 (E-health services Everywhere and for Everybody) vise à développer une plate-forme basée sur le Web pour offrir une large gamme de services de télémédecine, qui seront également accessibles à partir de terminaux grand public. Une large accessibilité est ciblée, en particulier dans des conditions de réseau dégradées. Le service devrait rester disponible tout en gardant un niveau de QoE acceptable. En ce qui concerne les aspects de codage vidéo, cette exigence d'évolutivité implique des capacités d'adaptation des débits afin de fournir la meilleure qualité pour une bande passante limitée donnée. Afin d'adapter le débit de compression de la vidéo, différentes approches peuvent être envisagées. Classiquement, on s'attend à ce que trois paramètres soient utilisés à des fins de contrôle de débit : la résolution spatiale, la fréquence d'images et le pas de quantification. Dans le contexte de la visioconférence, en raison de la contrainte de faible latence, la fréquence d'images devrait rester aussi élevée que possible, en particulier dans les interventions endoscopiques ORL. Ainsi, face à une diminution des conditions de transmission, le mécanisme d'adaptation doit décider soit de diminuer la résolution spatiale, soit d'augmenter le facteur de quantification. Par conséquent, afin d'aider à mettre en œuvre la meilleure stratégie, cette étude fournit des résultats comparatifs des évaluations subjectives obtenues à partir de contenus vidéos médicaux compressés à différentes résolutions (SD et FHD) et de différents niveaux de quantification. Dans un contexte proche, les performances de qualité de HEVC ont été évaluées dans le cadre de vidéos échographiques [6]. Dans notre étude, nous essayons de montrer les améliorations apportées par HEVC à la qualité vidéo par rapport à MPEG-4 AVC/H.264, pour les résolutions FHD et SD. Cet article est organisé comme suit : Les différents outils et méthodes utilisés au cours de cette étude pour l'évaluation de la qualité sont présentés dans la section 2. La section 3 est consacrée aux résultats expérimentaux. Enfin, nous concluons cette étude dans la section 4.

2 Matériel et Méthodes

2.1 Encodage MPEG-4 AVC/H.264 vs HEVC

Cette étude présente une comparaison de performance entre les deux dernières normes de codage vidéo HEVC et AVC/H264.

1. http ://hipermed.eu

— 2.1.1 Brève description des deux standards :

Dans la plate-forme HIPERMED, les différents flux ont été compressés à l'aide du standard MPEG-4 AVC/ITU-T H.264 [7] (note : l'acronyme officiel AVC sera utilisé tout au long de ce document) en raison de ses performances en comparaison de l'encodage vidéo par les normes précédentes. Il est actuellement le standard le plus utilisé dans les applications de réseau et de transport. AVC a été développé conjointement par l'UIT-T et l'ISO/IEC et est le produit d'un effort de partenariat connu sous le nom de Joint Video Team (JVT). Avec la plate-forme HIPERMED, nous utilisons l'encodeur x264 [8], qui permet d'encoder et de transmettre les vidéos en temps réel. Considéré comme le successeur d'AVC, le nouveau standard de compression HEVC [9] (High Efficiency Video Codage) améliore le compromis entre le taux de compression et la qualité visuelle. Il a été développé et finalisé en janvier 2013 par l'équipe JVT. Il peut supporter les résolutions ultra haute définition 4k (3840 x 2160) et 8K (7680 x 4320) et permet un traitement en parallèle, profitant des architectures multicœur. Cette norme est basée sur l'unité d'arbre d'encodage (CTU) avec des tailles de macro bloc plus grandes allant de 16 à 64, offrant plus d'efficacité et de flexibilité par rapport à AVC qui utilise des blocs plus petits (16x16, 16x8, 8x16, 8x8, 8x4, 4x8, 4x4). En outre, il applique 3 filtres : Filtre de déblocage, Décalage d'échantillon (SAO) et Filtrage de boucle adaptatif (ALF), contrairement à son prédécesseur qui n'applique que le filtre de déblocage. Ainsi, il diminue davantage les artefacts dans les images reconstruites. Au lieu d'utiliser un codage entropique différent, HEVC utilise uniquement le CABAC. Dans notre étude, nous utilisons l'encodeur x265 [10], l'implémentation principale « open source » de HEVC qui peut être configurée pour s'exécuter en temps réel sur les dernières architectures x86, diffusées sous GNU GPL, la même licence que x264.

— 2.1.2 Paramètres d'encodage AVC/H.264 et HEVC :

Avec l'aide du Dr. Gallet, chirurgien ORL (OtoRhino-Laryngologie), à l'hôpital universitaire de Nancy, France, nous avons sélectionné 4 séquences ORL originales, acquises à partir d'une caméra endoscopique « Storz » avec la tête de caméra « S1 » et l'unité de contrôle d'image « Image 1 HUB ». Les séquences sont liées à des chirurgies ORL réelles et identifiées comme critiques vis à vis de la qualité. Ces séquences durent 10 secondes et sont initialement encodées à 1,99 Gbits/s dans une résolution Full HD (1920x1080 - 1080p60 - 4 : 2 : 2 - 8 bits). Deux des 4 vidéos originales (séquence 1 et 2 de la figure 1) ont été soumises comme séquences de référence d'imagerie médicale pour le développement HEVC au groupe vidéo JCT-VC [11]. Elles sont présentées dans la figure 1. Pour être compatibles avec les contraintes liées au temps réel et à la latence pour le codage AVC et HEVC, nous nous sommes basés sur les paramètres de compression, résumés dans le tableau 1.

2.2 Présentation de l'évaluation de la qualité

— 2.2.1 Protocole des tests subjectifs :

Les tests subjectifs effectués par un panel d'experts sont essentiels pour évaluer l'impact du post-traitement sur les vidéos médicales, en particulier pour les applications sensibles telles que le diagnostic ou la chirurgie. Ici, nous proposons de suivre le protocole ITU-BT.500-13 (de l'union internationale des télécommunications) [12], qui fournit des méthodologies pour l'évaluation de la qualité de l'image et de la vidéo, y compris les méthodes générales de test, les échelles de classement et les conditions de visualisation. En se basant sur cette norme, il est recommandé d'effectuer la méthode de l'évaluation de la qualité à double stimulus (DSCQS) [12] en utilisant une échelle continue comme indiqué sur la figure 2.

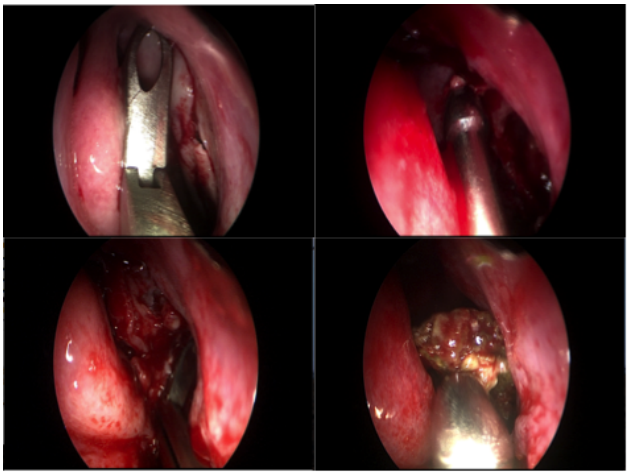


Figure 1 – *Séquences endoscopiques originales ORL numérotées de 1 à 4 dans le sens des aiguilles d'une montre en haut à gauche*

Paramétre	Format Pixel	Résolution	Fréquence
Valeur	uyvy422	1920x1080/720x404	60
Paramètre	Intra-refresh	Key-int	Latence
Valeur	1	60	zerolatency

Tableau 1 – *La configuration d'encodage*

— 2.2.2 Les outils d'évaluation de la qualité subjective :

Pour effectuer des tests subjectifs, nous avons utilisé le « living lab » PROMETEE [2], une plate-forme d'innovation permettant d'étudier et de gérer efficacement la qualité des vidéos associée à leur usage médical. Cette plate-forme, très bien équipée fournit un environnement hautement efficace et normalisé pour se conformer aux conditions générales d'observation des évaluations subjectives fixées par les recommandations ITU-BT.500-13 [12]. Ainsi, des séances de notation subjectives ont été menées dans ce «living lab», où les médecins regardaient des vidéos médicales (encodées à 5 taux de compression AVC/HEVC différents et résolutions FHD/SD) sur un écran FHD standard 42". Ils notaient chaque vidéo (séquence originale et codée) à l'aide d'un outil numérique sur une tablette, pendant 2x24 minutes. Pour plus d'informations sur le processus de test, le lecteur pourra se référer aux travaux [4, 5].

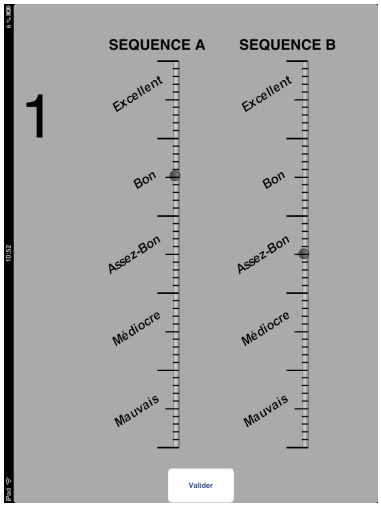


Figure 2 – *Echelle continue : Application de notation GUI sur tablette (Mauvais=0, Excellent=1)*

— 2.2.3 Analyse de la base de donnée des observateurs :

Comme cela est recommandé dans la norme ITU-BT.500-13 [12], certaines séquences ont été doublées et mélangées aléatoirement. Cela permet de faire une évaluation initiale de la cohérence des observateurs et, si la même personne, lors de la même session, donne des notations trop différentes pour la même séquence, elle sera écartée. Un second test [12] est également effectué pour les personnes restantes afin de normaliser la capacité des panélistes à répondre de manière cohérente par rapport à l'ensemble du panel. S'il s'avère qu'un observateur répond systématiquement différemment que le reste du panel, il sera également rejeté. Une fois ces étapes terminées, nous avons une base de données contenant les notes d'observateurs jugées cohérentes. Cela nous permet de définir la note d'opinion moyenne MOS (Mean Opinion Score) 1, représentant pour chaque séquence vidéo le score moyen des observateurs. MOS est donné par :

$$\bar{u_j}k = 1/N_{obs} \sum_{i=1}^{N_{obs}} u_{ijk} \quad (1)$$

Où N_{obs} est le nombre d'observateurs et u_{ijk} est la note de l'observateur i correspondant au taux de compression AVC/HEVC j de la séquence vidéo k. Ce score moyen

2. PROMETEE : PeRceptiOn utilisateur pour les usages du MultimÉdia dans les applications mÉdicalEs. La plate-forme est située TELECOM Nancy, Ecole d'Ingénieur en Sciences du Numérique de l'Université de Lorraine (France).

d'opinion est l'unité de perception subjective de qualité obtenue pour un panel d'observateurs qui ont réalisé un test strictement identique. Dans le contexte DSCQS, nous avons calculé le DMOS (MOS différentiel), c'est-à-dire la différence des notes attribuées à la séquence d'origine et aux séquences encodées.

2.3 2.3 Evaluation de la qualité objective

Malgré la pertinence des tests subjectifs dans l'évaluation de la qualité des vidéos médicales, cette méthode est encore très coûteuse en termes de temps et de ressources humaines. Comme alternative, nous pouvons utiliser des mesures objectives appropriées, à condition qu'elles soient fortement corrélées à la perception humaine, ici à la perception des experts médicaux. Jusqu'à récemment, ces mesures étaient limitées à celles de faible performance, telles que le PSNR. Les recherches menées récemment ont conduit à la mise en place d'outils psychovisuels qui ont permis de mieux comprendre le comportement du système visuel humain (HVS) et d'affiner les modèles associés. Ainsi, des mesures efficaces pour l'évaluation objective de la qualité sont apparues ces dernières années comme SSIM [13], PSNR-HVS [14] et HDR-VDP [15] par exemple. Dans cette étude, nous comparons un ensemble de métriques de qualité objectives aux scores des participants recueillis lors des tests subjectifs. La plupart d'entre elles sont disponibles dans la bibliothèque « Matlab Metrix Mux » [16]. Certaines métriques efficaces récentes ont été ajoutées à cette bibliothèque comme PSNR-HVS [14] et PSNR-HVS-M [14], qui tentent de modéliser le HVS. Enfin, nous avons mis en œuvre des métriques sans référence (BRISQUE [17] et NIQE [18]), ce qui permet de mesurer la présence d'artefacts de compression comme «effet de blocage» traditionnel liés à la mise en œuvre du DCT (transformée cosinus discrète). Le lecteur peut se référer à [19] pour plus de détails sur toutes ces mesures. Dans la section 3, nous présenterons les résultats.

3 Résultats expérimentaux

Notre étude est basée sur un panel de 16 observateurs (5 femmes et 11 hommes) ayant différentes années d'expérience dans un programme de médecine ORL (interne, externe, résident, médecin, professeur). Le tableau 2 résume les données des observateurs en termes d'années d'expérience.

Années d'expérience	[1 ; 5]	[6; 10]	[11 ; 25]
Nombre d'observateurs	7	4	5

Tableau 2 – *Caractéristiques des observateurs de notre étude*

Notons que dans le test décrit ici, nous avons trouvé deux observateurs non cohérents avec le reste du panneau. Ainsi, dans ce qui suit, les résultats s'appuient sur les 14 autres observateurs.

3.1 Performances en qualité de x265 vs x264

Tout d'abord, nous nous concentrons sur la résolution vidéo FHD. Pour mesurer la qualité subjective des vidéos encodées, nous représentons la courbe de l'évolution du score DMOS par rapport au débit binaire de compression AVC / HEVC. Nous avons interpolé les points (notes subjectives des médecins) en utilisant la fonction « pchip » (« Piecewise Cubic Hermite Interpolating Polynomial ») comme représenté sur la figure 3. Ensuite, nous déterminons le seuil de qualité en choisissant $DMOS_{min} = +0.1$ (10% de l'échelle de notation), une valeur minimale estimée comme une variation qui ne modifie pas la qualité technique à des fins médicales. En d'autres termes, nous considérons que les observateurs tolèrent la "qualité médicale" de la vidéo compressée lorsque la valeur DMOS est inférieure à 10%. Ainsi, cette valeur nous permet de trouver le débit minimum de compression AVC / HEVC qui peut être utilisé pour coder ce type de vidéo médicale.

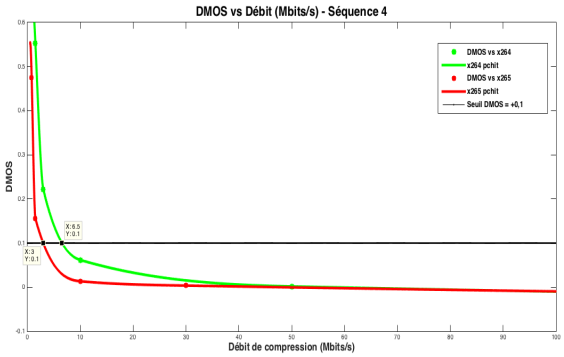


Figure 3 – *DMOS en fonction de débit de compression (AVC/HEVC) (Séquence 4 de la figure 1 en FHD)*

Comme le montre la figure 3, nous notons que l'encodage vidéo HEVC est meilleur que celui d'AVC en termes de qualité. Fixer le $DMOS_{min} = +0.1$ comme seuil nous conduit à avoir 6,5 Mbits/s comme limite de débit de compression AVC. A ce même seuil de DMOS, nous pouvons voir que HEVC offre un débit de compression égal à 3 Mbits/s, ce qui représente environ 50% du gain de débit par rapport à AVC. Les différents seuils de débit de compression pour toutes les séquences sont présentés dans le tableau 3 où l'on voit la supériorité d'HEVC par rapport à AVC.

Séquence	1	2	3	4
HEVC seuils (Mbits/s)	4.19	6.53	5	3
AVC seuils (Mbits/s)	5.15	7.49	7.3	6.5

Tableau 3 – *Débits de compression (FHD) pour un seuil de DMOS=+0,1*

3.2 Comportement de qualité sur les vidéos de résolutions SD vs Full HD

En plus des résultats précédents, ce test subjectif nous a permis de comparer le comportement de qualité de la com-

pression HEVC / AVC sur les vidéos de résolution FHD / SD. En raison des contraintes de temps de protocole ITU-BT.500-13 (<1 heure), le test a été effectué uniquement sur la séquence 1, jugée par les chirurgiens comme la plus pertinente parmi les 4 séquences. Nous pouvons voir, sur la figure 4, que cette séquence vidéo médicale peut être encodée en utilisant une résolution SD au lieu de FHD à faible débit (<0,96 Mbits/s). Ce résultat est très intéressant pour le projet E3, montrant que nous pouvons passer de FHD à la résolution SD dans un contexte de transmission à faible débit. Passer de FHD à la résolution SD permettra de réduire le traitement du temps de codage. Ceci est d'une importance majeure pour les terminaux mobiles avec l'encodeur HEVC utilisé en temps réel.

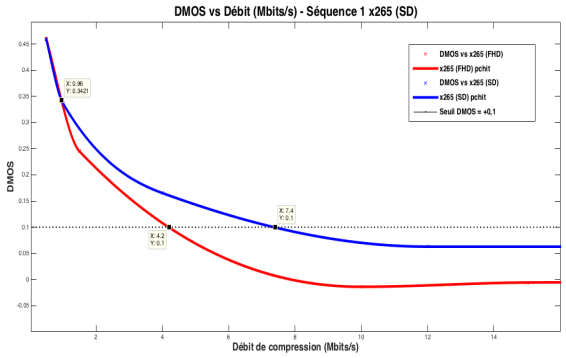


Figure 4 – *DMOS en fonction du débit de compression (HEVC – FullHD /SD) – Séquence 1 de la figure 1*

3.3 3.3 Corrélation avec les métriques objectives

Comme nous l'avons indiqué à la section 2.3, les mesures objectives peuvent être une alternative à l'évaluation subjective de la qualité. Cependant, nous devons trouver des critères appropriés pour être fortement corrélés à la perception des médecins. Ainsi, nous avons mesuré les corrélations entre les métriques objectives, mentionnées dans la section 2.3 et le DMOS afin de définir laquelle de ces mesures serait la plus appropriée pour l'évaluation de la qualité de la vidéo médicale ORL. Pour chacune d'elles, on calcule le coefficient d'évaluation de corrélation de Pearson (coefficient de corrélation linéaire LCC), indiquant la qualité de la régression linéaire. Plus le LCC est proche de 1, plus la corrélation entre les notes subjectives et objectives est meilleure. Nous présentons les différents résultats de LCC dans le tableau 4.

Sur la base du tableau 4, nous pouvons conclure que les critères objectifs les plus efficaces en termes de corrélation avec DMOS sont MSE, SSIM / MSSIM et NQM dans notre étude. MSE est une mesure simple, qui calcule l'erreur quadratique moyenne entre l'original et la vidéo encodée. Il est donc probable que les médecins soient plus sensibles à la qualité globale de la vidéo médicale (quantité de bruit dans la vidéo) qu'à la structure spécifique de l'image. Les chirurgiens ORL regardent la vidéo endoscopique en temps réel pendant qu'ils opèrent. Ainsi, il semble qu'ils se concentrent plutôt sur la qualité de la vidéo globale. Par contre, d'autres spécialistes, tels que les radiologues, devront analyser en détail toutes les parties de l'image et avoir le temps de faire le diagnostic. NQM (Mesure de qualité du bruit) semble être également efficace pour ce type d'images, car la vision humaine est sensible à la variation de la luminance et du contraste. De plus, SSIM et MSSIM ont de bons classements en raison de leur approche structurelle. Si nous comparons ce résultat à la dernière étude [4, 5], nous trouvons les mêmes résultats montrant que, en plus des derniers critères, NIQE (Naturalness Image Quality Evaluator) et BRISQUE sont parmi les indicateurs les plus corrélés avec les valeurs DMOS pour un codage AVC, car ils mesurent les artefacts d'encodage tels que l'effet de bloc. Cependant, ces deux dernières métriques ne sont pas efficaces en corrélation avec les valeurs DMOS pour HEVC, confirmant l'efficacité des 3 filtres utilisés par cette norme.

PEARSON	Moy(AVC)	Rg(AVC)	Moy(HEVC)	Rg(HEVC)
SSIM	**0,9236**	**5**	**0,9065**	**4**
UQI	0,8879	9	0,8809	6
PSNR	0,8651	10	0,8652	8
WSNR	0,8892	8	0,8839	5
VSNR	0,8215	12	0,8510	10
HDRVDP	0,9063	7	0,8720	7
IFC	0,7089	16	0,7827	16
MSE	**0,9757**	**2**	**0,9580**	**1**
MSSIM	**0,9282**	**6**	**0,9126**	**3**
NIQE	0,9855	**1**	0,8004	15
NQM	**0,9499**	**3**	**0,9513**	**2**
PSNRHVS	0,8363	11	0,8557	9
PSNRHVSM	0,8176	13	0,8458	11
VIF	0,7464	15	0,8126	14
VIFP	0,7951	14	0,8308	12
BRISQUE	**0,9437**	**4**	0,8263	13

Tableau 4 – *Moyenne Pearson LCC entre les mesures objectives et subjectives pour les 4 séquences vidéo médicales en FHD*

4 Conclusion et perspectives

Dans cet article, nous avons comparé les performances de compression AVC/H.264 et HEVC en termes de qualité dans un cadre médical sensible. Comme prévu, le nouveau standard de compression HEVC est plus efficace que AVC/H.264. Nous avons montré que l'encodeur x265 (HEVC) permettait de réduire les débits de compression. En effet, dans le cadre de notre application, pour un DMOS fixe = + 0.1, le gain en débit de compression varie entre 13% et 54% par rapport à x264 (encodeur AVC). Un test supplémentaire a été effectué en comparant le comportement de compression avec les deux résolutions SD et FHD, montrant que nous pouvons encoder les vidéos médicales ORL à une résolution SD au lieu d'utiliser la résolution Full HD dans un contexte de faible débit (<1Mbits/s). Ces résultats sont confirmés par des mesures objectives appropriées. En conclusion, HEVC peut être une solution efficace pour une transmission à faible débit, en particulier pour les réseaux mobiles ou les réseaux à faible bande pas-

sante dans les zones moins favorisées en termes de réseau. Il peut donc être un outil approprié pour les scénarios de télémédecine comme ceux conçus dans le projet européen E3 ou plus généralement, pour la télémédecine sur des réseaux avec une bande passante contrainte. Dans le futur proche, nous avons l'intention de généraliser ces résultats à d'autres types de vidéos médicales.

Remerciements

Cette étude a été menée dans le cadre du projet Européen Celtic E3. En conséquence, nous sommes profondément reconnaissants à tous les partenaires E3. Les résultats de ce travail s'appuient fortement sur la plate-forme PROMETEE, référencée au programme Innovation de l'Institut Français des Mines-Télécom en 2011, ce qui nous a permis d'effectuer des tests subjectifs dans les conditions recommandées. Enfin, nous remercions chaleureusement le Dr Gallet et tous les employés du service ORL de CHRU Nancy, qui nous ont permis d'avoir et d'utiliser des séquences endoscopiques issues de chirurgies ORL réels et authentiques et à tous les observateurs qui ont participé à cette étude. À savoir, nous tenons à remercier : Dr J. Huang, Dr G. Koch, Dr A. Bolzer, Dr B. Toussaint, Dr DT Nguyen, Dr T. De Saint Hilaire, Dr J. Chauvelot , Dr A. Bey, Dr J. Rebois, Dr A. Russel, Dr C. Rumean, Dr L. Dhaine, Dr Y. Abu-Shama, Dr L. Coffinet, Dr S. Botti .

Références

[1] N. Nouri, D. Abraham, J. M. Moureaux, M. Dufaut, J. Hubert, et M. Perez. Subjective MPEG2 compressed video quality assessment : Application to Telesurgery. Dans *2010 IEEE International Symposium on Biomedical Imaging : From Nano to Macro*, pages 764–767, Avril 2010.

[2] Yann Gaudeau et Jean-Marie Moureaux. Lossy compression of volumetric medical images with 3d dead-zone lattice vector quantization. *annals of telecommunications - annales des télécommunications*, 64(5-6) :359–367, Juin 2009.

[3] Canadian Association of Radiologists. Car standards for irreversible compression in digital diagnostic within radiology, Juin 2011.

[4] A. Chaabouni, Y. Gaudeau, J. Lambert, J. M. Moureaux, et P. Gallet. Subjective and objective quality assessment for H264 compressed medical video sequences. Dans *2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–5, Octobre 2014.

[5] A. Chaabouni, Y. Gaudeau, J. Lambert, J. M. Moureaux, et P. Gallet. H.264 medical video compression for telemedicine : A performance analysis. *IRBM*, 37(1) :40–48, Février 2016.

[6] Manzoor Razaak et Maria G. Martini. Rate-distortion and Rate-quality Performance Analysis of HEVC Compression of Medical Ultrasound Videos. *Procedia Computer Science*, 40 :230–236, Janvier 2014.

[7] H. Schwarz, D. Marpe, et T. Wiegand. Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9) :1103–1120, Septembre 2007.

[8] http ://www.videolan.org/developers/x264.html.

[9] G. J. Sullivan, J. R. Ohm, W. J. Han, et T. Wiegand. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12) :1649–1668, Décembre 2012.

[10] http ://x265.org/.

[11] Didier Nicholson, Piotr Pawałowski, et Jean-Marie Moureaux. Selected medical imaging sequences for HEVC development.

[12] ITU-R. Recommendation 500-13. Methodology for the subjective assessment of the quality of television pictures, ITU-R Rec – BT.500, 2012, Janvier 2012.

[13] Zhou Wang, A. C. Bovik, H. R. Sheikh, et E. P. Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4) :600–612, Avril 2004.

[14] Z. Wang, E. P. Simoncelli, et A. C. Bovik. Multiscale structural similarity for image quality assessment. Dans *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, Novembre 2003.

[15] Rafat Mantiuk, Kil Joong Kim, Allan G. Rempel, et Wolfgang Heidrich. HDR-VDP-2 : A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions. Dans *ACM SIGGRAPH 2011 Papers*, SIGGRAPH '11, pages 40 :1–40 :14, New York, NY, USA, 2011. ACM.

[16] M Gaubatz. Metrix Mux visual quality Assessment-Package.

[17] A. Mittal, A. K. Moorthy, et A. C. Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12) :4695–4708, Décembre 2012.

[18] A. Mittal, R. Soundararajan, et A. C. Bovik. Making a #x201c ;Completely Blind #x201d ; Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3) :209–212, Mars 2013.

[19] A. K. Moorthy, L. K. Choi, A. C. Bovik, et G. de Veciana. Video Quality Assessment on Mobile Devices : Subjective, Behavioral and Objective Studies. *IEEE Journal of Selected Topics in Signal Processing*, 6(6) :652–671, Octobre 2012.

INDICE D'ÉVALUATION AVEC RÉFÉRENCE DE LA QUALITÉ DES MAILLAGES 3D BASÉ SUR LA SAILLANCE VISUELLE

Anass Nouri, Christophe Charrier, Olivier Lézoray

Normandie Université, UNICAEN, ENSICAEN, GREYC UMR CNRS 6072, Caen, France

RÉSUMÉ

Dans ce papier, nous proposons un nouvel indice perceptuel indépendant de la vue pour l'évaluation avec référence de la qualité des maillages 3D. Celui-ci s'inspire de la méthode proposée par Wang *et al.* [1] comparant les informations structurelles du maillage de référence et du maillage dégradé. Afin d'extraire ces informations structurelles, nous utilisons une carte de saillance multi-échelle à partir de laquelle nous calculons des statistiques locales.

Les résultats expérimentaux attestent de la correlation importante entre les scores de qualité objectifs fournis par notre indice et les scores de qualité fournis par les observateurs humains. Nous comparons également notre approche avec celles proposées dans l'état de l'art et concluons que cette dernière est très compétitive.

Mots-clés— Evaluation objective de la qualité, maillage 3D, Saillance visuelle, graphes.

1. INTRODUCTION

Avec le développement des techniques d'acquisition 3D, une large quantité d'objets 3D est représentée, dans la plupart des cas, sous la forme de maillages 3D triangulaires qui sont utilisés dans plusieurs paysapplications s'articulant dans des domaines différents.

Ce progrès couplé avec le fait que l'être humain se base fortement sur sa vision, requiert que les maillages 3D représentant les cibles à analyser ou à traiter soient de bonnes qualité. Un maillage 3D peut être amené à subir plusieurs traitements avant d'être présenté à un observateur humain comme la compression, le tatouage, le lissage, etc. Dès lors, à supposer qu'une ou plusieurs des distorsions précédemment énumérées soient appliquées, une évaluation de la qualité perceptuelle devient nécessaire pour quantifier l'impact visuel de ces distorsions sur la géométrie du maillage présenté au consommateur final, et qui est en l'occurence l'observateur humain.

Une première approche pour l'évaluation de la qualité des maillages 3D s'inscrit dans l'évaluation subjective de la qualité lors des campagnes psychovisuelles permettant de collecter les jugement humains. Cependant, cette approche d'évaluation subjective de la qualité est lente, fastidieuse et inadéquate pour les applications réelles.

Une approche alternative s'inscrit dans l'évaluation objective de la qualité ayant pour objectif la prédiction de la qualité perçue d'une manière algorithmique. Cette approche permet de calculer des scores de qualité qui doivent łtre fortement corrélées avec les scores des observateurs humains. Dans la littérature, les métriques d'évaluation de la qualité sont regroupées en trois grandes familles : les mthodes avec référence (la version de référence est intégralement disponible pour la comparaison), les algorithmes avec référence réduite (des informations partielles du maillage de référence sont disponibles) et les approches sans référence (aucune information relative au maillage de référence n'est disponible).

Il est important de noter que notre métrique s'intègre dans la catégorie de l'évaluation avec référence de la qualité et qu'il est plus adequat d'utiliser le terme de *mesure de similarité* ou *mesure de fidélité* au lieu de *métrique de qualité* puisque l'objectif final revient à mesurer le degré de conformité du maillage dégradé au maillage de référence.

Les métriques perceptuelles jouent un rôle significatif dans plusieurs applications graphiques telles que l'optimisation et l'évaluation des performances des algorithmes de compression et de restauration, la mise en place de benchmarks autours des algorithmes de traitement des maillages 3D, etc. Dans ce contexte, nous proposons une métrique d'évaluation avec référence de la qualité basée sur la saillance visuelle nommée SMQI (Saliency-based Mesh Quality Index). En effet, la saillance visuelle permet de détecter les régions perceptuellement importantes sur lesquelles l'attention visuelle humaine est focalisée. Ainsi, si les régions perceptuellement importantes sur la surface du maillage sont dégradées, alors la qualité perçue globale du maillage est affectée et vice-verça (voir figure 1).

La suite du papier est organisée de la manière suivante : la section 2 présente l'état de l'art. La section 3 décrit le lien entre la saillance visuelle et l'évaluation de la qualité. Dans la même section, nous présentons le synopsis de la métrique proposée et les détails associés : la carte de saillance visuelle multi-échelle, la carte de rugosité et la distance perceptuelle utilisée. Dans la section 4, nous présentons les résultats des expérimentations en prenant en compte deux bases de maillages évaluées subjectivement et comparons notre approche avec l'état de l'art.

2. ETAT DE L'ART

Alors que les métriques d'évaluation de la qualité des images 2D sont très développées, la littérature est moins importante dans le domaine de l'informatique graphique. L'utilisation des métriques 2D pour les maillages 3D n'est pas pertinente comme démontré par Rogowitz et Rushmeir [2]. Ceci est principalement dû à la non prise en compte de la profondeur et du mouvement des maillages 3D. Par conséquent, nous citons ici uniquement les métriques indépendantes de la vues performant directement sur la surface du maillage 3D. Dans [3], Corsini *et al.* proposent une métrique basée sur les variation de la rugosité globale. La rugosité est calculée avec la variance des angles dièdres. Lavoué *et al.* [4] propose une extension de la métrique 2D SSIM vers les maillages 3D. Des différences de statistiques sont calculées à partir des cartes de courbures des deux maillages comparés au lieu des intensités des pixels utilisés par la métrique SSIM. Dans [5], Lavoué propose une amélioration de la métrique MSDM nommée MSDM2. Cette fois, l'aspect multi-échelle est pris en compte et une étape d'association de noeuds est intégrée dans le pipeline de la métrique pour pouvoir évaluer des maillages de différentes connectivités. Wang *et al* [6] proposent une métrique basée sur les variations de la rugosité locale dérivée du Laplacien de la courbure gaussienne. Torkhani *et al.* [10] proposent une métrique basée sur la différence des tenseurs de courbure. A la différence de MSDM2, celle-ci considère non seulement les amplitudes du tenseur mas également ses directions principales. Aucune approche de létat de l'art ne prend en compte la saillance visuelle qui pourtant représente une information primordiale pour systme visuel humain dans la section de l'information visuelle.

3. LA MÉTRIQUE SMQI

3.1. L'attention visuelle et l'évaluation de la qualité

À chaque regard en direction d'une scène ou d'un objet, notre attention visuelle se fixe sur des régions particulières distinctes de leurs voisinages. Ces zones, essentiellement proéminentes dans le contexte d'un objet 3D sont dépendantes du contenu de l'objet ou de la scène et indépendantes du comportement ou du vécu de l'observateur [7]. Ceci a été le constat de deux études dans lesquelles des séries d'expérimentations psychovisuelles ont confirmé qu'une dégradation est davantage perçue lorsqu'elle est située sur une région saillante du contenu [8]. Le même résultat peut être constaté sur la figure 1. Trois métriques significatives de l'état de l'art ([4], [5] et [10]) ont été testée dans ce cas de figure et ont échoué à évaluer la qualité du contenu d'une manière similaire à la perception humaine. Ceci est principalement dû à la non prise en compte de l'information relative à la saillance.

3.2. Synopsis de la métrique

L'indice de qualité proposé est inspiré de la métrique 2D SSIM [1] et de la métrique MSDM [4] pour les maillages

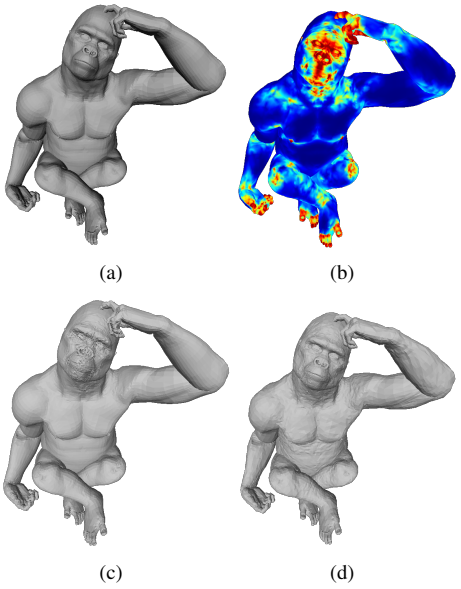


Fig. 1: Comparaison de la qualité de différents maillages 3D. (a) Maillage 3D Gorille de référence. (b) Carte de saillance de (a) avec [11]. (c) Maillage 3D Gorille bruité sur les zones saillantes. FMPD=0.15, MSDM2=0.36, SMQI=0.49. (d) Maillage 3D Gorille bruité sur des zones moins saillantes. FMPD=0.54, MSDM2=0.414, SMQI=0.41. (Notons qu'un score de qualité important fait référence à une mauvaise qualité et vice-verça). SMQI est l'indice proposé.

3D. Cependant, à la différence des métriques SSIM calculant les statistiques locales à partir des intensités des pixels et MSDM calculant les statistiques locales reflétant les informations structurelles à partir d'une carte de courbures, nous proposons de calculer une carte de saillance multi-échelle qui sera utilisée comme base pour le calcul des statistiques locales sur les surfaces des deux maillages. En effet, nous supposons que la qualité d'un maillage 3D est fortement associée à la modification de la saillance locale et globale du maillage 3D. De plus, pour les deux maillages, nous considérons une carte de rugosité à partir de laquelle nous calculons les differences de la moyenne locale de rugosité de chaque noeud. Ceci nous permet de prendre en compte le masquage visuel qui peut avoir lieu lorsqu'une région rugueuse est apte à dissimuler une distorsion. Par la suite, nous introduisons quatre fonctions de comparaison entre les voisinages locaux correspondants sur les surfaces des deux maillages afin d'évaluer les différences de structures. Finalement, nous combinons ces fonctions en utilisant la somme de Minkowski pondérée pour obtenir le score de qualité final.

3.3. Carte de saillance multi-échelle

Afin de d'estimer les régions perceptuellement importantes sur la surface d'un maillage 3D, nous utilisons notre approche de détection de la saillance proposée dans [11]. C'est une mesure de saillance visuelle multi-échelle basée sur une car-

actéristique importante du système visuel humain (SVH) qui est la forte sensibilité aux fortes fluctuations et aux fortes discontinuités. Nous présentons dans cette section les principales étapes de cette mesure de saillance. Premièrement, nous représentons un maillage $\mathcal{M}$ par un graphe non orienté $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ où $\mathcal{V} = \{v_1, \ldots, v_N\}$ est l'ensemble des nœuds N et $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ représente l'ensemble des arêtes. L'ensemble des arêtes est déduit à partir des faces connectant les nœuds. Pour chaque nœud v_i sont associés des coordonnées 3D $\mathbf{p}_i = (x_i, y_i, z_i)^T \in \mathbb{R}^3$. Pour un nœud 3D cible sur la surface du maillage, nous estimons son plan tangent 2D. Ensuite nous considérons un voisinage sphérique de rayon ε centré autour du nœud cible et projetons sur le plan 2D tous les nœuds voisins localisés dans la sphère. Une fois les noeuds projetés, nous définissons la taille du patch relatif au nœud cible en calculant la distance maximale entre les projections 2D le long des axes x et y. Par la suite, nous divisons le patch en un nombre de cellule l et les incrémentons par le champ des hauteurs de projections des nœuds voisins compris dans la sphère. L'étape suivante consiste à calculer les similarités entre le patch du nœud cible et les patchs associés aux nœuds voisins. Ces similarités sont affectées aux poids des arêtes reliant les les nœuds voisins au nœud cible :

$$w_{\mathcal{P}}(v_i, v_j) = \exp\left[-\frac{\kappa(v_j) \cdot ||\boldsymbol{\mathcal{P}}(v_i) - \boldsymbol{\mathcal{P}}(v_j)||_2^2}{\sigma_{\mathcal{P}}(v_i) \cdot \sigma_{\mathcal{P}}(v_j) \cdot ||\mathbf{p}_i - \mathbf{p}_j||_2^2}\right] \quad (1)$$

où $\boldsymbol{\mathcal{P}}(v_i) \in \mathbb{R}^{l \times l}$ est le vecteur des hauteurs cumulées dans les cellules du patch, $\kappa(v_j)$ est la courbure du nœud v_j, et $||\mathbf{p}_i - \mathbf{p}_j||_2^2$ représente la distance Euclidienne entre les nœuds v_i et v_j. Nous proposons de calculer localement le paramètre d'échelle $\sigma_{\mathcal{P}}(v_i)$ par $\sigma_{\mathcal{P}}(v_i) = \max_{v_k \sim v_i}(||\mathbf{p}_i - \mathbf{p}_k||_2)$. Par conséquent, la saillance visuelle mono-échelle du nœud cible est définie son degré moyen :

$$\text{Saillance mono-échelle}_{\mathcal{P}}(v_i) = \frac{1}{|v_j \sim v_i|} \sum_{v_i \sim v_j} w_{\mathcal{P}}(v_i, v_j) \quad (2)$$

Afin d'améliorer la qualité de la mesure de saillance, nous calculons celle-ci sur trois échelles en variant le rayon ε du voisinage sphérique lors de la constructions des patchs locaux adaptatifs. La saillance multi-échelle d'un nœud v_i, $MS(v_i)$, est définie par la moyenne des valeurs de saillance mono-échelles pondérées par leurs entropies respectives. La figure 1(b) présente un exemple de résultat de saillance multi-échelle.

3.4. Carte de rugosité et distance perceptuelle

Afin de calculer les statistiques locales reflétant les informations structurelles d'un maillage 3D, nous utilisons la carte de saillance multi-échelle décrite. Pour un voisinage local $N(v_i)$ représentant l'ensemble des nœuds adjacents au nœud cible v_i sur la surface du maillage, nous définissons la moyenne locale de la saillance et l'écart-type local de la saillance notés respectivement $\mu_{N(v_i)}$ et $\sigma_{N(v_i)}$ par :

$$\mu_{N(v_i)} = \frac{1}{|N(v_i)|} \sum_{v_j \in N(v_i)} MS(v_j) \quad (3)$$

$$\sigma_{N(v_i)} = \sqrt{\frac{1}{|N(v_i)|} \sum_{v_j \in N(v_i)} (MS(v_j) - \mu_{N(v_i)})^2} \quad (4)$$

où $|N(v_i)|$ est la cardinalité du voisinage $N(v_i)$. Pour deux voisinages correspodants $x = N_{\mathcal{M}_1}(v_i)$ et $y = N_{\mathcal{M}_2}(v_i)$ de deux maillages 3D $\mathcal{M}_1$ et $\mathcal{M}_2$, nous définissons la covariance σ_{xy} par :

$$\sigma_{xy} = \frac{1}{|x|} \sum_{v_j \in x,y} (MS_{\mathcal{M}_1}(v_j) - \mu_x)(MS_{\mathcal{M}_2}(v_j) - \mu_y) \quad (5)$$

où $MS_{\mathcal{M}_1}$ et $MS_{\mathcal{M}_2}$ representent respectivement les cartes de saillance multi-échelles des deux maillages $\mathcal{M}1$ et $\mathcal{M}2$.

Ensuite, similairement à [4], nous définissons trois fonctions de comparaisons entre deux voisinages correspondants x et y sur les deux maillages $\mathcal{M}_1$ et $\mathcal{M}_2$ pour quantifier la déformation affectant les informations structurelles du maillage dégradé :

$$L(x, y) = \frac{||\mu_x - \mu_y||_2^2}{max(\mu_x, \mu_y)} \quad (6)$$

$$C(x, y) = \frac{||\sigma_x - \sigma_y||_2^2}{max(\sigma_x, \sigma_y)} \quad (7)$$

$$S(x, y) = \frac{||\sigma_x \sigma_y - \sigma_{xy}||_2^2}{\sigma_x \sigma_y} \quad (8)$$

où L, C et S font référence respectivement à la fonction de comparaison de la saillance, la fonction de comparaison du contraste, et la fonction de comparaison de structure. Ayant ces trois fonctions de comparaison, nous avons remarqué que le masquage visuel n'était pas suffisamment pris en compte lorsqu'une région rugueuse était présente sur la surface du maillage. En effet, ayant une surface rugueuse et une surface lisse, une distorsion serait davantage visible sur la surface lisse que sur la surface rugueuse dans la mesure o une surface rugueuse masque la distorsion. Pour prendre en compte l'effet du masquage visuel, nous avons utilisé une carte de rugosité basée sur le Laplacien de la courbure Gaussienne [12]. Nous avons introduit ensuite une quatrième fonction basée sur la comparaison de la moyenne locale de la rugosité. L'objectif de cette fonction est d'induire une large différence lorsqu'une région lisse devient une région rugueuse :

$$R(x, y) = \frac{||\delta_x - \delta_y||_2^2}{max(\delta_x, \delta_y)} \quad (9)$$

où $\delta_x = \frac{1}{|x|} \sum_{v_j \in x} RoughnessMap(v_j)$ o $RoughnessMap$ reprsente la carte de rugosit. Il est important de noter qu'une carte de saillance est différente d'une carte de rugosité dans la mesure où la carte de saillance fait ressortir uniquement les informations nouvelles et non-redondantes. Finalement, notre indice d'évaluation avec référence de la qualité (SMQI)

entre deux maillages 3D $\mathcal{M}_1$ et $\mathcal{M}_2$ est défini par la somme de Minkowsky de leurs distances locales pondérée :

$$SMQI(\mathcal{M}_1, \mathcal{M}_2) = \left(\frac{1}{|V|}\sum^{|V|} L(x,y)\right)^{\alpha} + \left(\frac{1}{|V|}\sum^{|V|} C(x,y)\right)^{\beta} + \left(\frac{1}{|V|}\sum^{|V|} S(x,y)\right)^{\gamma} + \left(\frac{1}{|V|}\sum^{|V|} R(x,y)\right)^{\delta} \quad (10)$$

où α, β, γ et δ sont obtenus à partir d'une optimisation basée sur les algorithmes génétiques. Les détails associés à cette optimisation sont présentés dans la section suivante.

4. RÉSULTATS EXPÉRIMENTAUX

4.1. Bases de maillages

Pour comparer la métrique proposée avec les méthodes de l'état de l'art, deux bases de maillages évaluées subjectivement sont utilisées : 1) Liris/Epfl General-Purpose [4] et 2) Liris-Masking [15]. La première base contient quatre maillages de référence. Ces derniers sont affectés par deux types de distorsions : un bruit additif et un lissage. Ces distorsions sont appliquées suivant trois degrés d'intensité sur différentes régions du maillage : 1) uniformément sur la surface du maillage, 2) spécifiquement sur les zones rugueuses ou lisses du maillage (pour la simulation de l'effet du masquage) et 3) spécifiquement sur les zones de transitions entre les zones rugueuses et les zones lisses. Au total, 22 maillages 3D dégradés de chaque maillage de référence sont générés et évalués par 12 observateurs humains. La base de maillages Liris Masking consiste en 4 maillages de référence dégradés par un bruit additif suivant trois degrés d'intensités sur les zones lisses ou rugueuses pour générer 6 versions dégradées de chaque maillage de référence. 12 observateurs humains ont évalué cette base. La performance de notre méthode est mesurée par le coefficient de corrélation de Spearman (SROCC : Spearman Rank Ordered Correlation Coefficient). Le choix de la corrélation de Spearman est motivée par la non vérification de la normalité des scores objectifs pour pouvoir calculer la correlation avec les scores subjectifs.

4.2. Resultats

Avant d'évaluer la performance de notre approche et étant donné que notre métrique dépend de quatre paramètres indépendants (α, β, γ et δ) dont l'optimisation empirique ou manuelle serait difficile et inefficace, nous avons choisi d'utiliser une optimisation basée sur un algorithme génétique pour les fixer. Il est important aussi de souligner que le nombre de maillages 3D que contiennent les deux bases de maillages subjectivement évaluées (et décrites pralablement) sont de tailles très réduites en comparaison des bases d'images 2D. Pour faire face à ce problème, nous effectuons un apprentissage de type Leave-One-Out sur les deux bases de maillages. L'objectif est alors d'effectuer un apprentissage du modèle sur $k-1$ observations et de le valider sur la k-ème. Ce processus est répété k x 1000 fois. Dans notre contexte, une observation fait référence aux MOS (Mean-Opinion-Score) d'un maillage 3D de référence et de ses versions dégradées. La fonction de fitness utilisée pour effectuer l'optimisation génétique est définie par : $f(\alpha,\beta,\gamma,\delta) = \sqrt{\sum_{i=0}^{k-1}(MOS_i - SMQI_i(\mathcal{M}_1,\mathcal{M}_2))^2}$ où MOS_i est le vecteur des valeurs MOS de l'observation i et $SMQI_i(\mathcal{M}_1,\mathcal{M}_2)$ représente la distance perceptuelle calculée avec l'equation 10. Après l'optimisation, nous obtenons : $\alpha = 23.63$, $\beta = 3.26$, $\gamma = 5.04$ et $\delta = 0.77$. Notons que sur la figure 1, le maillage 3D Gorille a été évalué avec ces paramètres.

Le tableau 1 présente la performance de notre approche en terme de corrélation de Spearman avec les scores subjectifs fournis par la base Liris/Epfl General Purpose. Nous pouvons remarquer que notre métrique SMQI produit d'importantes valeurs de corrélation pour tous les maillages 3D et plus particulièrement pour les maillage Venus et RockerArm où les valeurs de corrélations sont les plus élevées. Le résultat de la régression psychométrique entre les scores objectifs et subjectifs est présenté sur la figure 2(a) et confirme ce résultat pour le maillage 3D Venus. La fonction choisie dans notre étude est la fonction cumulative de la loi normale Gaussienne :

$$g(m,n,R) = \frac{1}{\sqrt{2\pi}}\int_{m+nR}^{\infty} e^{-t^2}dt \quad (11)$$

où m et n sont estimés avec une régression non-linéraire par la méthode des moindres carrés et R représente le score objectif. De plus, il apparait que la métrique $SMQI$ est classée deuxième meilleure métrique derrière $TPDM$ en ce qui concerne la corrélation sur toute la base de maillages (les valeurs de corrélations sont 89.6% pour $TPDM$, 84.6% pour $SMQI$ et 80.4% pour FMPD). Cette corrélation importante associée la base Liris/EPFL General-Purpose peut être confirmée par la courbe présentée sur la figure 2(b) où les points $SMQI - MOS$ sont très proches de la courbe psychométrique. Cette courbe indique aussi une forte capacité de généralisation sur tous les maillages 3D.

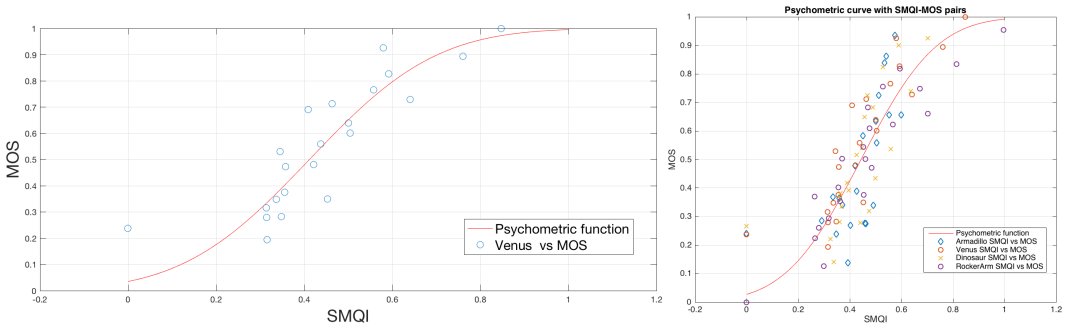
Nous avons également testé et comparé notre métrique avec les métriques de l'état de l'art sur la base de maillages Liris-Masking. Le tableau 2 présente les valeurs de corrélation de Spearman des différentes métriques sur cette base de maillage. A partir de ces résultats, trois observations peuvent êtres formulées. La première est que SMQI est très compétitive avec TPDM et MSDM2 et réussit à prendre en compte le masquage visuel. La deuxième observation concerne la légère infériorité de la valeur de corrélation de Spearman associée au maillage 3D Lion en comparaison avec les valeurs de corrélations de FMPD et MSDM2. Ceci pourrait être expliqué par le fait que la carte de saillance multi-échelle du maillage dégradé sur laquelle sont calculées les

Liris/Epfl General-Purpose	HD	RMS	3DWPM1	3DWPM2	MSDM2	FMPD	TPDM	SMQI
Armadillo	69.5	62.7	65.8	74.1	81.6	75.4	**84.9**	77.5
Venus	1.6	90.1	71.6	34.8	89.3	87.5	90.6	**91.6**
Dinosaure	30.9	0.3	62.7	52.4	85.9	89.6	**92.2**	84.8
RockerArm	18.1	7.3	87.5	37.8	89.6	88.8	**92.2**	91.8
Base entière	13.8	26.8	69.3	49.0	80.4	81.9	**89.6**	84.6

Table 1: Valeurs de corrélation de Spearman (%) des différentes métriques sur la base de maillages Liris/Epfl General Purpose database

Liris Masking	HD[2][3]	RMS[2][3]	3DWPM1	3DWPM2	MSDM2	FMPD	TPDM	SMQI
Armadillo	48.6	65.7	58.0	48.6	**88.6**	**88.6**	**88.6**	**88.6**
Lion-vase	71.4	71.4	20.0	38.3	**94.3**	**94.3**	82.9	83.0
Bimba	25.7	71.4	20.0	37.1	**100.0**	**100.0**	**100.0**	**100.0**
Dinosaure	48.6	71.4	66.7	71.4	**100.0**	94.3	**100.0**	**100.0**

Table 2: Valeurs de corrélation de Spearman (%) des différentes métriques sur la base de maillages Liris-Masking



(a) (b)

Fig. 2: Fonction de régression psychométrique tracée pour les paires SMQI-MOS appliquée sur les maillages de référence et les maillages dégradés du corpus Liris/Epfl General-purpose : (a) Maillage 3D Venus et (b) Corpus intégral.

statistiques locales ne reflète pas convenablement les zones saillantes dégradées. En effet, le maillage de référence et le maillage dégradé utilisent le même rayon (défini empiriquement) des voisinages sphériques lors du calcul des patchs locaux adaptatifs pour l'estimation de la carte de saillance multi-échelle. Nous pensons qu'une méthode automatique définissant un rayon propre à chaque maillage mènera à une meilleure estimation de la carte de saillance pour le maillage dégradé. Ceci constitue une de nos perspectives. La troisième observation se rapporte aux valeurs de corrélation sur tout le corpus de la base Liris-Masking. Nous ne présentons pas les résultats de corrélation sur tout le corpus car comme confirmé dans [16], le protocole utilisé lors des évaluations subjectives spécifiait un référentiel d'évaluation différent pour chaque maillage et par conséquent les valeurs de corrélation sur tout le corpus des maillages 3D ne sont pas significatives. A partir des résultats et comparaisons précédents, il apparait que la métrique SMQI est fortement corrélée à la perception humaine en raison de l'intégration de la saillance et du masquage visuel à la fois. En sus, SMQI est classée deuxième en terme de corrélation sur la base Liris/Epfl General-Purpose et est concurrentielle sur la base Liris-Masking.

5. CONCLUSION

Nous avons proposé dans ce papier une nouvelle métrique d'évaluation avec référence de la qualité des maillages 3D appelée SMQI. Cet indice compare les information structurelles d'un maillage 3D de référence avec sa version dégradée. Pour cela, nous utilisons une carte de saillance multi-échelle sur laquelle nous calculons les statistiques locales reflétant les structures du maillage. Dans le but de prendre en compte l'effet du masquage visuel, une carte de rugosité est calculée pour mesurer les différences des moyennes de rugosité. Par conséquent, nous combinons quatre fonctions de comparaisons en utilisant la somme de Minkowski pondérée afin de fournir un score quantifiant la similarité visuelle entre deux maillages. Les résultats expérimentaux ainsi que les comparaisons avec les méthodes de l'état de l'art ont montré la forte corrélation de notre approche avec les scores de qualité fournis par les observateurs humains et attesté de sa forte compétitivité.

6. REFERENCES

[1] Z. Wang, A.C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE T IMAGE PROCESS*, vol. 13, no. 4, pp. 600–612, 2004.

[2] Bernice E. Rogowitz and Holly E. Rushmeier, "Are image quality metrics adequate to evaluate the quality of geometric objects?," in *Proc. SPIE*, 2001, vol. 4299, pp. 340–348.

[3] M. Corsini, E.D. Gelasca, T. Ebrahimi, and M. Barni, "Watermarked 3D mesh quality assessment," *IEEE T MULTIMEDIA*, vol. 9, no. 2, pp. 247–256, Feb 2007.

[4] G. Lavoué, E. Drelie Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, "Perceptually driven 3D distance metrics with application to watermarking," in *Proc. SPIE*, 2006, vol. 6312, pp. 63120L–63120L–12.

[5] G. Lavoué, "A multiscale metric for 3D mesh visual quality assessment," *Computer Graphics Forum*, vol. 30, no. 5, pp. 1427–1437, 2011.

[6] K. Wang, F. Torkhani, and A. Montanvert, "Technical section: A fast roughness-based approach to the assessment of 3D mesh visual quality," *Comput. Graph.*, vol. 36, no. 7, pp. 808–818, Nov. 2012.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[8] F. Boulos, B. Parrein, P. Le Callet, and D. Hands, "Perceptual effects of packet loss on H.264/AVC encoded videos," in *VPQM workshop*, 2009.

[9] U. Engelke, R. Pepion, P. Le Callet, and H Zepernick, "Linking distortion perception and visual saliency in h.264/avc coded video containing packet loss," in *Proc. SPIE*, 2010.

[10] Fakhri Torkhani, Kai Wang, and Jean-Marc Chassery, "A curvature-tensor-based perceptual quality metric for 3d triangular meshes," *Machine Graphics and Vision*, pp. 1–25, 2014.

[11] A. Nouri, C. Charrier, and O. Lézoray, "Multi-scale mesh saliency with local adaptive patches for viewpoint selection," *Signal Processing: Image Communication*, vol. 38, pp. 151–166, 2015.

[12] K. Wang, F. Torkhani, and A. Montanvert, "Technical section: A fast roughness-based approach to the assessment of 3D mesh visual quality," *Comput. Graph.*, vol. 36, no. 7, pp. 808–818, Nov. 2012.

[13] N. Aspert, Diego S. Cruz, and T. Ebrahimi, "Mesh: Measuring error between surfaces using the hausdorff distance," in *ICME*, 2002, vol. 1, pp. 705–708.

[14] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: Measuring Error on Simplified Surfaces," *Computer Graphics Forum*, vol. 17, no. 8, pp. 167–174, 1998.

[15] G. Lavoué, "A local roughness measure for 3D meshes and its application to visual masking," *ACM Trans. Appl. Percept.*, vol. 5, no. 4, pp. 21:1–21:23, Feb. 2009.

[16] G. Lavoué and M. Corsini, "A comparison of perceptually-based metrics for objective evaluation of geometry processing," *IEEE T MULTIMEDIA*, vol. 12, no. 7, pp. 636–649, Nov 2010.

Un réseau neuronal convolutif pour l'évaluation de la qualité des maillages 3D

I. Abouelaziz[1] A. Chetouani [2] M. El Hassouni[1] H. Cherifi[3]

[1] LRIT, Associated Unit to CNRST (URAC No 29)- Faculty of Sciences,
Mohammed V University in Rabat, B.P.1014 RP, Rabat, Morocco.
[2] University of Orleans – PRISME laboratory - Orleans, France
[3]LE2I UMR 6306 CNRS, University of Burgundy, Dijon, France.

ilyass.abouelaziz@gmail.com
aladine.chetouani@univ-orleans.fr
mohamed.elhassouni@gmail.com
hocine.cherifi@u-bourgogne.fr

Résumé

Dans cet article, nous proposons une nouvelle méthode sans référence pour l'évaluation de la qualité des objets 3D basée sur une phase d'apprentissage. Pour ce faire, nous extrayons d'abord des caractéristiques visuelle en calculant les angles dièdres et la courbure locale du maillage dégradé. Ensuite, nous déterminons à partir de ces fonctionnalités un ensemble de patches 2D qui sont appris via un réseau neuronal convolutif (CNN) qui se compose de deux couches convolutives et de deux couches de max-pooling. Un perceptron multicouches (MLP) avec deux couches entièrement connectées est ensuite intégré pour résumer la représentation apprises vers un unique nœud de sortie. Avec cette structure, l'apprentissage et la régression sont utilisés pour prédire le score de qualité d'un maillage déformé sans avoir besoin du maillage de référence. Les expériences sont menées sur la base de données à utilisation générale (LIRIS general-purpose database) et la base de masquage (LIRIS masking database). Les résultats montrent de bons taux en termes de corrélation avec les scores de jugement humain.

Mots clefs

Réseau neuronal convolutif (CNN), évaluation sans référence de la qualité des objets 3D , système visuel humain, la courbure moyenne, les angles dièdres.

1 Introduction

La qualité perçue de maillages 3D est le résultat de différentes opérations liées à la transmission et le traitement géométrique (tatouage, simplification, compression, ...) [1, 2]. La qualité perceptuelle d'un maillage 3D est subjectivement défini comme la moyenne des évaluations effectuées par des sujets humains (MOS : Mean Opinion Score). Cependant, l'évaluation subjective est coûteuse, laborieuse et prend un certain temps. Les méthodes objectives d'évaluation sont la solution la plus adéquate pour évaluer automatiquement la qualité visuelle [3].

Le problème de l'évaluation de la qualité visuelle des maillage 3D a connu des progrès considérables au cours des dernières années. Les premiers travaux ont utilisé des similarités simples entre le maillage de référence et sa version déformée telle que l'erreur quadratique moyenne (RMS) [4] et la distance de Hausdorff (HD) [5]. Ce type de méthodes a généralement échoué car il calcule une distance géométrique négligeant les opérations principales du système visuel humain (SVH) [6]. Afin d'intégrer l'information perceptuelle, plusieurs méthodes utilisent différents principes pour une meilleure estimation de la qualité visuelle perçue. Dans [7], une métrique perceptuelle basée sur l'analyse de courbure appelée mesure de distorsion structurelle de maillage (MSDM) a été proposée. Afin d'évaluer la qualité des maillages tatoués, Corsini et *al.* ont développé une métrique perceptuelle utilisant la variation de rugosité [8]. Une autre mesure perceptuelle appelée FMPD (Fast Mesh Perceptual Distance) a été proposée dans [9]. Cette métrique est basée sur une mesure de rugosité locale dérivée de la courbure gaussienne. Ces méthodes, avec référence, atteignent une corrélation très élevée avec la perception humaine. Cependant, leur principal inconvénient est l'indisponibilité du maillage de référence dans les applications réelles.

La capacité de prévoir automatiquement la qualité est un problème difficile, en particulier dans de nombreuses applications pratiques de vision par ordinateur lorsque la référence n'est pas disponible. Étant donné uniquement le maillage déformé, l'approche sans référence tente de prédire la qualité perçue et assure une bonne corrélation avec les jugements humains sans prendre en compte les modèles de référence. Ce concept est largement abordé dans l'évaluation de la qualité de l'image, et plusieurs méthodes évaluent avec succès le score de qualité des images dégradées en exploitant les méthodes d'apprentissage [10]. Cependant, dans l'évaluation de la qualité des maillages 3D, il existe un manque distinct de méthodes sans réfé-

rence. L'exploitation de ce concept serait un gain important, en particulier avec les résultats prometteurs obtenus dans notre travaille précédent concernant l'évaluation de la qualité à l'aide d'un modèle SVR (support vector regression) [11]. Motivé par les observations ci-dessus, nous proposons dans cet article une méthode sans référence basée sur une phase d'apprentissage. L'apprentissage profond a été largement utilisé pour la tâche d'évaluation de la qualité des images [12, 13]. De bonnes performances ont été obtenues grâce à sa capacité à apprendre des caractéristiques discriminantes.

La méthode proposée utilise des patches 2D calculés à partir des caractéristiques extraites de la courbure et des angles dièdre, l'apprentissage est basée sur un réseau neuronal convolutif (CNN) pour prédire la qualité perçue des maillages 3D déformées avec différents types de distorsion.

Ce document est organisé comme suit. Dans la section. 2, nous décrivons la méthode proposée. Les résultats expérimentaux et la discussion sont présentés dans la Section. 3, pour finir avec la conclusion dans la section. 4.

2 Méthode proposée

2.1 Méthodologie

Les différentes étapes de la méthode proposée d'évaluation de la qualité visuelle des maillages 3D sont décrites par la Fig. 1. Tout d'abord, deux caractéristiques perceptuelles sont extraites : la courbure et les angles dièdres. Ces caractéristiques sont ensuite réorganisées en patches 2D afin d'adapter l'entrée à notre modèle CNN. Après cela, une architecture CNN est proposée suivie d'une méthode de régression. Le modèle de régression formé estime le score de qualité pour chaque patch. Le score global de qualité est obtenu par la moyenne des scores prédits.

2.2 Extraction des caractéristiques et préparation des patches 2-D

Dans notre travail, nous visons à apprendre une représentation de maillage compacte et efficace à partir de caractéristiques bas niveau. Par conséquent, deux types de caractéristiques perceptuelle sont extraites : courbure moyenne et angles dièdres. La courbure moyenne est une caractéristique perceptuelle importante représentant l'aspect visuel d'un maillage 3D. Elle décrit la quantité d'écart d'une surface plane et fournit plusieurs caractéristiques visuelles d'un modèle 3D, en particulier l'acuité, la rugosité ou la douceur d'une région. L'aspect structurel du maillage 3D est représenté par l'angle dièdre qui est utilisé pour construire le concept de rugosité globale. Ces caractéristiques géométriques sont largement utilisées dans de nombreuses applications de traitement de maillage [15, 9] et peuvent décrire le maillage 3D à partir des différentes perspectives. Les caractéristiques de bas niveau extraites peuvent être concaténées dans un vecteur de caractéristiques d'une dimension élevée. Cependant, cette représentation peut conduire à un ajustement excessif. En outre, cela ne correspond pas à la propriété convolutive des modèles CNN. Ainsi, nous proposons de réorganiser le vecteur de caractéristiques en patches 2D.

2.3 Architecture du modèle CNN utilisé

L'architecture du modèle CNN proposé est composée de sept couches, telles que représentées dans la Fig. 2.

Les différentes couches sont présentées comme suit :

- **Input :** patches 2-D de taille 28×28.
- **Couche de convolution 1 :** la première couche est une couche convolutionnelle qui filtre le patch d'entrée avec 32 noyaux. Chaque noyau est de taille (5×5). Cette couche fournit 32 cartes de caractéristiques de taille 24×24.
- **Couche de max-pooling 1 :** la deuxième couche est une couche de pooling qui applique le processus de max-pooling sur chaque carte de caractéristiques avec une fenêtre locale de taille 2×2. Cette couche produit 32 cartes de caractéristiques de dimension de 12×12.
- **Couche de convolution 2 :** la troisième couche est une autre couche convolutionnelle qui filtre la sortie de la couche max-pooling avec 32 noyaux de taille 5×5. Cette couche produit 32 cartes de caractéristiques de taille 8×8.
- **Couche de max-pooling 2 :** la quatrième couche est une autre couche max-pooling avec une fenêtre locale de taille 8×8. En conséquence, cette couche produit un vecteur de taille 1×32.
- **Couches entièrement connectées :** la cinquième et sixième couches sont deux couches entièrement connectées de 250 nœuds chacune.
- **Couche de sortie :** la septième couche est une couche de régression linéaire avec une sortie unidimensionnelle qui fournit le score de qualité.

2.4 Apprentissage et estimation de la qualité

Notre réseau est appris sur des patches 2D sans chevauchement de taille (28×28) obtenus à partir des caractéristiques extraites des maillages 3D. Nous attribuons pour chaque patch un score identique à celui du maillage original. De même que dans [17], nous adoptons la fonction d'apprentissage définie comme suit :

$$L = \frac{1}{N} \sum_{n=1}^{N} ||S(p_n; \omega) - MOS_n||_{l1} \qquad (1)$$

$$\hat{\omega} = \min_{\omega} L$$

Avec MOS_n est le score moyen d'opinion assigné à un patch d'entrée donné p_n. $S(p_n; \omega)$ est le score prédit de p_n avec des poids du réseau ω.

Les paramètres du réseau neuronal convolutif sont appris en utilisant une descente de gradient stochastique (SGD) et la technique de rétro-propagation en minimisant la fonction objective définie dans Eq. 1. Dans nos expériences, nous

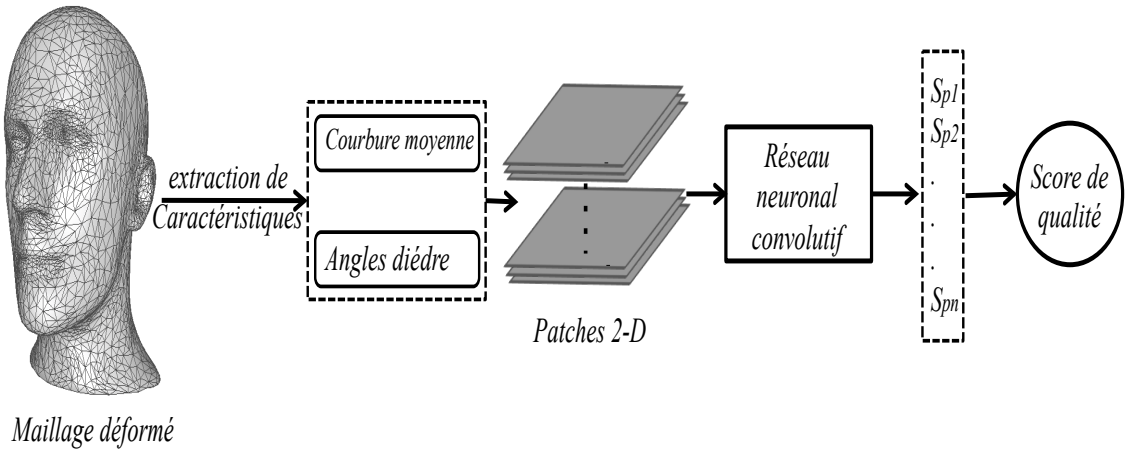


Figure 1 – *Schéma général de la méthode proposée.*

effectuons une descente de gradient stochastique pendant 40 epochs.
Dans la phase de test, le score de qualité pour un patch donné est obtenu en utilisant les paramètres du modèle qui assurent la meilleure corrélation avec les MOS. Enfin, le score global de qualité pour un maillage déformé donné est la moyenne de tous les scores des patches.

3 Résultats expérimentaux

3.1 Bases de données et protocole de validation

Pour tester la performance d'une méthode de qualité visuelle, une base de données de maillages dégradés notés par des observateurs humains est nécessaire. Notre méthode d'évaluation de qualité a été testée et validée à l'aide de deux bases de données accessibles au public spécialement conçus pour l'évaluation des méthodes de qualité.

– La base de données LIRIS Masking [18] qui contient 4 modèles de référence et 24 modèles déformés obtenus par l'addition locale de bruit avec différents niveaux.
– La base de données General-purpose [19] qui contient 4 modèles de référence et 84 modèles déformés obtenus par l'addition locale du bruit et le lissage avec différents niveaux.

Notez que la représentation de maille utilisée dans notre méthode fournit un nombre important de patches 2D qui rendent l'ensemble de données assez important pour le processus d'apprentissage.

Pour évaluer la performance de la méthode proposée, deux coefficients de corrélation sont couramment utilisés, à savoir le coefficient de corrélation linéaire de Pearson (r_p) (précision de prédiction) et le coefficient de corrélation de Spearman (r_s) (monotonie de prédiction) [20].

3.2 Résultats et discussion

Pour évaluer la performance de la méthode sans référence proposée, une comparaison a été effectuée avec les méthodes décrites dans la littérature : HD [5], RMS [4], 3DWPM2 [8], MSDM2 [7] and FMPD [9]. Le tableau 1 présente les coefficients de corrélation r_s et r_p des méthodes comparées obtenus sur les deux bases de données considérées. Les corrélations sur l'ensemble du **corpus** sont calculées entre les scores objectifs de tous les objets dans le corpus et leurs MOS correspondants . On peut remarquer que les méthodes basées sur les distances géométriques HD et RMS ne reflètent pas la qualité perçue et ne correspondent pas bien avec la perception humaine. D'autre part, les méthodes perceptuelles MSDM2, FMPD, 3DWPM2 et la méthode proposée atteignent des corrélations élevées et montrent une bonne performance dans l'évaluation de la qualité perçue.

En ce qui concerne la base de données de masquage LIRIS, nous pouvons constater que la méthode proposée a le score r_s le plus élevé (88.2%) et le deuxième meilleur score r_p (85.4%). Ainsi, il surpasse deux des méthodes les plus efficaces dans l'état de l'art qui sont MSDM2 et FMPD. En outre, notre méthode produit des corrélations élevées pour chaque maillage individuellement, notamment pour Armadillo (r_s (95.2%) et r_p (97.6%)).

Les bonnes performances de la méthode proposée sont aussi constatées sur la base à usage général. En particulier, le score r_s qui le plus élevé sur toute la base (83.6%) et les scores les plus élevées pour les modèles Armadillo, Venus et Rocker. En comparaison avec la base de données de masquage LIRIS, la base de données à usage général contient un nombre important de modèles dégradés (21 versions dégradées pour chaque modèle ainsi qu'une variété de types de distorsion).

4 Conclusion

Dans cet article, nous avons présenté un réseau neuronal convolutif (CNN) pour l'évaluation sans référence de la qualité visuelle des maillages 3D . Le réseau est ali-

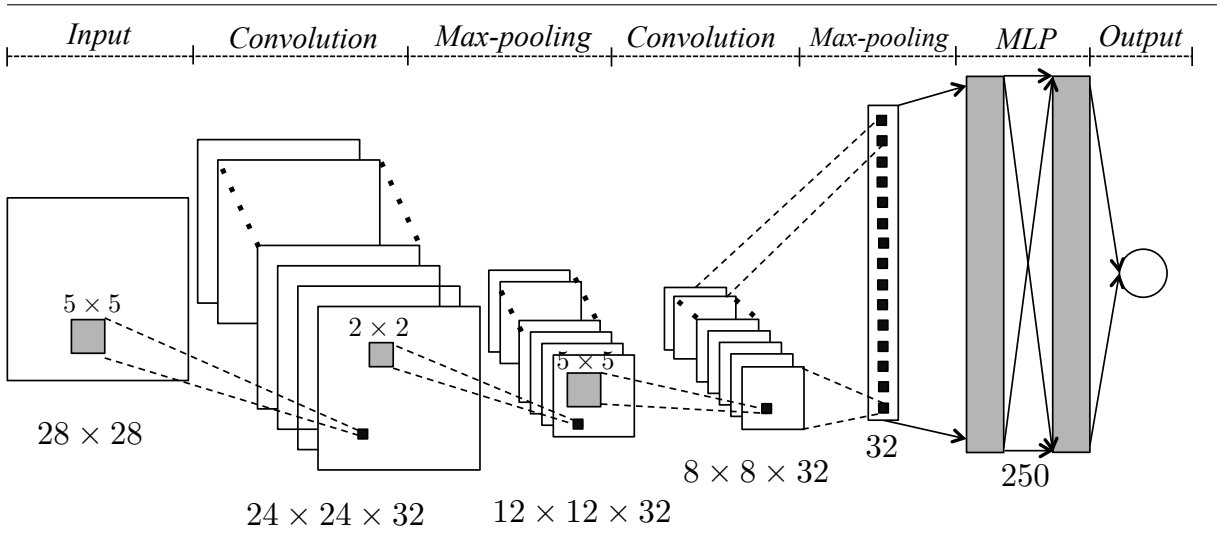


Figure 2 – *Configuration du réseau neuronal convolutif pour l'évaluation de la qualité des maillage 3D.*

Tableau 1 – *Les coefficients de corrélation r_s (%) et r_p (%) de différentes méthodes objectives.*

Base de données	Méthodes	HD [5]		RMS [4]		3DWPM2 [8]		MSDM2 [7]		FMPD [9]		Méthode proposée	
		r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p	r_s	r_p
LIRIS masking	Armadillo	48.6	37.7	65.6	44.6	48.6	37.9	81.1	88.6	94.2	88.6	95.2	97.6
	Lion	71.4	25.1	71.4	23.8	38.3	22.0	93.5	94.3	93.5	94.3	89.4	91.6
	Bimba	25.7	7.5	71.4	21.8	37.1	14.4	96.8	100	98.9	100	93.4	98.7
	Dyno	48.6	31.1	71.4	50.3	71.4	50.1	95.6	100	96.9	94.3	96.3	89.9
	Répertoire entier	26.6	4.1	48.8	17.0	37.4	18.2	87.3	89.6	80.8	80.2	88.2	85.4
General purpose	Armadillo	69.5	30.2	62.7	32.3	74.1	43.1	81.6	85.3	75.4	83.2	87.2	84.3
	Dyno	30.9	22.6	0.3	0.0	52.4	19.9	85.4	85.7	89.6	88.9	86.4	86.2
	Venus	1.6	0.8	90.1	77.3	34.8	16.4	89.3	87.5	87.5	83.9	92.2	85.6
	Rocker	18.1	5.5	7.3	3.0	37.8	29.9	89.6	87.2	88.8	84.7	91.3	85.2
	Répertoire entier	13.8	1.3	26.8	7.9	49.0	24.6	80.4	81.4	81.9	83.5	83.6	82.7

menté par des caractéristiques perceptuelles extraites des maillages 3D et disposées en patches 2-D pour répondre aux exigences de notre modèle CNN. L'architecture proposée est composée de plusieurs couches de convolution et de max-pooling. En outre, la couche MLP avec deux couches entièrement connectées est intégrée pour résumer la représentation et produire le score final de la qualité. Les résultats expérimentaux ont prouvé que le réseau formé prédit avec succès la qualité visuelle. Il convient de noter que la méthode proposée est sans référence et ne nécessite pas le maillage de référence. Contrairement aux méthodes de référence complètes et réduites concurrentes, notre méthode peut être utile dans des situations pratiques.

Références

[1] Garland, Michael, and Paul S. Heckbert. "Surface simplification using quadric error metrics." Proceedings of the 24th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 1997.

[2] Wang, Kai, et al. "A comprehensive survey on three-dimensional mesh watermarking." IEEE Transactions on Multimedia 10.8 (2008) : 1513-1527.

[3] Lavoué, Guillaume, and Massimiliano Corsini. "A comparison of perceptually-based metrics for objective evaluation of geometry processing." IEEE Transactions on Multimedia 12.7 (2010) : 636-649.

[4] Cignoni, Paolo, Claudio Rocchini, and Roberto Scopigno. "Metro : Measuring error on simplified surfaces." Computer Graphics Forum. Vol. 17. No. 2. Blackwell Publishers, 1998.

[5] Aspert, Nicolas, Diego Santa-Cruz, and Touradj Ebrahimi. "Mesh : Measuring errors between surfaces using the hausdorff distance." Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on. Vol. 1. IEEE, 2002.

[6] Breitmeyer, Bruno G. "Visual masking : past accomplishments, present status, future developments." Advances in cognitive psychology 3.1-2 (2007) : 9.

[7] Lavoué, Guillaume, et al. "Perceptually driven 3D distance metrics with application to watermarking." SPIE Optics+ Photonics. International Society for Optics and Photonics, 2006.

[8] Corsini, Massimiliano, et al. "Watermarked 3-D mesh quality assessment." IEEE Transactions on Multimedia 9.2 (2007) : 247-256.

[9] Wang, Kai, Fakhri Torkhani, and Annick Montanvert. "A fast roughness-based approach to the assessment of 3D mesh visual quality." Computers & Graphics 36.7 (2012) : 808-818.

[10] Li, Chaofeng, Alan Conrad Bovik, and Xiaojun Wu. "Blind image quality assessment using a general regression neural network." IEEE Transactions on Neural Networks 22.5 (2011) : 793-799.

[11] Abouelaziz, Ilyass, Mohammed El Hassouni, and Hocine Cherifi. "No-reference 3D mesh quality assessment based on dihedral angles model and support vector regression." International Conference on Image and Signal Processing. Springer International Publishing, 2016.

[12] Hou, Weilong, et al. "Blind image quality assessment via deep learning." IEEE transactions on neural networks and learning systems 26.6 (2015) : 1275-1286.

[13] Zhang, Wei, et al. "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network." Pattern Recognition 59 (2016) : 176-187.

[14] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 5 2015.

[15] Gal, Ran, and Daniel Cohen-Or. "Salient geometric features for partial shape matching and similarity." ACM Transactions on Graphics (TOG) 25.1 (2006) : 130-150.

[16] Corsini, Massimiliano, et al. "Watermarked 3-D mesh quality assessment." IEEE Transactions on Multimedia 9.2 (2007) : 247-256.

[17] Mittal, Anish, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain." IEEE Transactions on Image Processing 21.12 (2012) : 4695-4708.

[18] Lavoué, Guillaume, et al. "Perceptually driven 3D distance metrics with application to watermarking." SPIE Optics+ Photonics. International Society for Optics and Photonics, 2006.

[19] Lavoué, Guillaume. "A local roughness measure for 3D meshes and its application to visual masking." ACM Transactions on Applied perception (TAP) 5.4 (2009) : 21.

[20] Wang, Zhou, and Alan C. Bovik. "Modern image quality assessment." Synthesis Lectures on Image, Video, and Multimedia Processing 2.1 (2006) : 1-156.

Nouvelle approche d'estimation de la qualité sans référence des images compressées JPEG2000

A. Chetouani[1] A. Beghdadi[2]

[1]PRISME, Polytech'Orleans, Orléans
[2]L2TI, Institut Galilée, Université Paris XIII - Villetaneuse

aladine.chetouani@isir.upmc.fr, azeddine.beghdadi@univ-paris13.fr

Résumé

Les métriques de qualité sans référence proposées dans la littérature sont généralement développées pour un type dégradation particulier. Cette spécificité restreint ainsi considérablement le champ d'utilisation de ces mesures. Pour pallier cette limitation, nous proposons dans cette étude une métrique de qualité d'image sans référence multi-dégradations (pour l'effet de Gibbs 2D [ou "rinring"] et le flou) basées sur une pondération neuronale. Pour une image dégradée donnée, la quantité de dégradation (flou et ringing) contenue dans l'image est tout d'abord estimée via un réseau de neurones artificiel (ANN). L'index de qualité final est ensuite obtenu en fusionnant différentes métriques de qualité d'image sans référence dont les paramètres sont déduits du modèle ANN. L'efficacité de la méthode proposée a été évaluée en termes de classification et de corrélation avec l'appréciation subjective.

Mots Clef

Qualité d'image, Réseau de neurones artificiel, notes subjectives, artefacts.

Abstract

No Reference Image Quality Metrics proposed in the literature are generally developed for a specific degradation, which limits highly its application. To overcome this limitation, we propose in this study a NR-IQM for ringing and blur distortions based on neural weighting. For a given image, we first estimate the quantity of blur and ringing degradations contained in the image using an Artificial Neural Networks (ANN) model through a learning step. Then, the final index quality is given by combining a blur and a ringing metrics using to the obtained weights. The proposed method has been evaluated in terms of classification and correlation with subjective judgments.

Keywords

Image Quality, Artificial Neural Networks, Subjective Scores, Artifacts.

1 Introduction

De nos jours, l'évaluation de la qualité d'image joue un rôle important. En effet, les images subissent généralement différents traitements (acquisition, compression, transmission etc.) qui peuvent affecter lourdement sa qualité visuelle et dont, l'impact visuel varie et dépend des caractéristiques de la distorsion. Parmi les dégradations les plus courantes : les effets de blocs, le flou et l'effet de Gibbs 2D (plus connu sous le nom de « *ringing* »).

Les effets de bloc se manifestent visuellement au niveau des frontières entre blocs et apparaissent comme des contours verticaux et horizontaux dont la visibilité dépend de la distribution spatiale du signal image. La régularité de ces contours occasionne une gêne importante et rend alors cet artefact très visible. Cette distorsion est la conséquence d'un découpage de l'image en blocs et de leurs traitements de façon indépendant. C'est généralement le cas des méthodes de compression par bloc telles que JPEG ou la quantification vectorielle. Du fait de la grande popularité de ce type de méthodes de compression, elle est probablement la dégradation la plus étudiée.

Un autre artefact gênant est l'effet Gibss 2D ou « ringing », qui affecte la netteté des bords. Elle est due en général à l'étape de quantification ou de décimation des coefficients hautes fréquences et se manifeste sous forme d'oscillations au voisinage des régions à fort contraste. Elle est souvent définie comme un bruit autour de ces régions. Son niveau de visibilité varie en fonction du contraste et de son contenu fréquentiel. Plus le contraste est élevé, plus la distorsion incommode l'observateur. Au voisinage des zones texturées, ce phénomène est légèrement masqué mais perturbe la cohérence spatiale du signal image. A proximité d'une région homogène, la dégradation devient très gênante. On retrouve ce type de distorsion notamment dans les images compressées JPEG2000 [1].

Le flou est aussi un artefact gênant qui se manifeste essentiellement au niveau des détails et des transitions dans l'image. L'effet de lissage au niveau des contours et des textures qui en résulte affecte sensiblement la qualité de l'image par une diminution du contraste. Les origines de cette dégradation sont diverses et peuvent se produire à différents niveau de la chaîne d'acquisition, de traitements et de transmission tels qu'un flou de bougé, de défocalisation, un mouvement, une compression ou bien encore un filtrage de type passe-bas.

Afin d'estimer l'impact visuel de ces artefacts, différentes mesures de qualité ont été proposées dans la littérature. On distingue essentiellement trois approches : les mesures avec référence qui nécessitent l'image originale et sa version dégradée telles que SSIM [2], VIF [3]. Les mesures avec référence réduite où uniquement certains attributs de l'image originale sont exploités [4]. Et les mesures les plus attractives, les métriques sans référence, qui requièrent uniquement l'image dégradée [5, 6].

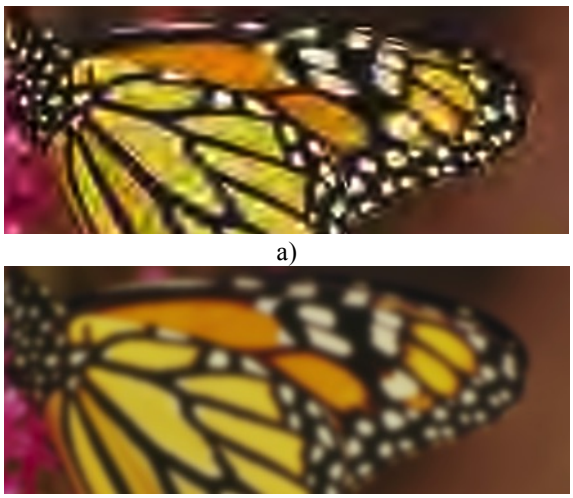
Dans cet article, notre travail se focalise sur le développement d'une approche originale d'estimation de la qualité sans référence des images compressées JPEG2000. Nous proposons une métrique de qualité efficace permettant de mesurer à la fois le flou et l'effet de « *ringing* ». En effet, les mesures de qualité d'image sans référence pour ce type d'images considèrent généralement l'effet de « *ringing* » comme étant l'unique dégradation. Cependant, le flou est aussi une dégradation inhérente de ces images et peut devenir, selon le taux de compression, la distorsion principale. Pour ce faire, nous calculons tout d'abord un poids pour chacune des distorsions considérées (flou ou « *ringing* ») par l'intermédiaire d'une modélisation de la dégradation. Ces poids représentent la quantité de dégradation contenue dans l'image dégradée. Ils sont ensuite utilisés pour estimer la qualité d'une image dégradée donnée. La modélisation de la dégradation est ici réalisée en utilisant un réseau de neurones artificiel (ANN) de type MLP (« *Multi Layer Perceptron* »). Il est à noter qu'un autre type de classifieur peut être choisi (SVM). Cependant, le point clé de ce travail n'est pas le classifieur en lui-même, mais plutôt l'approche proposée, permettant de prendre en considération les différents types de dégradations inhérents aux images compressées JPEG2000.

Cet article est organisé comme suit: Dans la section 2, nous décrivons en détails l'approche proposée. Dans la section 3, nous présentons et discutons des résultats obtenus. Nous terminons ensuite, section 4, par une conclusion et des perspectives.

2 Méthode proposée

Ce travail a été motivé par le fait que dans les images compressées JPEG2000, l'effet de « *ringing* » et le flou apparaissent simultanément dans l'image à certains taux de compression (voir figure 1). Cependant, les mesures de qualité d'image sans référence généralement proposées dans la littérature ne considèrent qu'une seule distorsion. Cet a priori limite fortement l'utilisation de ce genre de mesures. Ainsi, le fait de ne pas se restreindre à un seul type de distorsion va améliorer l'évaluation la qualité de ce type d'images.

Dans cette étude, nous proposons de pallier cette limitation en estimant pour une image dégradée donnée, le poids, en terme de quantité, de chaque dégradation considérée. Les poids obtenus sont ensuite utilisés pour calculer un indice de qualité d'image unique. Le schéma synoptique de la méthode proposée est présenté par la figure 2.



a)

b)

Fig. 1. Images dégradées : a) « ringing » et b) flou.

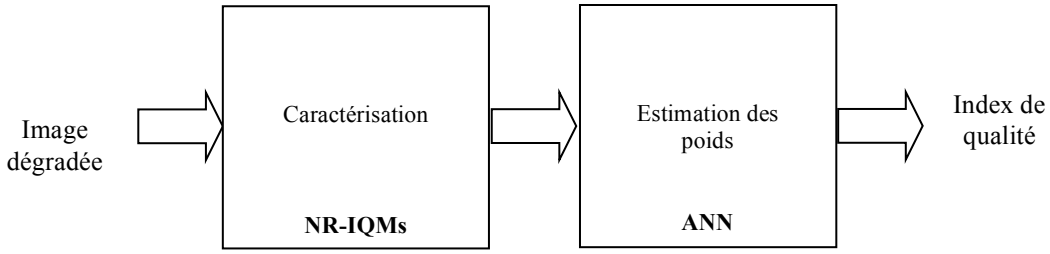


Fig. 2. Schéma synoptique de la méthode proposée.

Dans cette section, la base de données d'images utilisée est tout d'abord présentée. Puis, les descripteurs ainsi que le calcul des poids sont décrits.

2.1 Base de données d'images utilisées

Les images utilisées, à la fois pour l'apprentissage et les tests, sont issues de la base d'images LIVE (Laboratory Image and Video Engineering) [7]. Cette base de données est composée de 5 types de dégradations différents (bruit, flou, effet de bloc, « ringing » et « fast fading »). Il est à noter que nous considérons dans cette étude uniquement le flou et le « *ringing* », soit au total 401 images (227 images compressées JPEG 2000 et 174 images floues). On dispose aussi des notes subjectives moyennes, appelées MOS (Mean Opinion Scores), de chaque image. Cette base est utilisée en appliquant le principe de la validation croisée (60% pour l'apprentissage et 40% pour les tests).

2.2 Estimation des poids

C'est l'étape cruciale de l'ensemble du système proposé. Pour quantifier le flou et le « ringing » contenus dans une image dégradée donnée, différents descripteurs sont extraits et utilisés comme entrées du réseau de neurones artificiel (ANN). Les descripteurs, le réseau de neurones utilisé ainsi que la fusion des mesures de qualité d'images sans référence sont décrits dans cette section.

2.2.1 Descripteurs

Différents descripteurs peuvent être considérés pour caractériser les dégradations contenues dans l'image. Nous proposons ici d'utiliser directement des index de qualité d'image sans référence comme descripteurs. En effet, ces mesures sans dédiées généralement à une dégradation spécifique. Elles prennent ainsi en considération les caractéristiques intrinsèques de chaque dégradation.

La méthode proposée par Sheikh at al. [8], basée sur une phase d'apprentissage dans le domaine des ondelettes, a été sélectionnée. A l'issue de l'apprentissage, différents paramètres et seuils sont définis et utilisés pour décrire les caractéristiques de la dégradation. La mesure de qualité est donnée via les modèles statistiques de la distribution des coefficients pour différents niveaux de décomposition.

La deuxième métrique utilisée est aussi basée sur une décomposition en ondelettes [9]. Pour chaque niveau de décomposition, une carte de contours est tout d'abord déduite à partir des coefficients hautes fréquences. Chaque carte est ensuite analysée afin de déterminer la forme des contours de l'image, permettant ainsi d'estimer la dégradation.

La méthode proposée par Crêté et al. a aussi été considérée comme descripteur [10]. L'idée repose sur l'observation suivante: "Une image floue est moins affectée par l'ajout de flou qu'une image nette initialement". En d'autres termes, l'impact d'un flou ajouté à une image est différent selon que l'image contient initialement du flou ou non. Par conséquent, il suffit d'introduire un flou à l'image et analyser son comportement face à cette distorsion ajoutée. La mesure globale est ainsi déduite de cette analyse.

Basée sur le même principe, le quatrième descripteur sélectionné est obtenu après une analyse fréquentielle de la distribution d'énergie de l'image à évaluer et de sa version lissée [11]. La distribution spatiale du spectre d'énergie dans le plan de Fourier est analysée en utilisant la méthode décrite dans [12].

Dans [13], Marziliano et al. proposent d'analyser la distribution des intensités des pixels à proximité des points contours et d'en déduire une métrique de qualité. Une fois les points contours détectés, l'index de qualité est obtenu en estimant la longueur de chaque contour à partir des extrema locaux.

Le dernier descripteur utilisé est quant à lui déduit à partir de tests psycho-visuels [6]. L'objectif est d'estimer le seuil juste visible du flou (JNB : Juste Noticeable Blur) en fonction du changement de contraste. Pour ce faire, un stimulus uniforme sur fond gris est présenté à 16 observateurs. Il leur est ensuite demandé de modifier l'intensité du flou de façon à déterminer le seuil JNB. Une modélisation de la variation du seuil JNB en fonction du contraste permet ensuite de définir une métrique de qualité.

2.2.2 Caractéristiques du réseau de neurones artificiel

Une fois les descripteurs définis, il convient de modéliser les dégradations afin d'estimer le poids de chacune des dégradations considérées. Pour ce faire, nous avons utilisé un réseau de neurones artificiel de type Perceptron Multi Couches (MLP) avec une couche cachée (voir figure 3), dont les caractéristiques sont listés dans le tableau 1. Le nombre d'entrées correspond aux descripteurs présentés précédemment et est fixé à 6. Le nombre de sorties est égal au nombre de types de dégradations considérés, c'est à dire 2. Elles représentent le poids de chacune des distorsions (flou et « ringing »

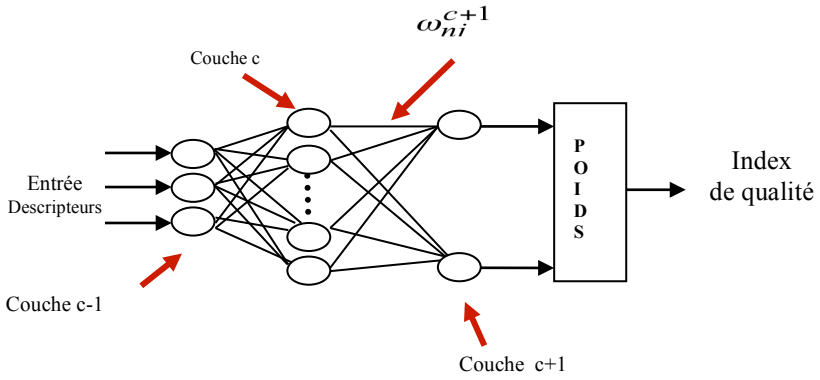


Fig. 3. Réseau de neurones artificiel utilisé

Tableau 1. Caractéristiques du réseau de neurones artificiel.

Entrées	6 (i.e. 6 descripteurs)
Couche cachée	1 avec un nombre de neurones égal à 9
Sorties	2 (flou et « ringing »)
Apprentissage	Rétro propagation
Fonction d'activation	Sigmoïde

2.2.3 Index de qualité d'image multi dégradations.

Après avoir calculé les poids de chacune des dégradations considérées, l'indice de qualité d'image sans référence est obtenu en combinant deux métriques. La mesure est donnée par l'équation suivante:

$$Index = w_b \ . \ BlurMetric \ + \ w_r \ . \ RingingMetric \qquad (1)$$

où w_b et w_r sont, respectivement, les poids obtenus pour la distorsion de type flou et de type « ringing ». La métrique *BlurMetric* utilisée est ici basée sur des tests subjectifs [14]. La mesure *RingingMetric* a été brièvement décrite dans la section précédente [8] et est basée sur une modélisation des coefficients d'ondelettes à différents niveaux de décomposition.

3 Tests et validation

Afin de tester l'efficacité de la méthode proposée, la base de données d'images LIVE, décrite précédemment, est utilisée. Nous présentons tout d'abord les résultats obtenus en termes de quantification des dégradations dans l'image à travers les poids obtenus sur différentes images dégradées. La méthode proposée est ensuite évaluée en termes de corrélation entre notre mesure de qualité et les notes d'appréciations subjectives issues de la base LIVE.

La figure 4 montre les poids obtenus pour les 5 images dégradées présentées figure 5. Il est à noter que nous nous sommes limités à 5 images pour l'affichage, mais des résultats similaires ont été constaté sur l'ensemble de la base.

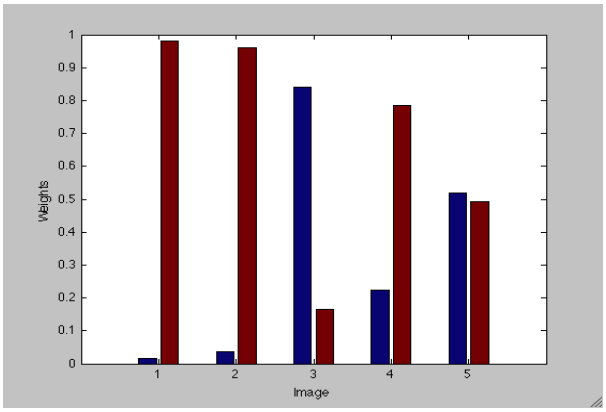


Fig. 4. Poids obtenus pour le flou (bleu) et le « ringing » (rouge).

A travers ces premiers résultats, nous pouvons clairement constater que les poids obtenus sont bien représentatifs de la distorsion contenue dans l'image. En effet, pour l'image 3 (figure 5.c), les poids obtenus sont égaux à 0,84 et 0,16 pour le flou et pour le « ringing », respectivement. En d'autres termes, cela signifie que le flou est dominant dans cette image mais qu'il y a présence de « ringing » à faible quantité. Tandis que pour l'image 1 (figure 5.a), le « ringing » est considéré comme la dégradation dominante. Effectivement, lorsque l'on présente une version zoomée de l'image, on s'aperçoit que le « ringing » est très important dans l'image (voir figure 6).

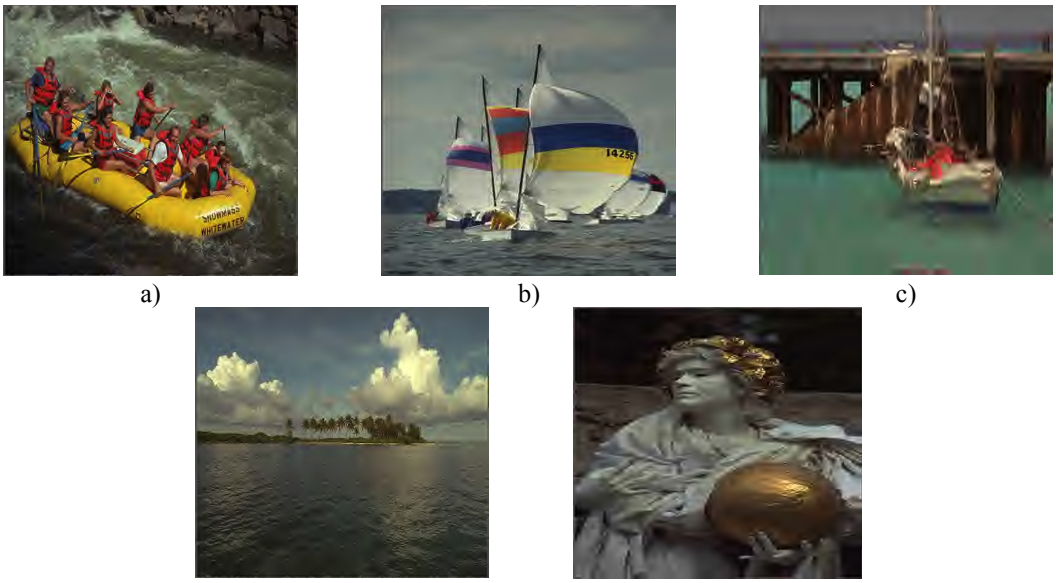


Fig. 5. a-e) Images dégradées correspondantes aux images 1-5) de la figure 4.

Fig. 6. Version zoomée de l'image figure 5.a.

La métrique proposée est ensuite évaluée en termes de corrélation avec l'appréciation subjective. Les coefficients de corrélation de Pearson et de Spearman obtenus pour les deux dégradations considérées sont présentés dans le tableau 2. Nous pouvons constater que ces corrélations sont proches de 1. En d'autres termes, la métrique proposée est en concordance avec l'appréciation subjective.

Tableau 2. Coefficients de corrélation de Pearson and Spearman.

Type de dégradation	**Corrélations de Pearson**
Ringing	0.90
Blur	0.91
Type de Dégradation	**Corrélations de Spearman**
Ringing	0.91
Blur	0.93

Ainsi, la métrique proposée peut être utilisée pour estimer la qualité des images sans référence contenant à la fois du flou et du « ringing », notamment pour les images compressées JPEG2000. Le système global est illustré par la Fig. 7.

4 Conclusion

Dans cette étude, nous avons proposé une approche originale permettant de mesurer l'impact visuel du flou et du « ringing » sans référence. Nous estimons d'abord le poids des distorsions considérées en termes de quantité et ceci à travers une modélisation neuronale. L'indice de qualité est ensuite calculé en utilisant ces poids comme paramètres de fusion. Les résultats ainsi obtenus ont permit de mettre en évidence la pertinence d'un tel schéma ainsi que son efficacité.

Bibliographie

[1] D. Taubman and M. Marcellin, "JPEG2000: Image compression fundamentals, standards and practice", *Boston, Kluwer Academic Publishers*, 2001.

[2] Z. Wang E.P. Simoncelli, and A.C. Bovik, Multi-scale structural similarity for image quality assessment. *IEEE Asilomar Conference on Signals, Systems and Computers*, Vol. 2, pp. 1398-1402, 2003.

[3] H.R. Sheikh and A.C Bovik, "Image information and visual quality", *IEEE Transactions on Image Processing*, Vol.15, pp. 430-444, 2006.

[4] Z. Wang and E.P. Simoncelli, Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. *Human Vision and Electronic Imaging X, Proc. SPIE,* Vol. 5666, pp. 149-159, 2005.

[5] A. Chetouani, G. Mostafaoui and A. Beghdadi, A New Free Reference Image Quality Index Based on Perceptual Blur Estimation. *IEEE Pacific Rim Conference on Multimedia,* pp. 1185-1196, 2009.

[6] R. Ferzli and J.L. Karam, A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur. *IEEE Transactions on Image Processing*, Vol. 18, no 4, pp. 717-728, 2009.

[7] H.R Sheikh, Z. Wang, L. Cormack and A.C. Bovik, LIVE Image Quality Assessment Database. *http://live.ece.utexas.eduesearch/quality*

[8] H.R. Sheikh, A.C. Bovik and L.K. Cormack, No-Reference Quality Assessment Using Natural Scene Statistics: JPEG2000. *IEEE Transactions on Image Processing*, Vol. 14, No. 12, pp. 1918-1927, 2005.

[9] H. Tong, M. Li, H. Zhang and C. Zhang, Blur detection for digital images using wavelet transform. *IEEE International Conference on Multimedia and Expo,* Vol. 1, pp. 17-20, 2004.

[10] F. Crête, Estimer, mesurer et corriger les artefacts de compression pour la télévision. *Thesis report, Université Joseph Fourier*, 2007.

[11] A. Chetouani, A. Beghdadi and M. Deriche, A new free reference image quality index for blur estimation in the frequency domain. *IEEE International Symposium on Signal Processing and Information Technology*, pp. 155-159, 2009.

[12] A. Beghdadi, and M. Deriche, "Features extraction from fingerprints using frequency analysis", *IEEE Workshop On Signal Processing and Applications*, Vol. 14-15, 2000.

[13] P. Marziliano, F. Dufaux, S. Winkler and T. Ebrahimi, A no-reference perceptual blur metric. *IEEE International Conference on Image Processing*, Vol. 3, pp. 57-60, 2002.

[14] N.D. Narvekar and L.J. Karam, A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD). *To appear in the IEEE Transactions on Image Processing*, 2011.

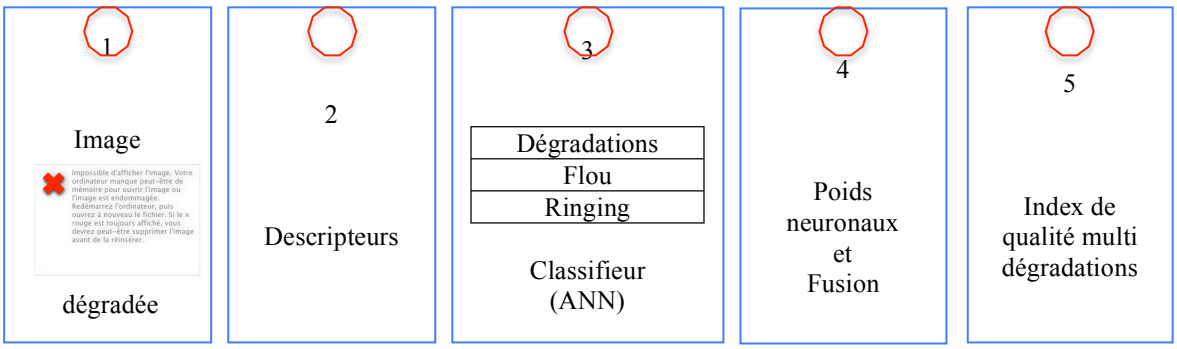


Fig. 7. Schéma global de la méthode proposée.

Biométrie, forensics et protection du contenu

Session 1

Yedroudj-Net : un réseaux de neurones efficace pour la stéganalyse spatiale

M. Yedroudj[1] M. Chaumont[1,2] F. Comby[1]

[1] UNIVERSITE MONTPELLIER, UMR5506-LIRMM, F-34095 Montpellier Cedex 5, France

[2] UNIVERSITE DE NIMES, F-30021 Nîmes Cedex 1, France

{yedroudj, chaumont, comby}@lirmm.fr

Résumé

Pendant environ 10 ans, l'approche classique pour détecter la présence d'un message secret inséré dans une image était d'utiliser un ensemble de classifieurs alimentés par des vecteurs de caractéristiques issues des images à traiter. Ces dernières années, des études telles que Xu et al. ont indiqué que des réseaux de neurones convolutionnels (CNN) bien conçus peuvent atteindre des performances comparables aux approches classiques d'apprentissage automatique.

Dans cet article, nous proposons un CNN qui dépasse les performances de l'état de l'art en terme de probabilité d'erreur de classification. La proposition s'inscrit dans la continuité de ce qui a été proposé récemment et consiste en une fusion intelligente de briques importantes proposées dans divers articles. Parmi les éléments essentiels du CNN proposé, on peut citer l'utilisation : d'un ensemble de filtres pour le prétraitement de l'image d'entrée, de la troncature comme fonction d'activation, d'au moins cinq couches convolutionnelles avec une normalisation par lot (Batch normalization layer) et une couche de mise à l'échelle (Scale layer) ainsi qu'une couche entièrement connectée correctement dimensionnée.

Mots clefs

Stéganographie, stéganalyse, réseau de neurones convolutionnel, classification.

1 Introduction

Les premières tentatives d'utilisation des méthodes de Deep Learning pour la stéganalyse remontent à 2014 avec l'utilisation d'auto-encodeurs [1]. Un an plus tard, Qian *et al.* [2] et Pibre *et al.* [3] proposent d'utiliser des CNNs pour la stéganalyse. En 2016, les premiers résultats, similaires à ceux de l'état de l'art actuel, sont obtenus avec un ensemble de CNNs [4]. Le CNN Xu-Net [5] y est alors utilisé comme brique élémentaire.

D'autres réseaux ont été proposés en 2017 pour la stéganalyse des images JPEG. Dans l'article [6], les auteurs proposent d'utiliser un pré-traitement inspiré des Spatial Rich Models (SRM) ainsi qu'une grande base de données d'apprentissage. Les résultats obtenus sont alors proches de ceux de l'état de l'art. Dans l'approche décrite dans [7], le réseau est construit en tenant compte du découpage en bloc de 8×8 dû à la compression JPEG (notion de phase). Un ensemble de CNNs est requis pour obtenir des résultats légèrement meilleurs à ceux de l'état de l'art. Dans [8], un CNN inspiré de ResNet [9], avec un système de "shortcut connection" et vingt couches, permet également d'obtenir de meilleurs résultats que ceux de l'état de l'art.

Ces résultats sont très encourageants. Toutefois, si l'on compare aux améliorations obtenues dans d'autres domaines du traitement des images utilisant le Deep Learning [10], les résultats pour la stéganalyse ne sont pas "10%" meilleurs que ceux obtenus en utilisant un ensemble de classifieurs [11] avec un SRM [12, 13] ou un SRM avec connaissance du canal de sélection [14, 15]. En 2017, les principales voies explorées pour améliorer les résultats des CNNs sont : l'utilisation d'un ensemble de CNNs et la modification de la topologie en imitant le procédé d'extraction des SRM. Dans la plupart des cas, l'effort architectural (structure du réseau) ou expérimental (taille de la base) est très élevé pour une faible amélioration des performances.

En revenant sur les bonnes pratiques du Deep Learning et les études récentes, nous avons cherché à construire, expérimentalement, un réseau plus efficace que ceux de l'état de l'art, robuste au type d'images (non compressée ou compressée JPEG...) et sans nécessité d'utiliser un ensemble de CNNs (qui est connu pour améliorer les résultats mais au prix d'une complexité accrue). Dans cet article, nous présentons un CNN conçu pour la stéganalyse dans le domaine spatial qui permet d'obtenir de bons résultats sans avoir recours à une ou plusieurs astuces pour améliorer ses performances tel que : l'apprentissage par transfert [16], l'augmentation virtuelle de la taille de la base de données [17], etc. De plus, le réseau proposé est peu sensible à l'initialisation des hyper-paramètres et converge facilement (voir Section 4). Nous nommerons ce réseau "Yedroudj-Net CNN" et comparerons ses performances avec celles de deux autres réseaux : Xu-Net [5] et Ye-Net [17], et également avec un ensemble de classifieurs [12] combiné à un SRM [11] pour la stéganalyse dans le domaine spatial.

2 Stéganographie et stéganalyse

La stéganographie est l'art de dissimuler des informations au sein d'un support de sorte que ces informations soient indétectables pour un observateur. De nos jours, les méthodes stéganographiques les plus sûres **s'adaptent au contenu**. La plupart des algorithmes intègrent les données secrètes dans les régions avec un contenu complexe où les zones d'insertion sont moins détectables. Parmi les méthodes les plus performantes d'insertion dans le domaine spatial on trouve : WOW [18] et S-UNIWARD [19].

La stéganalyse, quant à elle, est l'art qui permet de détecter et d'extraire des données dissimulées au sein d'un support. La stéganalyse peut également servir de moyen efficace pour juger des performances de sécurité des techniques de stéganographie. En d'autres termes, une bonne méthode stéganographique doit être imperceptible non seulement pour le système de vision humain, mais indétectable aussi par analyses statistiques.

Jusqu'à présent, l'état de l'art de la stéganalyse utilise l'apprentissage automatique en appliquant deux étapes :

— l'extraction de vecteurs de caractéristiques pertinents à partir des images en utilisant un SRM [12],
— et l'utilisation d'un classifieur qui, à partir des vecteurs de caractéristiques, apprend un modèle permettant de distinguer les images modifiées (stégos) des images initiales (covers) [11].

3 Réseau de neurones convolutifs (CNN)

Un réseau neuronal est un modèle mathématique dont la conception s'inspire du fonctionnement des neurones biologiques. Il est composé de trois parties appelées couches qui sont elles-mêmes exclusivement composées de neurones. Un neurone effectue un produit scalaire entre les valeurs en entrée et ses paramètres (poids) et applique ensuite une fonction d'activation au résultat. Un réseau est composé de :

— la première couche qui est la couche dans laquelle les données à analyser sont injectées,
— la partie intermédiaire qui est un ensemble de couches appelées couches cachées,
— la dernière couche qui est la couche de sortie. Dans le cas d'une classification, elle affecte un score d'appartenance pour chacun des classes du problème.

Dans un réseau neuronal convolutif, une couche se décompose en trois étapes consécutives :

— l'application de convolution(s),
— l'application d'une fonction d'activation,
— la mise en commun des données.

Les données en entrée sont des images appelées carte de caractéristiques. Une couche de convolution produit également une carte de caractéristique

4 Yedrouj-Net

La figure.1 illustre la structure de notre CNN. Le réseau est composé d'un bloc de **prétraitement**, un bloc de **convolution** comprenant cinq couches, et finalement un bloc **entièrement connecté** comprenant trois couches et une couche de perte **« softmax »**. Le réseau prend en entrée des images de taille 256×256 pixels et fournit une distribution de probabilité pour les deux classes de sortie (image stego ou non).

Le bloc de **prétraitement** a pour objectif de ne conserver que les résidus de bruits hautes fréquences présents dans l'image d'entrée. L'image ainsi traitée est ensuite passée au CNN. Plusieurs articles [2, 3] ont remarqué que sans cette couche préliminaire le réseau convergeait beaucoup plus lentement. Le but de cette couche est de supprimer le contenu de l'image, réduire la dynamique de l'image et ainsi augmenter le rapport signal à bruit entre le signal stégo de faible amplitude (s'il est présent) et le signal lié à l'image. De cette façon, le CNN apprend sur un signal plus compact et plus robuste.

En s'inspirant des avantages de la *diversité* présentés dans [17], nous n'utilisons pas un seul filtre dans l'étape de prétraitement (comme présenté dans [2, 3, 5]) mais la banque de 30 filtres passe-haut présentés dans [12] dans le Spatial Rich Model (SRM). De ce fait, nous obtenons en sortie 30 cartes de caractéristiques. Il faut noter que les poids constituant les noyaux des filtres de l'étape de prétraitement ne sont pas optimisés (appris) durant la phase d'entraînement du réseau. Ces filtres ont été définis de façon à ce qu'ils n'augmentent pas trop la complexité structurelle du réseau. Ainsi, chaque noyau a été étendu sur un support de 5×5 avec en partie centrale les valeurs des poids données dans [12] et le reste a été complété avec des « 0 ». Aucune normalisation des noyaux n'a été effectuée. Le reste de notre CNN peut être divisé deux parties : la partie convolutionnelle visant à transformer l'image d'entrée en un *vecteur caractéristique* et la partie classification qui détermine si l'image est stéganographiée ou non. Comme dans le réseau de Xu *et al.*[5] (Xu-Net), la partie convolutionnelle est composée de cinq blocs repérés par « Bloc 1-5 » dans la figure Fig. 1. Ces blocs extraient des caractéristiques pertinentes des images permettant de déterminer ultérieurement si l'image est stégo ou non. Chaque bloc est composé de toutes ou une partie des fonctionnalités suivantes :

1- **Une couche de convolution**. Comme dans [5] la taille des noyaux des filtres de convolution est fixée à 5×5 pour les blocs 1 et 2 et est ensuite réduite à 1× 1 pour les blocs 3 à 5. Comme dans les réseaux Res-Net [9] et Xu-Net [5], aucun biais n'est utilisé (le terme biais est positionné à faux dans la configuration du réseau). Ce terme est géré dans la couche de mise à l'échelle (voir plus loin)
2- **Une couche d'activation ABS** (uniquement pour le premier bloc comme dans [5]) qui impose au modèle statistique de prendre en compte les symé-

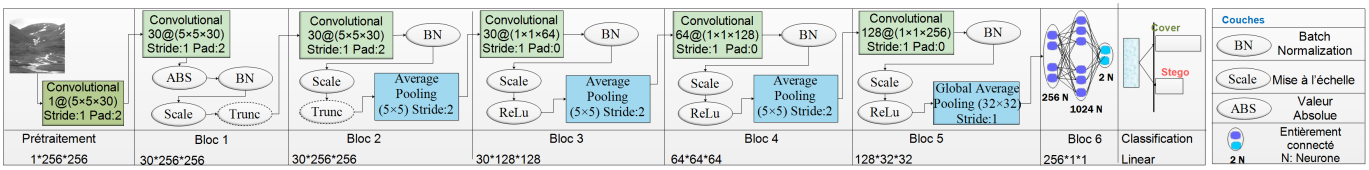


Figure 1 – *Yedrouj-Net CNN.*

tries de signe des résidus du bruit. L'utilité de cette couche a été montrée sur le réseau Xu-Net [5].

3- **Une étape de normalisation en lots (batch normalization)** qui vise à normaliser la distribution de chaque caractéristique afin de les rendre toutes comparables (voir Eq. 1). Cela transforme la distribution des caractéristiques en une distribution centrée (moyenne nulle) et de variance unitaire. Soit une variable aléatoire X dont la réalisation est une caractéristique $x \in \mathbb{R}$, la normalisation en lot de X est donné par l'équation Eq.1 :

$$BN(X, \gamma, \beta) = \beta + \gamma \frac{X - E[X]}{\sqrt{Var[X] + \epsilon}}, \quad (1)$$

avec $E[X]$ l'espérance mathématique de X, Var[X] sa variance et γ et β deux facteurs respectivement pour le décalage et la mise à l'échelle. Dans l'article original [20] les auteurs préconisaient de calculer $E[X]$ et Var[X] et d'apprendre γ et β avec le reste des paramètres du réseau. La normalisation rend l'apprentissage des paramètres moins sensible à l'initialisation [20], et permet d'utiliser un taux d'apprentissage plus grand, ce qui accélère la phase d'apprentissage et améliore la précision de détection [7]. Nos expériences ont montré que séparer la normalisation de la mise à l'échelle augmentait légèrement la précision du réseau. On imposera donc $\beta = 0$ et $\gamma = 1$ pour cette couche de batch normalisation.

4- **Une couche de mise à l'échelle**. Comme dans ResNet [9] et à la différence du Xu-Net [5] qui utilise la couche de normalisation pour apprendre les valeurs de γ et β , nous déportons cette capacité à la couche de mise à l'échelle. On aura donc la fonction de mise à l'échelle suivante :

$$Scale(X, \gamma, \beta) = \beta + \gamma X, \quad (2)$$

5- **Une couche d'activation non linéaire**. Pour les blocs 1 et 2, une fonction *troncature* est utilisée pour limiter la dynamique des valeurs et empêcher les couches plus profondes d'utiliser de grandes valeurs éparses et statistiquement non significatives. La fonction troncature $Trunc$ est donnée par l'équation 3 :

$$Trunc(x) = \begin{cases} -T, si & x < -T \\ x, si & -T \leq x \leq T \\ T, si & x > T \end{cases} \quad (3)$$

où $T \in \mathbb{N}$ est un seuil. Cette suppression des valeurs aberrantes, proposée dans [17], rend le processus plus robuste. Pour les blocs 3 à 5 la fonction d'activation ReLU (*Rectified Linear Unit*) a été utilisée car elle donne de bonnes performances et que le calcul de son gradient est rapide.

6- **Une couche de pooling moyenné**. Cette couche est seulement utilisée dans les couches 2 à 5. Elle permet de faire un sous-échantillonnage des cartes de caractéristiques et ainsi de réduire leur taille. On perd ainsi des informations spatiales mais on limite le risque de sur-apprentissage [21].

Les caractéristiques extraites du module convolutif sont transmises au module de classification qui se compose de trois couches entièrement connectées. Le nombre de neurones dans la première et la deuxième couche est de 256 et 1024 respectivement, et la dernière couche entièrement connectée n'a que deux neurones correspondant au nombre de classes de sortie du réseau. À la fin de ce module, une fonction d'activation softmax est utilisée pour retourner un score pour les deux classes cover et stego.

5 Expérimentations

Nous présentons dans cette section le contexte expérimental ainsi qu'une comparaison des performances de notre réseau avec 3 autres approches.

5.1 Base de données d'images et plateformes logicielles

Pour l'insertion de données dans le domaine spatial nous avons utilisé deux méthodes très connues s'adaptant au contenu de l'image : S- UNIWARD [19] et WOW [18].

Pour la partie stéganalyse, nous avons comparé notre réseau *Y*edrouj-Net avec deux approches de l'état de l'art par CNN : les réseaux Xu-Net [5] et Ye-Net [17], et avec une méthode basée SRM [12] couplé avec un ensemble de classifieurs [11].

Pour que la comparaison soit significative, toutes les méthodes de stéganalyse ont été testées sur des images sous échantillonnées de la base de données d'images BOSS-Base v.1.01 [22] (voir la section 5.2). Les algorithmes d'insertion utilisent les codes en Matlab disponibles en ligne [1] avec un simulateur pour l'insertion et une clef d'insertion aléatoire pour chaque insertion. Nous évitons ainsi toute erreur d'utilisation des codes en C++, i.e. l'utilisation d'une clef unique d'insertion comme mentionné dans [3].

1. http ://dde.binghamton.edu/download/

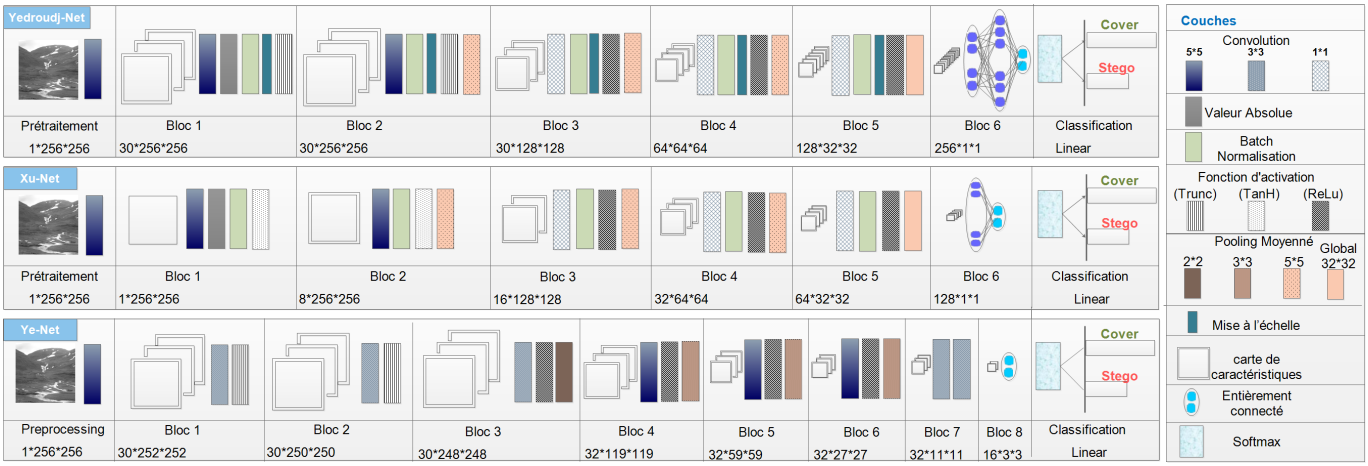


Figure 2 – *Comparison entre Yedroudj-Net, Xu-Net, et Ye-Net.*

Toutes les expérimentations avec les CNNs ont été réalisées en utilisant la toolbox **Caffe** [23] avec les modifications nécessaires et digits V5. La machine ayant servi pour ces tests est équipée d'une carte graphique NVidia Titan X.

5.2 Apprentissage, Validation et tests

Pour prendre en compte les limitations de notre carte graphique ainsi que limiter le temps de calcul, nous avons réalisé les expérimentations sur des images de taille 256×256 pixels (à l'instar de [17]). Pour ceci, toutes les images 512x512 ont été redimensionnées en utilisant la fonction *imresize()* de Matlab avec les paramètres par défaut (interpolation bicubique).

Notre base d'images issues de la BOSSBase est divisée en deux sous-ensembles : la moitié des paires cover/stego pour l'apprentissage et l'autre moitié pour le test. Seules 4000 paires d'images sur les 5000 disponibles de l'ensemble d'apprentissage sont sélectionnées aléatoirement pour l'apprentissage des poids du réseau ; les autres 1000 restantes sont utilisées pour la validation. Les images de l'ensemble de test ne sont jamais utilisées lors de l'entraînement du réseau.

Pour les CNNs, nous avons imposé le nombre maximum d'époque à 500 durant l'apprentissage. Néanmoins, la plupart du temps nous stoppions manuellement la convergence dès lors qu'un phénomène de sur-apprentissage apparaissait. Ceci se manifestait généralement par une valeur de fonction de perte (*Loss*) qui décroit sur l'ensemble d'apprentissage et au contraire qui augmente sur l'ensemble de validation. L'observation de l'évolution de la courbe de la fonction de perte calculée sur le sous-ensemble d'images pour la validation permet de conserver deux modèles du réseaux : celui dont la valeur de la fonction de perte est minimum (resp. maximum) sur les 5 dernières époques. Ces deux modèles de réseaux sont évalués sur l'ensemble d'images de test et on retourne la moyenne de la probabilité d'erreur de détection de ces deux modèles. Pour le SRM plus l'ensemble de classifieurs, nous avons utilisé l'ensemble de caractéristiques SRM de dimension 34671 [12] et l'ensemble de classifieurs de [11]. Nous donnons les résultats de la moyenne sur 10 tests de la probabilité d'erreur minimum, avec a priori égales.

5.3 Hyper-parameters

Afin d'entraîner notre CNN, nous appliquons une descente de gradient stochastique (SGD) sur les mini-batchs. Le momentum est fixé à 0.95 et le weight decay à 0.0001. il n'y a pas de dropout. La taille du mini-batch dans le processus d'apprentissage est initialisé à 16 soit 8 paires cover/stego. Toutes les couches sont initialisées en utilisant l'initialisation de "Xavier" qui initialise les poids de telle sorte que la variance de la distribution gaussienne de l'entrée et de la sortie soit identique [24].

Durant l'apprentissage, on utilise la stratégie itérative de caffe "step policy" pour ajuster le taux d'apprentissage qui est initialisé à 0.01. Une telle stratégie induit une réduction du taux d'apprentissage d'un facteur gamma (égal à 0.1) tout les 10% du nombre total d'époques. Les seuils T des fonctions de troncature valent respectivement 3 et 2 pour la première et la deuxième couche (voir Eq 3).

5.4 Résultats

Le Tableau 1, présente les probabilités d'erreurs obtenues en opérant une stéganalyse avec deux algorithmes d'insertion WOW et S-UNIWARD pour des payloads respectivement de 0.2 et 0.4 bits par pixels (bpp). Les quatre méthodes de stéganalyse comparées sont : le Yedroudj-Net, le Xu-Net [5], le Ye-Net [17] et le SRM+EC [11, 12]. Pour l'algorithme WOW, le Yedroudj-Net a des probabilités d'erreur plus basses de 8% et de 11% respectivement pour des payloads de 0.2 bpp et 0.4 bpp, comparé au SRM+EC. Les résultats restent bons pour la stéganalyse de S-UNIWARD avec une probabilité d'erreur égale pour un payload de 0.2 bpp et 2% plus faible pour un payload de 0.4 bpp.

Notre CNN présente de meilleurs résultats que les autres algorithmes de CNN. Le Yedroudj-Net est meilleur de 2%

Tableau 1 – *Comparaison de Yedroudj-Net et trois méthodes de stéganalyse état de l'art. Nous rapportons la probabilité d'erreur obtenue en utilisant les deux algorithmes d'insertion WOW [18] et S-UNIWARD [19], à 0,2 bpp et 0,4 bpp. Les méthodes de stéganalyse sont Yedroudj-Net, Xu-Net [5], Ye-Net [17] et SRM+EC [11, 12].*

	BOSS 256×256			
	WOW		**S-UNIWARD**	
	0.2 bpp	0.4 bpp	0.2 bpp	0.4 bpp
SRM+EC	36.5 %	25.5 %	**36.6** %	24.7 %
Yedroudj-Net	**27.8** %	**14.1** %	**36.7** %	**22.8** %
Xu-Net	32.4 %	20.7 %	39.1 %	27.2 %
Ye-Net	33.1 %	23.2 %	40.0 %	31.2 %

à 6% que le Xu-Net pour les deux algorithmes et les deux payloads. Ses performances sont de loin meilleures que celle du Ye-Net de 3% à 9%. Notons cependant que les deux autres CNNs ne sont pas plus performants que le SRM+EC. Afin d'être meilleurs que le SRM+EC, il est nécessaire qu'ils utilisent un ensemble de CNNs comme dans [4] ou qu'ils augmentent la base d'apprentissage [6]. L'initialisation du taux d'apprentissage du Ye-Net ainsi que la gestion de son évolution à travers les époques nécessitent beaucoup de minutie. En effet, une mauvaise initialisation empêche le réseau de converger. Dans le Yedroudj-Net et le Xu-Net, la normalisation en lot assure moins de sensibilité. En guise de conclusion à ces comparaisons générales, nous retenons que dans un scénario classique clairvoyant sans connaissance du canal de sélection et sans utiliser d'ensemble, de transfert d'apprentissage, de base de données plus grande ou virtuellement augmentée, le Yedroudj-Net a un avantage certain sur les autres méthodes de l'état de l'art.

6 Conclusion

Dans cet article, nous présentons et évaluons un nouveau CNN conçu pour la stéganalyse dans le domaine spatial, le Yedroudj-Net. Ce CNN est simple, et donne de meilleures performances que l'état de l'art dans un scénario classique clairvoyant sans connaissance du canal de sélection.
Les parties importantes de son architecture sont une banque de filtres pour l'étape de pré-traitement, la fonction d'activation de troncature et la normalisation en lots associée à la couche de mise à l'échelle.

Références

[1] S. Tan et B. Li. Stacked convolutional auto-encoders for steganalysis of digital images. Dans *Proceedings of Signal and Information Processing Association Annual Summit and Conference, APSIPA'2014*, pages 1–4, Siem Reap, Cambodia, Décembre 2014.

[2] Yinlong Qian, Jing Dong, Wei Wang, et Tieniu Tan. Deep Learning for Steganalysis via Convolutional Neural Networks. Dans *Proceedings of Media Watermarking, Security, and Forensics 2015, MWSF'2015, Part of IS&T/SPIE Annual Symposium on Electronic Imaging, SPIE'2015*, volume 9409, pages 94090J–94090J–10, San Francisco, California, USA, Février 2015.

[3] L. Pibre, J. Pasquet, D. Ienco, et M. Chaumont. Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover source-mismatch. Dans *Proceedings of Media Watermarking, Security, and Forensics, MWSF'2016, Part of I&ST International Symposium on Electronic Imaging, EI'2016*, pages 1–11, San Francisco, California, USA, Février 2016.

[4] Guanshuo Xu, Han-Zhou Wu, et Yun Q. Shi. Ensemble of CNNs for Steganalysis : An Empirical Study. Dans *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec'16, pages 103–107, Vigo, Galicia, Spain, Juin 2016.

[5] G. Xu, H.-Z. Wu, et Y. Q. Shi. Structural Design of Convolutional Neural Networks for Steganalysis. *IEEE Signal Processing Letters*, 23(5) :708–712, Mai 2016.

[6] Jishen Zeng, Shunquan Tan, Bin Li, et Jiwu Huang. Pre-training via fitting deep neural network to rich-model features extraction procedure and its effect on deep learning for steganalysis. Dans *Proceedings of Media Watermarking, Security, and Forensics 2017, MWSF'2017, Part of IS&T Symposium on Electronic Imaging, EI'2017*, page 6, Burlingame, California, USA, Janvier 2017.

[7] Mo Chen, Vahid Sedighi, Mehdi Boroumand, et Jessica Fridrich. JPEG-Phase-Aware Convolutional Neural Network for Steganalysis of JPEG Images. Dans *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec'17, page 10, Drexel University in Philadelphia, PA, Juin 2017.

[8] Guanshuo Xu. Deep Convolutional Neural Network to Detect J-UNIWARD. Dans *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec'17, page 6, Drexel University in Philadelphia, PA, Juin 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et Jian Sun. Deep residual learning for image recognition. Dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'2016*, pages 770–778, Las Vegas, Nevada, Juin 2016.

[10] Yann LeCun, Yoshua Bengio, et Geoffrey Hinton. Deep learning. *Nature*, 521(7553) :436–444, Mai 2015.

[11] J. Kodovský, J. Fridrich, et V. Holub. Ensemble Classifiers for Steganalysis of Digital Media. *IEEE Transactions on Information Forensics and Security, TIFS*, 7(2) :432–444, 2012.

[12] J. Fridrich et J. Kodovský. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security, TIFS*, 7(3) :868–882, June 2012.

[13] Chao Xia, Qingxiao Guan, Xianfeng Zhao, Zhoujun Xu, et Yi Ma. Improving GFR Steganalysis Features by Using Gabor Symmetry and Weighted Histograms. Dans *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec'17, page 11, Drexel University in Philadelphia, PA, Juin 2017.

[14] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, et J. Fridrich. Selection-channel-aware rich model for steganalysis of digital images. Dans *Proceedings of IEEE International Workshop on Information Forensics and Security, WIFS'2014*, pages 48–53, Décembre 2014.

[15] T. Denemark, M. Boroumand, et J. Fridrich. Steganalysis features for content-adaptive jpeg steganography. *IEEE Transactions on Information Forensics and Security*, 11(8) :1736–1746, Août 2016.

[16] Y. Qian, J. Dong, W. Wang, et T. Tan. Learning and transferring representations for image steganalysis using convolutional neural network. Dans *Proceedings of IEEE International Conference on Image Processing, ICIP'2016*, pages 2752–2756, Phoenix, Arizona, Septembre 2016.

[17] Jian Ye, Jiangqun Ni, et Yang Yi. Deep Learning Hierarchical Representations for Image Steganalysis. *Accepted to IEEE Transactions on Information Forensics and Security, TIFS*, page 13, 2017.

[18] V. Holub et J. Fridrich. Designing Steganographic Distortion Using Directional Filters. Dans *Proceedings of the IEEE International Workshop on Information Forensics and Security, WIFS'2012*, pages 234–239, Tenerife, Spain, Décembre 2012.

[19] V. Holub, J. Fridrich, et T. Denemark. Universal Distortion Function for Steganography in an Arbitrary Domain. *EURASIP Journal on Information Security, JIS*, 2014(1), 2014.

[20] Sergey Ioffe et Christian Szegedy. Batch normalization : Accelerating deep network training by reducing internal covariate shift. Dans *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456, 2015.

[21] Min Lin, Qiang Chen, et Shuicheng Yan. Network in network. *arXiv preprint arXiv :1312.4400*, 2013.

[22] P. Bas, T. Filler, et T. Pevný. 'Break Our Steganographic System' : The Ins and Outs of Organizing BOSS. Dans *Proceedings of the 13th International Conference on Information Hiding, IH'2011*, volume 6958 de *Lecture Notes in Computer Science*, pages 59–70, Prague, Czech Republic, Mai 2011. Springer.

[23] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, et Trevor Darrell. Caffe : Convolutional architecture for fast feature embedding. Dans *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[24] Xavier Glorot et Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. Dans *Aistats*, volume 9, pages 249–256, 2010.

Quels a priori sont importants pour attaquer un système biométrique embarqué ?

B. Vibert J.M. Le Bars C. Charrier C. Rosenberger

Normandie Univ, ENSICAEN, UNICAEN, CNRS, GREYC, 14000 Caen, France

{benoit.vibert, christophe.rosenberger}@ensicaen.fr
{jean-marie.lebars, christophe.charrier}@unicaen.fr

Résumé

Les attaques d'un système biométrique est un domaine d'étude très important. Généralement les attaques sont faites à différentes parties du système biométrique complet. La majorité des attaques sont effectuées à partir d'image, dans notre étude nous nous attardons plus particulièrement sur les a priori utiles pour un attaquant afin d'usurper l'identité d'un individu. Nous montrons l'utilité de ces a priori pour attaquer un système biométrique embarqué. Le type de l'empreinte digitale ainsi que la résolution de l'image aide un attaquant dans sa tâche.

Mots clefs

Attaques, Template de minuties, Empreinte digitale.

1 Introduction

Nous nous intéressons dans cette étude à la sécurité de systèmes biométriques embarqués sur SE. Nous proposons d'étudier les informations a priori que pourrait exploiter un attaquant afin de faciliter la génération d'un template biométrique pour usurper l'identité d'un individu. Nous souhaitons déterminer si l'utilisation d'a priori, comme le fait de connaître la classe de l'empreinte digitale, le type de capteur, la résolution de l'image ou le nombre de minuties présent dans la référence biométrique de l'individu à usurper peut aider un attaquant à réussir son attaque.

Comme il n'est pas possible de révoquer des données biométriques en cas d'attaque, ces informations sont très sensibles et doivent être protégées le mieux possible. C'est pourquoi le modèle d'empreinte digitale est souvent sauvegardé dans un élément sécurisé. En raison d'une limitation de la taille mémoire ainsi que des capacités de calcul, ce modèle biométrique n'inclut que les minuties stockées sur le SE suivant la représentation ISO Compact Card II [1]. Cette représentation est utilisée pour la mise en correspondance entre la référence et les échantillons capturés. La sécurité des systèmes biométriques embarqués sur SE est, dès lors, une exigence primordiale. L'attaque classique de ce type de système consiste à envoyer au système biométrique un template biométrique afin d'usurper l'identité d'un individu. L'attaquant doit alors générer un template biométrique pour réaliser cette attaque. Une attaque commune sur les algorithmes de comparaisons biométriques embarqués sur SE, consiste à envoyer des modèles aléatoires de minuties pour tenter de se faire passer pour un individu. Ces attaques sont nommées *force brute*, différents travaux ont été réalisés sur ce type d'attaque [2, 3]. Un autre type d'attaque simple existe, et consiste à utiliser un template biométrique calculé à partir de sa propre donnée biométrique, cette attaque est nommée « zéro effort ». Cette attaque n'a que très peu de chance de fonctionner, mais elle peut servir de base pour des attaques plus évoluées.

Les empreintes digitales sont généralement réparties suivant la classification proposée par Henry pour laquelle cinq classes ont été identifiées : Arche, Boucle à gauche, Boucle à droite, Tente et Spirale [4, 5]. En ce qui concerne la sécurité, une comparaison biométrique embarquée (OCC) présente de nombreuses vulnérabilités. Comme présenté par Ratha et *al.* [6] et plus récemment Jain *et al.* [7] ont classé les attaques d'un système biométrique générique en huit catégories (résumées dans la figure 1). Pour chacun des points identifiés, il existe différents types d'attaques. Uludag et Jain [2], Martinez [3] et Soutar [8] considèrent les points 2 et 4 pour effectuer une attaque dite de *hill-climbing*. Cette attaque peut être réalisée par une application qui envoie des données aléatoires (perturbées itérativement) au système. L'application récupère le score de correspondance, entre la référence biométrique et l'échantillon testé, et poursuit ses perturbations seulement lorsque le score de correspondance augmente et jusqu'au moment où on atteint le seuil d'acceptation. À notre connaissance, aucune étude sur les a priori exploitables par un attaquant sur l'empreinte digitale de la personne à usurper n'a été menée. Dans cette étude, nous considérons les attaques sur les points 1 et 2. Pour effectuer une telle attaque, nous devons remplacer le module *capteur* par notre propre mécanisme. Les informations disponibles sur un capteur biométrique sont définies ci-après, c'est pourquoi nous proposons d'étudier l'impact de ces informations en tant qu'a priori :

— Classe de l'empreinte digitale ;

— Type de capteur (utilisé à l'enrôlement) ;
— Résolution de l'image ;
— Nombre de minuties extraites.

Notre hypothèse est qu'un attaquant ait un accès logique au système et envoie de faux templates biométriques à l'élément sécurisé en exploitant ces a priori. Pour évaluer l'impact de ces a priori sur l'efficacité d'une attaque, nous utilisons la plateforme EVABIO pour caractériser son influence sur la décision de comparaison.

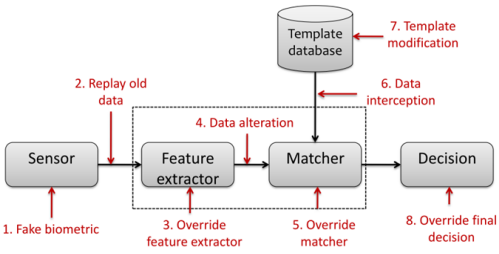


Figure 1 – *Localisation des vulnérabilités sur un système biométrique (définies par [6])*

Ce papier est organisé de la façon suivante : tout d'abord, nous présentons les a priori utilisés pour usurper l'identité d'un individu. Ensuite, nous déterminons quels a priori sont les plus pertinents pour réussir à attaquer un système biométrique embarqué. Nous finissons par une conclusion, ainsi que des perspectives à cette étude.

1.1 Plateforme EVABIO - module Attaque

Dans le cadre de la plateforme EVABIO, nous avons développé un nouveau module d'attaque pour mener notre étude, décrit dans la figure 2. Il offre aux développeurs ainsi qu'aux chercheurs différentes méthodes d'attaque lors de la comparaison d'empreintes digitales. En outre, la plateforme permet de tester les attaques sur des algorithmes embarqués (OCC) ainsi que sur des ordinateurs. Grâce à la modularité de la plateforme EVABIO, nous avons l'avantage de pouvoir seulement modifier le module d'Attaque, pour quantifier l'avantage qu'a un attaquant à connaître la classe d'empreintes digitales afin d'usurper l'identité d'un individu. Dans cette étude, le module `Attaque` est mis à jour car il contient des méthodes permettant de tester les connaissances utiles pour un attaquant comme le type de capteur d'empreinte, la résolution de l'image extraite par le capteur, la classe de l'empreinte digitale ou bien le nombre de minuties extraites de l'image. Compte tenu de toutes ces informations, nous déterminons si ce type de connaissances est important ou non pour qu'un attaquant réussisse à usurper l'identité d'individus. Ce module contient également une méthode pour générer un template biométrique aléatoire respectant la norme ISO à l'aide du logiciel SFinge [9], qui peut être utile pour l'attaque par *force brute*. Avec cette méthode, il est possible de générer des modèles d'empreintes digitales aléatoires pour attaquer l'algorithme de comparaison embarqué.

Figure 2 – *Schéma général de la plateforme EvaBio (défini dans [10]) avec le module Attacks developpé*

1.2 Les a priori sur une empreinte digitale

Nous supposons qu'un attaquant ne peut remplacer le module de capteur que par son propre "*Faux Capteur*", lui permettant ainsi d'avoir accès aux 4 informations le constituant et réparties suivant les vulnérabilités définies dans le modèle de Ratha : 1) en modifiant la résolution de l'image fournie en sortie du capteur influençant ainsi la partie extraction des minuties, 2) en fixant la classe de l'empreinte digitale, 3) en fournissant des informations sur le type de capteur utilisé pendant le processus d'enrôlement ; 4) lorsque l'attaque est effectuée juste après le processus d'extraction des minuties, l'attaquant peut connaître le nombre de minuties extraites et sauvegardées comme référence dans le module de comparaison embarqué (OCC). Toutes les minuties extraites sont stockées dans un template au format ISO Compact Card II.

Nous voulons quantifier dans quelle mesure la connaissance pour un attaquant des paramètres utilisés par le capteur augmente la probabilité de réussir une attaque. Cette probabilité est basée sur le taux de fausses acceptations (FAR) qui peut être considérée comme la probabilité d'une attaque réussie. Avec b_z le modèle de référence de l'utilisateur z et D un algorithme de comparaison basé sur une distance entre une référence et un échantillon biométrique. Le succès d'une attaque par un imposteur est donné par :

$$FAR_A(\epsilon) = P[D(b_z, A_z) \leq \epsilon)] \quad (1)$$

où FAR_A est la probabilité d'une attaque réussie pour un seuil de décision fixé à ϵ. La requête biométrique A_z est construite par l'imposteur en prenant en compte toutes les informations qu'il connaît sur l'utilisateur z ou sur le système biométrique. Notre but est alors d'estimer l'avantage pour un attaquant de construire A_z lorsqu'il connait la classe de l'empreinte digitale C_z, le type de capteur S_z, le nombre de minuties MN_z ou la résolution de l'image R_z de l'utilisateur z.

1.3 Les paramètres de l'expérimentation

Pour l'expérimentation, nous avons créé des bases de données biométriques spécifiques et nous avons utilisé Bozorth et MCC qui sont deux algorithmes de comparaison déjà étudiés dans la section 1.3.

Bases de données biométriques. Nous avons utilisé le logiciel SFinge [9] pour générer les différentes bases de données biométriques synthétiques, n'ayant pas à notre disposition de bases de données avec l'information sur les différents types de capteurs, ainsi que les différentes résolutions d'images, le nombre de minuties et surtout la classe de l'empreinte digitale. De plus, il a été démontré dans différents travaux [11, 12] que SFinge produit des empreintes digitales synthétiques avec des comportements similaires en terme de taux de reconnaissance à ceux obtenus à partir de bases de données réelles, c'est pourquoi nous l'utilisons ici.

Pour chacun des quatre a priori, deux types de bases de données sont conçus :

— **Base de données de référence** : Cette base de données simule les modèles de référence des utilisateurs. Nous avons généré un échantillon par utilisateur pour 500 individus. Cette base de données contient donc 500 empreintes digitales ;
— **Base de données d'attaque** : Nous avons généré une base de données avec 1000 échantillons d'empreintes digitales différents (un échantillon par utilisateur). Cette base de données est utilisée pour les attaques.

En utilisant SFinge, nous pouvons choisir le type de capteur parmi deux types, Capacitif et Optique. Cela induit la construction de quatre bases de données (une BDD de référence et une BDD d'attaque par type de capteur). En considérant le niveau de résolution de l'image, nous avons 3 valeurs (250dpi, 500dpi, 1000dpi) induisant 6 bases de données. En ce qui concerne le nombre de minuties, nous avons créé deux classes (nombre de minuties < 38 ou > 38) induisant 4 bases de données. Enfin, lorsque l'on considère la classe des empreintes digitales (Arche, Boucle à gauche, Boucle à droite, Tente et Spirale), 10 bases de données sont générées (2 par classe).

Afin d'estimer le seuil de décision ϵ utilisé dans l'équation 1 permettant de calculer la valeur EER, nous avons généré une base de données dédiée en utilisant SFinge avec les paramètres par défaut, que nous nommons BDD_SFinge. Les seuls paramètres que nous avons fixés sont le nombre d'utilisateurs (100) ainsi que le nombre de modèles par utilisateur (8). Enfin, nous obtenons un total de 800 empreintes digitales. C'est un choix arbitraire, ce point de fonctionnement est toujours accessible pour n'importe quel algorithme de comparaison.

Les algorithmes de comparaison. Dans cette étude, nous avons utilisé deux algorithmes de comparaison issus de la recherche :

— **Bozorth3** : La valeur EER de cet algorithme a été calculée en utilisant la base de données BDD_SFinge. La valeur obtenue est égale à 1,03% pour une valeur de seuil de décision $\epsilon = 26,8$;

— **Minutia Cylinder-Code (MCC) algorithm** : La valeur de l'EER de cet algorithme a aussi été calculée en utilisant la base de données BDD_SFinge. La valeur obtenue est égale à 0% pour un seuil de décision $\epsilon = 0,0315$.

Protocole expérimental. Pour toute attaque, un imposteur fournit une requête pour être authentifié en tant qu'utilisateur légitime. Deux scenarii sont mis en œuvre pour simuler une attaque :

1. **Scénario 1** : Nous simulons une attaque de type *force brute*. 500 modèles sont sélectionnés aléatoirement, en suivant une distribution uniforme, dans la base de données construite par notre outil de génération aléatoire de template biométrique, ce qui constitue la base de données de référence. La base de données d'attaques est générée en construisant 1000 templates biométriques aléatoires mais respectant le format ISO, lui même provenant de notre outil de génération de templates biométriques.

2. **Scénario 2** : Pour chacun des a priori donné, une base de données de référence est générée avec le logiciel SFinge contenant 500 modèles. De plus, pour chacun des a priori, une base de données d'attaque contenant 1000 templates biométriques est générée et sera comparée aux bases de données de référence. Par exemple, en considérant le type de capteur, nous obtenons quatre comparaisons comme représenté dans le tableau 1.

BDD de reference	BDD d'attaque
Capacitif	Capacitif
Capacitif	Optique
Optique	Capacitif
Optique	Optique

Tableau 1 – *Exemple du scénario 2 pour le type de capteur*

1.4 Résultats expérimentaux

Dans cette partie, nous présentons les résultats de l'expérimentation pour chaque a priori pris indépendamment.

Type de capteur. En considérant la connaissance du type de capteur utilisé pour générer la référence biométrique de l'individu, nous calculons la valeur FAR_A pour les deux scénarii décrits précédemment lorsque nous fixons la valeur du seuil de décision par rapport à l'algorithme de comparaison utilisé comme décrit dans la section 1.3.

Le tableau 2 donne la valeur de probabilité d'attaque réussie FAR_A pour chaque type de capteur et les deux algorithmes de comparaison. Nous pouvons clairement voir que la connaissance du type de capteur utilisé lors de l'enrôlement n'apporte pas d'aide à l'attaquant.
L'avantage qu'a un attaquant lorsqu'il connaît le type de capteur utilisé à l'enrôlement pour générer la référence n'est pas très significatif.

Nombre de minuties extraites. Avec la connaissance de cet a priori, (le nombre de minuties dans la référence biométrique de l'individu), nous calculons la valeur FAR_A pour les deux scénarii décrits précédemment lorsque nous fixons la valeur du seuil de décision par rapport à l'algorithme de comparaison comme décrit dans la section 1.3. Les résultats obtenus montrent que pour Bozorth3, la probabilité d'une attaque réussie est égale à 0,0141% avec la méthode de type *force brute* et 0,0162% lorsque l'on connait le nombre de minuties dans la référence biométrique. Pour l'algorithme MCC, la probabilité qu'une attaque soit réussie est égale à $1.63 \times 10^{-4}\%$ avec l'attaque de type *brute force* et $1.6 \times 10^{-4}\%$ en connaissant le nombre de minuties.

On peut voir dans les deux cas que l'attaquant obtient peu de résultats avec seulement cette information. Afin d'analyser si la connaissance du nombre de minuties de la référence biométrique a un impact sur l'efficacité de cette attaque, nous appliquons le scénario suivant : nous ne considérons que les scores entre le modèle de référence et les tests ayant le même nombre de minuties.

Dans ce cas, nous avons deux séries de $4 \times 800 = 3200$ scores de correspondances. Nous pouvons calculer la valeur FAR_A pour les deux classes du nombre de minuties. Si nous considérons l'algorithme de comparaison Bozorth3, les attaques réussissent plus pour $1 < \epsilon < 35$ lorsque le nombre de minuties est supérieur à 38. Pour l'algorithme de correspondance MCC, la même remarque peut être formulée pour $0.0011 < \epsilon < 0.0023$. Le tableau 3 donne la valeur de la probabilité d'attaque réussie FAR_A pour chaque classe du nombre de minuties pour les deux algorithmes de comparaison. Nous pouvons voir clairement que si nous avons plus de 38 minuties, cette information aide plus l'attaquant mais elle ne suffit pas à augmenter de façon importante le succès de l'attaque.

Résolution de l'image. En ce qui concerne la connaissance de la résolution de l'image originale, nous calculons FAR_A pour les deux scénarii lorsque nous fixons le seuil de décision pour obtenir la valeur à l'EER. Les résultats obtenus montrent que lorsque nous utilisons l'algorithme de comparaison Bozorth3, la probabilité d'une attaque réussie est égale à 0,019% avec une attaque de type *force brute* et 0,035% connaissant la résolution de l'image originale. En considérant l'algorithme MCC, la probabilité d'attaque réussie est égale à 0,51 $\times 10^{-3}\%$ avec une attaque de type *brute force* et $0,8 \times 10^{-3}\%$ connaissant la résolution de l'image originale.

Nous pouvons voir, dans les deux cas, le petit avantage pour un attaquant de connaître la résolution de l'image originale extraite par le capteur. Afin d'analyser si la résolution de l'image originale a un impact sur l'efficacité de cette attaque, nous appliquons le schéma suivant : nous ne considérons que les scores entre le modèle de référence et d'attaque ayant la même résolution d'images. Dans ce cas, nous avons 3 séries de $4 \times 800 = 3200$ scores de correspondances. On peut ainsi calculer la valeur FAR_A pour chaque classe de résolution d'image, comme le montre la figure 3.

Pour l'algorithme de comparaison Bozorth3, nous pouvons voir qu'il est tout à fait impossible de réussir l'attaque avec une image de haute résolution (1000dpi), contrairement à une image de faible résolution (250dpi). La même remarque peut être formulée pour l'algorithme MCC. Le tableau 4 donne la valeur de la probabilité d'attaque réussie FAR_A pour chaque résolution d'image pour les deux algorithmes de comparaison. Nous pouvons clairement voir que la basse résolution aide un attaquant avec plus de 3 fois plus d'attaques réussies que la résolution moyenne (500dpi). Il faut donc éviter d'utiliser des images de faible résolution pour limiter ce type d'attaque.

Type d'empreinte digitale. A partir de la connaissance de la classe d'empreintes digitales, nous calculons la valeur FAR_A pour les deux scénarii lorsque nous fixons la valeur du seuil de décision par rapport à l'algorithme de comparaison comme décrit dans la section 1.3. En considérant l'algorithme Bozorth3, la probabilité d'une attaque réussie est égale à 3% avec la méthode de type *brute force* et à 4,7% en connaissant la classe de l'empreinte digitale. Les résultats obtenus montrent que lorsque nous utilisons l'algorithme de comparaison MCC, la probabilité d'une attaque réussie est égale à 1,7% avec la méthode de type *brute force* et 2,6% avec la connaissance de la classe d'empreintes digitales. Nous pouvons en déduire que la connaissance de la classe d'empreintes digitales enrôlée sur l'élément sécurisé aide un attaquant à se faire authentifier sur le système. Cependant, nous devons étudier comment cette connaissance influence l'efficacité de l'attaque.

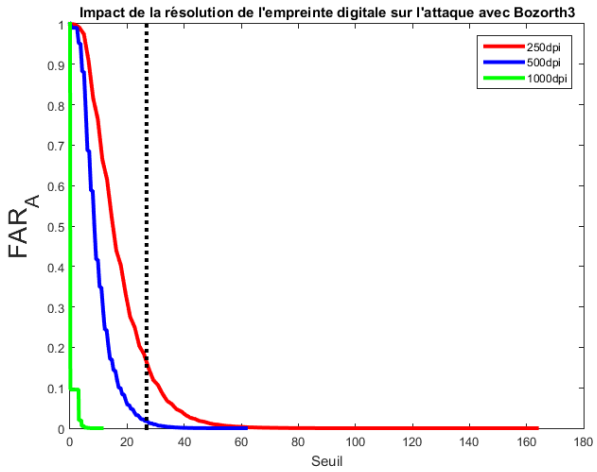
Afin d'analyser son impact, nous appliquons l'approche suivante : nous considérons uniquement les scores entre les modèles de référence et d'attaque ayant la même classe d'empreintes digitales pour calculer la valeur FAR pour

Algorithme de comparaison	Capacitif	Optique
Bozorth3	0.0158 %	0.016 %
MCC	0.13×10^{-3} %	0.23×10^{-3} %

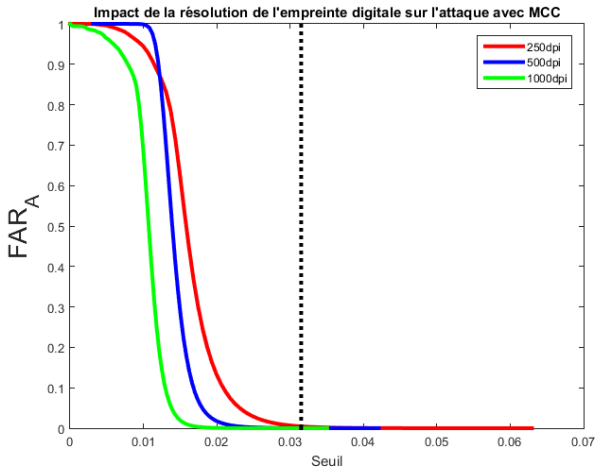
Tableau 2 – *Valeur de la probabilité d'une attaque réussie FAR_A pour chaque type de capteur pour les deux algorithmes de comparaison.*

Algorithme de comparaison	< 38	> 38
Bozorth3	0.0038 %	0.0391 %
Minutia CC	0.8×10^{-4} %	2.5×10^{-4} %

Tableau 3 – *Valeur de la probabilité d'attaque réussie FAR_A pour les deux classes du nombre de minuties pour les deux algorithmes de comparaison.*



(a) Impact de la résolution de l'image avec Bozorth3



(b) Impact de la résolution de l'image avec MCC

Figure 3 – *Évolution de l'efficacité des attaques en considérant les trois résolutions du capteur pour les deux algorithmes de comparaison.*

chacune des classes. Dans ce cas, nous avons 5 séries de $4 \times 800 = 3200$ scores de correspondances nous permettant de calculer la valeur FAR_A.

Les résultats sont présentés dans la figure 4. En considérant l'algorithme de comparaison Bozorth3, la figure 4(a) nous permet de déduire que la classe *Arche* présente le taux d'attaque avec le succès le plus élevé alors que la classe *boucle à droite* présente le taux le plus bas. Par contre, pour l'algorithme de comparaison MCC, nous observons dans la figure 4(b) que la classe *Spirale* présente le taux d'attaque réussie le plus élevé contrairement à la classe *boucle à droite* ayant le plus faible taux. Une première remarque que nous pouvons formuler est que les empreintes digitales appartenant à la classe *boucle à droite* sont les moins simple à usurper. Le tableau 5 donne la valeur de la probabilité d'attaque réussie FAR_A pour chaque classe d'empreintes digitales pour les deux algorithmes de comparaison. Nous pouvons clairement voir que certaines classes d'empreintes digitales sont plus faciles à attaquer en fonction de l'algorithme de comparaison utilisé. Par exemple avec Bozorth3, les empreintes de la classe *Arche* peuvent être usurpées dans 50% des cas, ce qui est très important. En conclusion Bozorth ne doit pas être utilisé car il est sensible aux attaques sur les empreintes digitales de type *Arche*.

Discussion

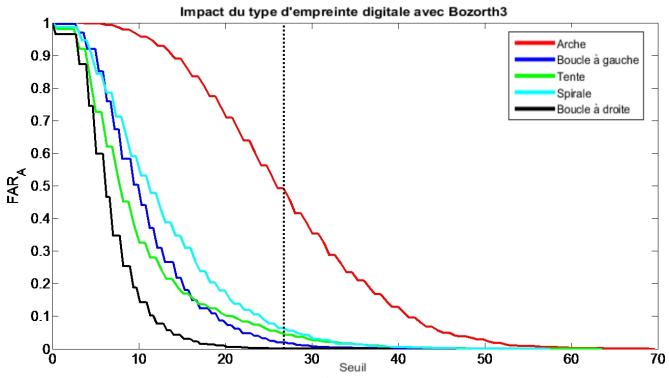
Dans cette étude originale, par rapport à la littérature, nous avons voulu savoir quels a priori étaient importants pour un attaquant lorsqu'il souhaite usurper l'identité de l'individu enrôlé sur l'élément sécurisé. Nous avons montré les connaissances aidant un attaquant à usurper l'identité d'un individu, ainsi que l'apport des informations lui permettant d'augmenter ses probabilités d'attaque réussie, comme la classe d'empreintes digitales ainsi que la résolution de l'image. Nos expériences montrent que si nous connaissons la classe d'empreinte digitale pour les individus enrôlés sur le système, nous augmentons en général la probabilité d'usurper leurs identités. D'autre part, le nombre de minuties ainsi que le type de capteurs (capacitif et optique) sont moins significatifs pour aider

Algorithme de comparaison	250dpi	500dpi	1000dpi
Bozorth3	0.165 %	0.047 %	0 %
Minutia CC	0.45×10^{-3} %	0.176×10^{-3} %	0 %

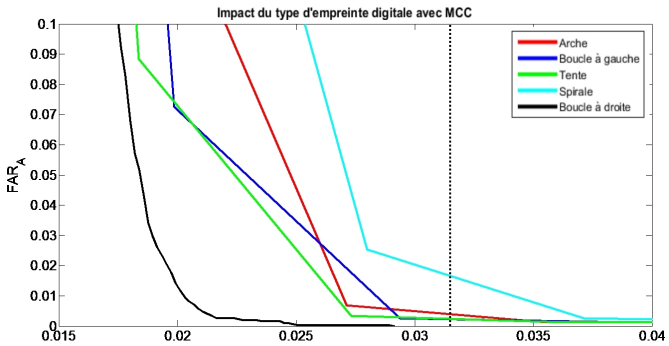
Tableau 4 – *Valeur de la probabilité d'une attaque réussie FAR_A pour chaque résolution des images originales pour les deux algorithmes de comparaison.*

Algorithme de comparaison	Arche	Boucle à droite	Boucle à gauche	Tente	Spirale
Bozorth3	50 %	0 %	2 %	5 %	6.3 %
Minutia CC	0.6 %	0 %	0.2 %	0.2 %	2 %

Tableau 5 – *Valeur de la probabilité FAR_A d'une attaque réussie pour chaque classe d'empreintes digitales pour les deux algorithmes de comparaison.*



(a) Impact du type d'empreinte digitale avec Bozorth3



(b) Impact du type d'empreinte digitale avec MCC

Figure 4 – *Évolution de l'efficacité des attaques en tenant compte de toutes les classes d'empreintes digitales pour les deux systèmes biométriques.*

un attaquant. L'algorithme sur lequel nous obtenons le meilleur taux d'usurpation d'identité est Bozorth avec les empreintes digitales de type Arche. Nous émettons l'hypothèse que cet algorithme, destiné à la recherche et au public, est moins performant et non optimisé pour les empreintes de type Arche. Si l'on regarde les autres types d'empreintes pour les deux algorithmes, nous remarquons que le taux d'usurpation est assez faible ce qui est assez logique et cohérent. De plus, pour les deux algorithmes le plus haut taux d'acceptation est sur les Spirales, qui sont de surcroit le type d'empreinte digitale le plus courant [13]. D'une manière générale, nous en déduisons que le taux de réussite d'une attaque dépend quasiment exclusivement du fonctionnement de l'algorithme de comparaison.

La connaissance du type de l'empreinte digitale de l'individu à usurper est une information importante pour un imposteur.

2 Conclusion et perspectives

Dans ce papier, nous avons évaluer les informations utiles, sur un système biométrique, pour un attaquant lorsque le capteur biométrique est corrompu. Sur ce type d'attaque, nous avons vu que seulement quatre paramètres sont disponibles : le type de capteur, la résolution, le nombre de minutie et le type de l'empreinte. Nous avons démontré que seulement deux a priori permettent d'avoir des informations permettant de se faire accepter par le système, la résolution de l'image et, le plus important, le type de template biométrique enrôlé sur le système.

Références

[1] ISO/IEC 19794-2. information technology - biometric data interchange format format - part 2 : Finger minutiae data, 2011.

[2] Umut Uludag and Anil K Jain. Attacks on biometric systems : a case study in fingerprints. In *Electronic Imaging 2004*, pages 622–633. International Society for Optics and Photonics, 2004.

[3] Marcos Martinez-Diaz, J Fierrez-Aguilar, Fernando Alonso-Fernandez, Javier Ortega-García, and JA Siguenza. Hill-climbing and brute-force attacks on biometric systems :

A case study in match-on-card fingerprint verification. In *Proceedings 2006 40th Annual IEEE International Carnahan Conferences Security Technology*, pages 151–159. IEEE, 2006.

[4] Anil K Jain, Salil Prabhakar, and Lin Hong. A multichannel approach to fingerprint classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(4) :348–359, 1999.

[5] Qinzhi Zhang and Hong Yan. Fingerprint classification based on extraction and analysis of singularities and pseudo ridges. *Pattern Recognition*, 37(11) :2233–2243, 2004.

[6] Nalini K. Ratha, Jonathan H. Connell, and Ruud M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM systems journal*, 40(3) :614–634, 2001.

[7] Anil K Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research : Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 2016.

[8] Colin Soutar et al. Biometric system security. *White Paper, Bioscrypt, http ://www. bioscrypt. com*, 2002.

[9] Raffaele Cappelli, D Maio, and D Maltoni. Sfinge : an approach to synthetic fingerprint generation. In *International Workshop on Biometric Technologies (BT2004)*, pages 147–154, 2004.

[10] B Vibert, Z Yao, Sylvain Vernois, Jm Le Bars, Christophe Charrier, and Christophe Rosenberger. Evabio platform for the evaluation biometric system : Application to the optimization of the enrollment process for fingerprint device. In *International Conference on Information Systems Security and Privacy*, 2015.

[11] Dario Maio, Davide Maltoni, Raffaele Cappelli, Jim L Wayman, and Anil K Jain. Fvc2004 : third fingerprint verification competition. In *Biometric Authentication*, pages 1–7. Springer, 2004.

[12] Julian Fiérrez-Aguilar, Loris Nanni, Javier Ortega-Garcia, Raffaele Cappelli, and Davide Maltoni. Combining multiple matchers for fingerprint verification : a case study in fvc2004. In *International Conference on Image Analysis and Processing*, pages 1035–1042. Springer, 2005.

[13] HandResearch. Fingerprints world map. http ://www.handresearch.com/news/fingerprints-world-map-whorls-loops-arches.htm.

Vers un Code Personnel d'Identité Respectueux de la Vie Privée

D. Migdal[1] C. Rosenberger[1]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

{denis.migdal, christophe.rosenberger}@ensicaen.fr

Résumé

De nombreuses applications sur Internet nécessitent d'avoir des informations sur l'internaute, pour vérifier qu'il a bien le droit à d'accéder à un service numérique (vérification d'une preuve d'identité comme un mot de passe), pour éviter des attaques (pédopornographie, usurpation de profil,...) ou pour donner de la confiance aux autres utilisateurs (réseaux sociaux). Nous proposons dans ce papier une méthode de génération d'une signature basée sur l'identité de l'individu. Elle est calculée à partir de 1) la collecte de données biométriques sur l'individu, sur son ordinateur, son navigateur web, 2) le pré-traitement de ces données et 3) la protection des données personnelles pour la génération d'un code binaire. Nous illustrons l'intérêt de la méthodologie proposée avec des résultats préliminaires sur des données personnelles réelles.

Mots clefs

Informations personnelles, biométrie comportementale, protection de données personnelles.

1 Introduction

La consommation de services numériques sur Internet est de nos jours importante que ce soit pour les réseaux sociaux, le commerce électronique ou les jeux en ligne. À titre d'exemple, en 2016, 96% des français ayant demandé un extrait du casier judiciaire l'ont fait sur Internet [1]. Néanmoins, plusieurs données personnelles peuvent être récupérées lors de l'usage d'un service numérique sur Internet soit fournies par l'internaute (notamment sur les réseaux sociaux) soit collectées automatiquement.

Les services numériques sur Internet collectent de plus en plus des données personnelles liées à l'internaute parfois à des fins légitimes (détection de fraude, examen à distance, ...) mais aussi à des fins non conformes aux conditions de collecte (vente à d'autres services, consolidation d'identités, ...). Ces données personnelles peuvent être liées à l'individu (donnée biométrique, nom, âge, ...), au navigateur (version, type, ...), à la machine de l'internaute (système d'exploitation, matériel, résolution de l'écran). Toutes ces informations parfois collectées dans un contexte légitime peuvent aller jusqu'à identifier l'individu ce qui pose un problème majeur de respect de la vie privée.

La principale contribution de ce papier est de proposer une méthode de génération d'un code binaire lié à l'identité numérique d'un individu. Ce code ne permet pas de remonter aux informations utilisées pour le calculer et permet également de réaliser des comparaisons avec d'autres codes. Nous présentons les différentes étapes de calcul de ce code. Les informations utilisées vont du navigateur utilisé, à la machine jusqu'à l'individu. Des pré-traitements sont réalisés sur ces données afin de calculer le code dans la dernière étape. Nous présentons brièvement quelques applications intéressantes de ce code (authentification ou identification d'attaques comme la détection de plusieurs comptes associés à une identité).

Cet article est organisé comme suit. La section 2 présente un état des travaux antérieurs sur la collecte et l'utilisation de données personnelles. La méthode proposée est décrite dans la section 3. La section 4 présente des résultats préliminaires sur des données réelles. Nous concluons et donnons quelques perspectives dans la section 5.

2 Travaux antérieurs

Le *Browser Fingerprinting* permet de suivre les utilisateurs dans leur navigation Internet grâce aux données discriminantes qu'un service donné peut récolter, souvent dans l'objectif de proposer des "services personnalisés" correspondant au profil-type de l'utilisateur. Les sites Panopticlick [4], IAmUnique [6], et UniqueMachine [3] permettent de calculer son *Browser Fingerprint* à partir des données collectées par le site, généralement via le réseau et l'API JavaScript afin de déterminer le degré d'unicité de l'empreinte calculée parmi celles déjà collectées. Plus le *browser fingerprint* est unique, plus un service aura capacité à le discriminer.

Cependant, le *browser fingerprint* peut varier, e.g. par le changement du navigateur, de sa configuration [7], ou tout simplement de machine. Le but n'est pas d'identifier l'utilisateur de façon certaine, mais d'identifier un ensemble de sessions de navigations appartenant à un même utilisateur. Les données utilisées pour le *browser fingerprinting* peuvent être liées, e.g., au matériel (e.g. carte graphique [3], écran), au système d'exploitation, au navigateur uti-

lisé, à sa configuration, aux polices installées [4, 6], à l'historique du navigateur [10], ou aux domaines bloqués [2].

3 Méthode proposée

L'objectif de la méthode proposée est de calculer un code binaire lié à une personne à partir d'informations personnelles (techniques et biométriques). Ce code doit répondre à différentes exigences :

- *Non inversible* : le code binaire de l'utilisateur ne doit pas donner d'informations sur les données personnelles collectées.
- *Confidentialité* : la valeur des attributs ne peut être connue, ni déduite, par le service.
- *Conservation de la similarité* : Si les données personnelles d'un individu sont similaires alors les codes binaires résultant doivent l'être.
- *Non-usurpation* : un tiers ne peut forger un code lui permettant d'usurper un utilisateur légitime.
- *Répudiation* : l'utilisateur légitime doit pouvoir révoquer un code binaire existant.

Dans le cadre de cet article, un score de confiance peut être calculé avec la distance de Hamming entre la preuve et l'engagement, tous deux vecteurs binaires de taille fixe. Aussi, nous considérerons, et approfondirons les modalités d'informations personnelles suivantes :

- ce que l'utilisateur est/sait faire : sa biométrie comportementale ;
- ce que l'utilisateur possède : son navigateur ;
- où l'utilisateur est : sa localisation physique et organisationnelle ;
- "ce que l'utilisateur préfère" : sa configuration.

La figure 1 présente le principe général de la méthode proposée. Un simple mot de passe est utilisé comme clé secrète [5]. Dans ce cas, Alice par la saisie du mot de passe consent à donner ce code binaire au service. Les différentes étapes de calcul sont présentées par la suite.

3.1 Collecte de données personnelles

A l'heure actuelle, il est possible de collecter un grand nombre de données personnelles. Nous détaillons les informations collectées par grande catégorie.

Navigateur. Afin d'authentifier un navigateur, une simple clé stockée sur ce dernier suffit. La clé, que nous nommerons *localkey*, est une valeur de n bits générée aléatoirement au premier usage du navigateur, utilisée ensuite pour l'authentifier. Pour n suffisamment grand, la probabilité de collision est négligeable, et la recherche exhaustive difficile. Dans le cadre de l'expérience, n=64, pour des besoins en sécurité plus importants, la taille de la clé peut-être augmentée, e.g. n=512.
La clé peut être stockée dans le localStorage [1] du navigateur, ou, idéalement, dans le simple-storage d'une WebExtension. Il est cependant possible à un attaquant de subtiliser la clé s'il a accès à la machine, ou à la session de l'utilisateur. Les clés étant générées aléatoirement, la compromission d'une clé ne compromet pas les clés des autres navigateurs possédés par l'utilisateur. Il est possible de protéger la clé, e.g. en la chiffrant, ainsi que d'en détecter l'utilisation frauduleuse, e.g. via les autres informations personnelles. Cependant, ceci ne sera pas abordé dans le cadre de cet article.

Localisation. Les adresses IP sont distribuées par plages, de l'IANA [2] aux RIR [3], des RIR aux RIL [4], et enfin des RIL aux utilisateurs. Il est ainsi possible d'en déduire le réseau de l'utilisateur, et sa position administrative (e.g. département) ou physique (e.g. position GPS). Cependant, le réseau TOR, un VPN, ou un proxy, peuvent être utilisés pour masquer l'adresse IP de l'utilisateur. Le réseau et positions déduites de l'adresse IP seront alors ceux du proxy, du VPN, ou du nœud TOR sortant.
Dans le cadre de cet article, les localisations administrative (pays, région, département, ville) et physique (latitude et longitude) sont déterminées via l'API Google Map à partir d'une adresse extraite de la base dp-ip [5]. Dans un travail futur, il serait aussi possible de déduire, soit le FAI (Fournisseur d'adresse Internet) de l'utilisateur, soit sa localisation structurelle au sein d'une entité (e.g. entreprise, université, centre de recherche, structures gouvernementales), à l'aide de requêtes DNS, reverse DNS, WHOIS IP, et WHOIS domain. Il est aussi possible d'avoir plus d'informations sur l'adresse IP à l'aide de DNSBL [6].

Données réseau. Les données envoyées au service par les protocoles de communication sont discriminantes et permettent, par des techniques de *browser fingerprinting*, d'identifier l'utilisateur [4, 6]. De manière analogue, ces données peuvent être exploitées pour authentifier l'utilisateur en les comparant avec les données d'enrôlement. Ainsi, cette modalité ne peut être utilisée si les données sont, à chaque échange, générées aléatoirement. Cependant, l'usurpation est triviale pour qui a connaissance de ces données, e.g. pour qui fournit un service à l'utilisateur. De même, l'utilisation de données normalisées, e.g. via l'utilisation du navigateur TOR, augmente la probabilité de collision. Cette modalité accorde ainsi peu de confiance en l'authentification de l'utilisateur, mais permet de détecter la réception de données inhabituelles.
Dans le cadre de cet article, les champs suivants sont extraits de l'en-tête HTTP :

- *User-Agent* : chaîne de caractère arbitraire définie par le navigateur.
- *Accept, Accept-Language, Accept-Encoding* : préférences (valeurs $\in [0, 1]$) quant aux formats, langues, et encodages à utiliser.
- *Referer* : URL de la page précédente, parfois retiré, tronqué, ou aléatoire.

1. fonctionnalité HTML5

2. Internet Assigned Numbers Authority
3. Registres Internet Régionaux
4. Registres Internet Locaux
5. download.dp-ip.com/free/dbip-city-2017-05.csv.gz
6. DNS Blacklist

Figure 1 – *Principe de la méthode proposée*

— *Cookie* : cookies envoyés par le navigateur.
— *DNT, Connection, Upgrade-Insecure-Requests* : autres paramètres.

Données biométriques. La biométrie comportementale de l'utilisateur peut être analysée à partir de ses actions claviers et souris, décrits, dans le navigateur, par les événements JavaScript . Dans le cadre de cet article, la dynamique de frappe de l'utilisateur est représentée à partir des 20 digrammes les plus fréquents : *"r ", "te", "nt", ", ", "n ", "en", " s", "le", " l", " c", "de", ('arrowleft', 'arrowleft'), " p", " d", "on", "t ", "es", "s ", "e ", ('backspace', 'backspace')*. Plus précisément, les durées suivantes seront étudiées :

— P_1R_1 : d'appui du premier caractère.
— P_2R_2 : d'appui du second caractère.
— P_1P_2 : entre les pressions des deux caractères.
— R_1R_2 : entre les relâchements des deux caractères.
— R_1P_2 : entre le relâchement du premier caractère et la pression du second.
— P_1R_2 : entre la pression du premier caractère et le relâchement du second.

3.2 Pré-traitement des données

Afin d'obtenir pour chaque modalité, un vecteur de réels de taille fixe, les données collectées sont converties en vecteurs de réels, puis concaténées. La distance entre deux vecteurs pouvant fortement être influencée par les valeurs extrêmes, ces dernières sont normalisées.

Navigateur. Localkey, clé de n bits, est convertie en un vecteur de n bits. Ainsi, la localkey de 16 bits, *"0x0123"*, est convertie en *[0,0,0,0, 1,0,0,0, 0,1,0,0, 1,1,0,0]*.

Localisation. L'adresse IP est convertie en un vecteur composé :

— d'un vecteur composé des bits de l'adresse IP divisés par 2^{32-p-1} avec p, poids du bit.
— d'un vecteur de bits des $128/2^k$ premiers octets du hash md5 du nom de chaque localité avec k=1 pour "pays", k=2 pour "région", k=3 pour "département", et k=4 pour "ville".
— d'un vecteur de 3 angles $\in [-90; +90]$ représentant la latitude (lat) et la longitude l (lng1, lng2) de la localisation GPS. lng1 et lng2 valent $sign(\alpha) * ||\alpha| - (|\alpha| > 90) * 180|$ avec $\alpha = l$ pour lng1 et $\alpha = rot90(l) = (l - 90)\%360 - 180$ pour lng2. Ces angles sont normalisés par la formule suivante : $(angle + 90)/180$.

Données réseau. *Referer, User-Agent, Connection* et *Cookie* sont convertis en histogrammes, vecteurs donnant pour chaque caractère son effectif. Seuls les caractères ASCII $\in [0x20, 0x7F[$, soit 95 caractères, sont considérés. *Accept, Accept-Encoding,* et *Accept-Language* sont convertis en vecteurs donnant la préférence pour chaque format, encodage, et langue présents dans une liste prédéfinie. Une valeur supplémentaire indique la présence d'espaces après les virgules présentes dans le champ. *DNT* et *Upgrade-Unsecure-Requests* sont convertis en vecteurs de un entier valant 1 si positionné, 0 sinon. Les listes prédéfinies sont :

— Accept : "text/html", "application/xhtml+xml", "application/xml", "image/webp", "image/jxr" ;
— Accept-Encoding : "gzip", "deflate", "br", "sdch" ;
— Accept-Language : "fr", "fr-FR", "en-US", "en".

Données biométriques. Les durées collectées sont converties en un vecteur donnant, pour chaque digraphe considéré, les moyennes des 6 durées. Ces moyennes sont converties en millisecondes, limitées à 1000 puis divisées par 1000.

3.3 Protection des données

L'enjeu que nous souhaitons adresser dans ce travail est la possibilité de répondre à des applications de services numériques sur Internet (authentification, détection d'attaque, ...) tout en préservant le respect de la vie privée de l'individu. À partir des données personnelles collectées,

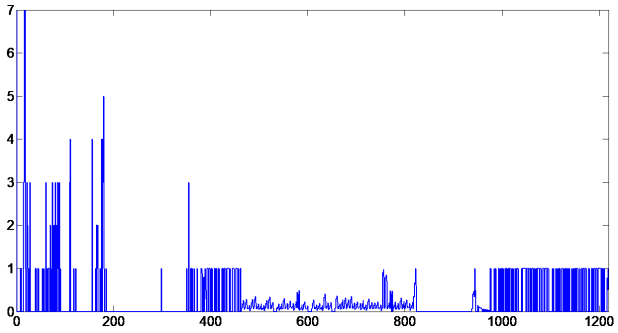


Figure 2 – *Un exemple des valeurs brutes après prétraitement (1218 nombres réels)*

nous souhaitons générer une signature binaire comme caractéristique dynamique d'un individu ayant perdu son caractère sémantique. Au final, le service numérique peut exploiter cette signature binaire sans connaître les informations utilisées pour la générer.

L'algorithme Biohashing est un algorithme bien connu dans le domaine de la biométrie. Il permet de transformer des données biométriques représentées par un vecteur à valeur réelle de longueur fixe et génère un modèle binaire appelé BioCode de longueur inférieure ou égale à la taille d'origine. Cette transformation est non inversible et permet de conserver la similarité des données en entrée. Cet algorithme a été initialement proposé pour le visage et les empreintes digitales par Teoh *et al.* dans [9]. L'algorithme de Biohashing est applicable sur toutes les modalités biométriques, voire données personnelles, pouvant être représentées par un vecteur de valeurs réelles de longueur fixe. Cette transformation nécessite un secret liée à l'utilisateur. Dans notre cas, il pourra s'agir d'un mot de passe saisi par l'utilisateur [5]. La comparaison des BioCodes est réalisée par le calcul de la distance de Hamming. L'algorithme de Biohashing transforme un vecteur de paramètres $T = (T_1, \dots T_n)$ dans un modèle binaire appelé BioCode $B = (B_1, \dots B_m)$, avec $m \leq n$, comme suit :

1. m vecteurs aléatoires orthonormés $V_1, \dots, V_m$ de la longueur n sont générés à partir d'un secret servant de germe du tirage aléatoire (typiquement avec l'algorithme de Gram Schmidt).
2. Pour $i = 1, \dots, m$, calcul du produit scalaire $x_i =< T, V_i >$.
3. Calcul du BioCode $B = (B_1, \dots, B_m)$ avec le processus de quantification :

$$B_i = \begin{cases} 0 & \text{if} \quad x_i < \tau \\ 1 & \text{if} \quad x_i \geq \tau, \end{cases}$$

Où τ est un seuil donné, généralement égal à 0.

La performance de cet algorithme est assurée par le produit scalaire avec les vecteurs orthonormés, tels que détaillés dans [8]. Le processus de quantification garanti la non-inversibilité des données (même si $n = m$), car chaque coordonnée de l'entrée T est une valeur réelle, alors que le BioCode B est binaire. Nous proposons d'utiliser cette transformation dans la protection des données personnelles.

4 Expérimentations

Dans cette partie, nous détaillons le protocole expérimental utilisé. Quelques résultats préliminaires sont donnés afin de montrer l'intérêt du calcul du code binaire.

4.1 Protocole expérimental

Une campagne de collecte a été organisée en mars 2017 sur le site trust.greyc.fr. Les participants ont été recrutés via les listes de diffusion du laboratoire GREYC et de l'école d'ingénieur ENSICAEN. De ce fait, les données collectées proviennent de membres sur un lieu assez unique. En effet, la majorité des participants sont localisés à Caen, utilisent les mêmes réseaux (ENSICAEN et UNICAEN), et ont donc la même adresse IP sortante. De plus, l'utilisation des postes des structures du GREYC et de l'ENSICAEN, font que les participants ont des configurations similaires, et ainsi des données réseau proches.

Avec seulement 22 participants, majoritairement localisés sur Caen, l'échantillon n'est pas représentatif, mais permet une première expérimentation du code personnel d'identité. Lors de la collecte, les participants sont invités à répondre à 8 questions relatives à la vie privée, puis à recopier un extrait de la Déclaration Universelle des Droits de l'Homme (voir 3). Afin d'éviter toute influence sur la dynamique de frappe au clavier, les participants ne sont informés de la collecte d'informations qu'à partir de l'étape 5 où ils sont invités à renseigner leurs informations personnelles. Toutes les données collectées sont stockées dans le sessionStorage du navigateur et ne sont soumises qu'après validation de l'utilisateur via la page de confirmation, présentant les types d'informations collectées, ainsi que le détail des informations collectées. Une fois les données soumises, une localkey est générée et stockée dans le localStorage du navigateur, ce afin de reconnaître ce dernier en cas de soumissions multiples. La localkey est aussi affichée à l'utilisateur afin qu'il puisse faire valoir ses droits quant à l'accès et la correction des données le concernant.

4.2 Résultats expérimentaux

A partir des 29 collectes issues de 22 personnes (8 ont été réalisées par la même personne dans des contextes différents), nous allons estimer dans quelle mesure ces informations permettent de mesurer une ressemblance des personnes. La figure 4 présente deux matrices de distance. La première (a) compare les données pré-traitées (sans protection) avec la distance du cosinus ($1 - cos(A, B)$, si A et B sont deux vecteurs de réels). Sur cette figure, nous pouvons constater deux choses. La première est que les

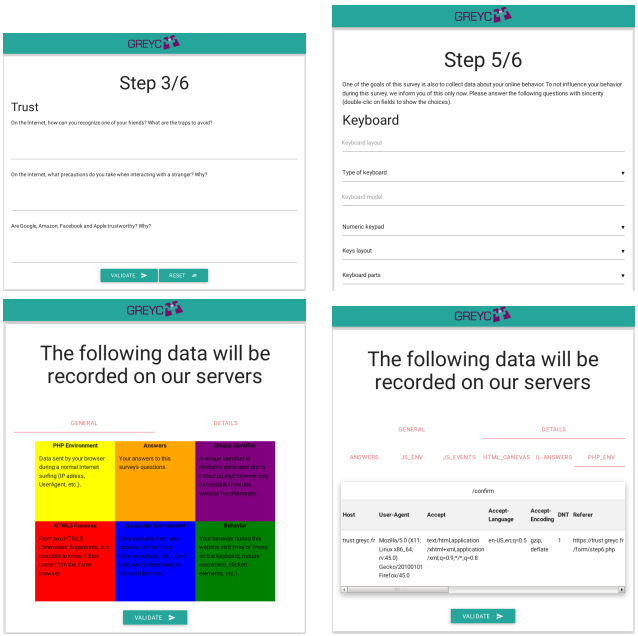


Figure 3 – *Écrans du questionnaire de collecte des données personnelles.*

signatures 4 et 5 sont jugés comme très similaires. Il s'agit en fait de la même personne dans le même contexte. La seule différence est la dynamique de frappe au clavier. Les signatures 3 à 10 ont été générées par le même individu mais dans des contextes différents (wifi, navigateur, ...), la ressemblance est plus contrastée. La seconde constatation importante est la relative similarité des signatures 4 et 5 avec d'autres signatures du tableau. Ceci peut s'expliquer par le fait que les données ont été acquises au sein du laboratoire par du matériel ayant une configuration proche et la même adresse IP sortante.

La figure 4 (b) présente la distance entre les codes binaires (signatures protégées) de la base en prenant pour chaque individu une clé secrète unique. Avec la protection et cette clé, on met en évidence très clairement la similarité entre individus. Pour les codes binaires reliés aux signatures de 3 à 10, on identifie bien une similarité en eux avec des degrés plus ou moins élevés en fonction des données personnelles similaires. Ceci démontre bien la capacité de la méthode proposée à produire un code exploitable pour des calculs de similarité d'informations personnelles.

Concernant les exigences énoncées au début, il est assez facile de vérifier qu'elles sont respectées. La transformation du BioHashing garantie la non inversibilité du code binaire calculé et le respect de la similarité. La confidentialité des données est obtenue par cette dernière transformation et l'usage d'une clé secrète (ici un mot de passe). Un imposteur ne pourra pas générer ce code binaire sans connaître la clé secrète, utiliser le même matériel.... Il pourra tout au mieux rejouer une donnée existante, des mécanismes de protection du canal de communication et de chiffrement de la donnée côté service peuvent résoudre ce problème. La répudiation du code est aisée en changeant de clé secrète (i.e. ici, mot de passe).

5 Conclusion et perspectives

Dans ce papier, nous proposons une méthode permettant de calculer un code personnel lié à un internaute respectueux de la vie privée. Ce code intègre différentes informations liées à son navigateur, sa façon de taper au clavier, ou sa localisation. Nous avons montré sur une base préliminaire de 29 collectes qu'il était possible d'obtenir un code binaire proche pour la même personne malgré des différences de contexte. Plusieurs applications sont envisageables à ce travail dont l'authentification d'un internaute, l'usage pour identifier des comptes multiples par un service (similarité de codes calculés avec une clé unique). Ces applications constituent les perspectives de cette étude.

Références

[1] Tableau de bord des services publics numériques – Édition 2017. http ://www.modernisation.gouv.fr/ladministration-change-avec-le-numerique/par-des-services-numeriques-aux-usagers/tableau-de-bord-des-services-publics-numeriques-edition-2017.

[2] Károly Boda, Ádám Földes, Gábor Gulyás, and Sándor Imre. User tracking on the web via cross-browser fingerprinting. *Information Security Technology for Applications*, pages 31–46.

[3] SL Yinzhi Cao and E Wijmans. Browser fingerprinting via os and hardware level features. *Network & Distributed System Security Symposium, NDSS*, 17, 2017.

[4] Peter Eckersley. How unique is your web browser ? *Privacy Enhancing Technologies*, 6205 :1–18, 2010.

[5] Patrick Lacharme and Aude Plateaux. Pin-based cancelable biometrics. *International Journal of Automated Identification Technology (IJAIT)*, 3(2) :75–79, 2011.

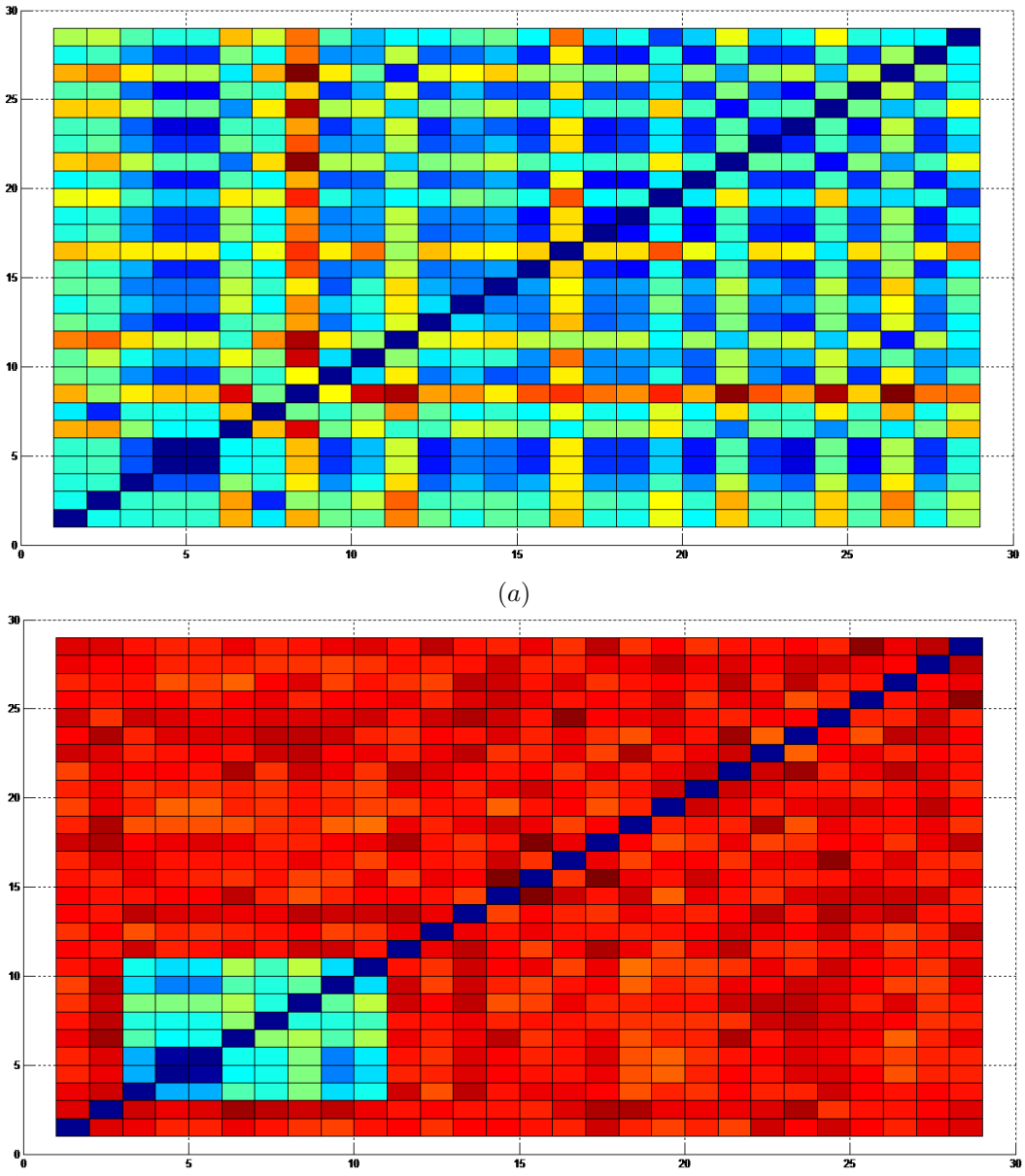
[6] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. Beauty and the beast : Diverting modern web browsers to build unique browser fingerprints. *Security and Privacy (SP)*, pages 878–894, 2016.

[7] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. Privaricator : Deceiving fingerprinters with little white lies. *Proceedings of the 24th International Conference on World Wide Web*, pages 820–830, 2015.

[8] A. B.J. Teoh, Y. W. Kuan, and S. Lee. Cancellable biometrics and annotations on biohash. *Pattern Recognition*, 41 :2034–2044, 2008.

[9] A.B.J. Teoh, D. Ngo, and A. Goh. Biohashing : two factor authentication featuring fingerprint data and tokenised random number. *Pattern recognition*, 40, 2004.

[10] Zachary Weinberg, Eric Y Chen, Pavithra Ramesh Jayaraman, and Collin Jackson. I still know what you visited last summer : Leaking browsing history via user interaction and side channel attacks. *Security and Privacy (SP)*, 2011.



(*a*)

(*b*)

Figure 4 – *Représentation de la distance entre les données pré-traitées (a) et après protection (b).*

Tatouage et codes en métrique rang

Pascal Lefèvre
Laboratoire XLIM
UMR CNRS 7252
Université de Poitiers
BP 30179, 86962 Futuroscope
pascal.lefevre@univ-poitiers.fr

Philippe Carré
Laboratoire XLIM
UMR CNRS 7252
Université de Poitiers
BP 30179, 86962 Futuroscope
philippe.carre@univ-poitiers.fr

Philippe Gaborit
Laboratoire XLIM
UMR CNRS 7252
Université de Limoges
87060 LIMOGES CEDEX
philippe.gaborit@xlim.fr

Résumé

Nous présentons une manière différente pour améliorer la robustesse du tatouage numérique. En utilisant la méthode Lattice QIM dans le domaine spatial des images en niveau de gris, nous analysons l'intérêt d'utiliser un nouveau type de codes correcteurs appliqués codes en métrique rang. Ces codes sont très utilisés en télécommunication dans le cadre du codage de réseaux mais sont, selon nous, encore inconnus dans le domaine du tatouage numérique. Dans cet article, nous montrons comment cette métrique permet de corriger des erreurs d'une structure spécifique et donc comment elle est adaptée pour contrer un certain type d'attaque. Dans cet article, nous proposons une première étude pour valider le concept de la métrique pour le tatouage numérique. Pour cela, nous utilisons ces codes pour obtenir une résistance théoriquement parfaite aux modifications constantes de luminance.

Mots clefs

Tatouage, codes correcteurs, métrique rang, erreurs structurées.

1 Introduction

Le tatouage numérique est un domaine de recherche important à cause la pleine expansion de l'utilisation de contenus multimédias en ligne. Pour assurer la protection des droits et la propriété intellectuelle de tels documents contre la distribution massive de ceux-ci sur internet, nous avons besoin de moyens efficaces pour contrôler cette communication, mettre un terme aux manipulations et duplications non autorisées qu'elles soient malicieuses ou non.

Pour être efficace, un tatouage doit être imperceptible, doit contenir autant d'information que possible et doit être robuste [1] aux attaques (fiabilité de transmission sur un canal non fiable) les plus communes tout en assurant une transmission sécurisée [2].

Un des outils les plus efficaces pour augmenter la robustesse d'une marque est d'utiliser des codes correcteurs d'erreur qui vont permettre de corriger les erreurs produites par une attaque donnée. En fonction de l'attaque et de la structure des erreurs induite par celle-ci, le type de codes utilisé sera plus ou moins efficace. Par exemple, si l'erreur induite sur la marque est aléatoire, les meilleurs résultats sont obtenus en utilisant des codes binaires comme par exemple les codes BCH.

Pour d'autres attaques, il arrive que les erreurs arrivent par paquet. Dans ce cas, il est préférable d'utiliser des codes plus structurés c'est à dire des codes définis sur un alphabet plus grand (par exemple $GF(2^m)$), tels que les codes de Reed-Solomon où il est possible de décoder les erreurs par paquet [3].

Alors, les erreurs ne se décodent plus indépendemment sur chaque bit mais sur des paquets de m bits de façon à ce que plusieurs erreurs dans le même paquet binaire ne compte que pour une seule erreur (une erreur de symbôle) dans un mot de code.

Ainsi, en fonction de l'attaque i.e. des erreurs produites, nous pouvons choisir un code correcteur adapté. C'est une idée déjà bien connue et a permis de nombreuses applications industrielles de ces codes utilisant la distance de Hamming.

Dans cet article, nous considérons l'utilisation d'une nouvelle métrique appelée la *métrique rang*. Ces codes sont déjà très utilisés en télécommunication dans le cadre du codage de réseaux [4] et en cryptographie [5]. Ils sont capables de corriger des erreurs d'une structure spécifique. Si on considère un code sur $GF(2^m)$ de taille m, chaque coordonnée d'un mot de code sur $GF(2^m)$ est encodée sur m bits et puisque le code est de longueur m, tout mot de code peut être vu comme une matrice de taille $m \times m$.

Comme la métrique utilisée est le rang d'une matrice binaire, la condition de décodage s'exprime avec cette métrique c'est à dire que les mots de code reçus dont les erreurs ont un rang faible peuvent être corrigés correctement. Par exemple, une attaque qui inverse tous les bits d'un marque (c'est à dire d'un mot de code) ne pourra pas du tout être corrigé par un code de Hamming. Par contre, avec des codes en métrique rang, l'erreur sera de rang 1 car l'erreur est une matrice remplie de symbôles binaires 1 et ainsi, on retrouve sans difficulté le mot de code qui a été transmis.

Notre contribution : Nous proposons d'introduire ce type de correction dans un schéma de tatouage. Nous commençons par définir les codes en métrique rang (section 2) puis

la méthode Lattice QIM (noté LQIM) (section 3). Ensuite, nous proposons une méthode de tatouage combinant la méthode Lattice QIM et les codes en métrique rang. Nous montrons que les modifications de luminance produisent une erreur structurée adaptée à la métrique rang. De plus, nous expliquons pourquoi cette structure existe et comment le décodeur LQIM est amélioré pour un décodage sans erreur.

2 Codes correcteur en métrique rang

2.1 Définitions et propriétés

Nous pourrons nous référer à [6] pour plus de détails sur les codes en métrique rang. Soient $B = (\alpha_1, ..., \alpha_m)$ une base de $GF(q^m)$ sur $GF(q)$ et $x = (x_1, ..., x_n) \in GF(q^m)^n$.
Le rang de x sur $GF(q)$ est le rang de la matrice $X = (x_{ij})$, où $x_{ij} = \sum_{i=1}^{m} x_{ij}\alpha_i$. Il est noté $Rank(x|GF(q))$, ou $Rank(x)$ quand il n'y a pas d'ambiguité.
Soient x et y deux éléments de $GF(q^m)^n$. On définit la distance rang d_R entre x et y tel que

$$d_R(x, y) = Rank(x - y).$$

Un code linéaire en métrique rang $\mathcal{C}$ de longueur n et de dimension k sur $GF(q^m)$ est un sous-espace de $GF(q^m)^n$.
La distance rang minimale de $\mathcal{C}$ est définie telle que :

$$d = \min(d_R(c_1, c_2), c_1, c_2 \in \mathcal{C}|c_1 \neq c_2)$$

2.2 Decodage d'un code en métrique rang

Les bornes de décodage classiques pour la distance de Hamming peuvent être adaptées pour celles de la métrique rang. Si un code en métrique rang $\mathcal{C}$ possède une distance minimale d et que, par exemple, un vecteur $y = c + e$ est reçu tel que $c \in \mathcal{C}$ et e un vecteur erreur de rang inférieur à $(d-1)/2$, alors il est possible d'avoir un décodage unique de y en c.
Contrairement aux codes de Hamming, on ne connaît que très peu de familles de codes en métrique rang dont le décodage est facile. Les codes de Gabidulin est une de ces familles et admettent pour paramètres $[n, k, n - k + 1]$ sur $GF(q^n)$. Ces codes peuvent tolérer un maximum de $(n-k)/2$ erreurs et peuvent être vu comme une famille de code analogue à celle des codes de Reed-Solomon. Différentes approches pour le décodage des codes de Gabidulin existent telles que [7, 8].

2.3 Les codes en métrique rang en pratique

En pratique, nous utilisons ces codes dans une extension $GF(q^m)$ de $GF(2)$ et nous pouvons associer un vecteur binaire de taille m à chaque coordonnée d'un mot de code de telle sorte qu'un mot de code c peut être représenté par une matrice binaire de taille $m \times m$.
Après une modification du tatouage, le mot de code inséré c est modifié par une erreur e qui peut aussi être vu comme une matrice binaire de taille $m \times m$.

Pour évaluer dans quel cas la métrique rang est plus efficace que la métrique de Hamming classique, nous comparons les comportements de ces dernières en observant quelques exemples d'erreurs. Soient $y = c + e$ un mot de code reçu après extraction d'une marque.

Exemple : $m = 4$ et un mot de code c :

$$c = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

Avec y tel que :

$$y = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

On a l'erreur suivante :

$$e = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

On voit que la matrice erreur e est de rang 2. Si un code en métrique rang est capable de corriger jusqu'à 2 erreurs alors on décode de manière unique y en c. Avec un code de Hamming de longueur 16, on aurait une erreur de poids égal à 4. Dans ce cas là, il est possible de trouver des codes performants avec les deux métriques pour une dimension raisonnable k.
Indépendemment de y et c, si on avait une erreur tel que :

$$e = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

En comparant de la même façon que dans le paragraphe précédent, le poids de Hamming de e est de 9 alors que le rang de e est de 4. Dans les deux cas, il est impossible de trouver des codes performants avec les deux métriques. Si e était la matrice identité, on aurait une erreur de rang plein et uniquement les codes de Hamming auraient été utiles.
En fait, les codes en métrique rang deviennent plus intéressants lorsque l'erreur possède une structure particulière. Par exemple :

$$e = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Avec la métrique de Hamming, on a 9 erreurs sur 16 bits transmis et il n'existe pas de code capable de décoder correctement alors qu'avec la métrique rang, e est de rang 1 seulement. Nous pouvons donc facilement décoder un telle erreur (par exemple, avec un code de paramètres $[4, 2, 3]$).

Bien sûr, ce type d'erreur se retrouve dans le cadre de certaines attaques. Les codes en métrique rang seraient, dans ce cas, une alternative d'application au tatouage numérique plus performante que les codes classiques de Hamming.
Nous allons maintenant présenter une méthode de tatouage pouvant être adaptée avec les codes en métrique rang afin de se protéger contre certains types d'attaques.

3 Lattice QIM (LQIM)

La quantification vectorielle a été introduite par B. Chen and Gregory W. Wornel ([9, 10]). A l'étape d'insertion, nous avons deux cosets du réseau euclidien ($\Delta\mathbb{Z}^L$) de dimension L et un quantificateur Q_m :

$$\Lambda_0 = -\frac{\Delta}{4} + \Delta\mathbb{Z}^L, \Lambda_1 = \frac{\Delta}{4} + \Delta\mathbb{Z}^L \tag{1a}$$

$$y = Q_m(x, \Delta) = \lfloor x/\Delta \rfloor \Delta + (-1)^{m+1}\Delta/4 \tag{1b}$$

avec $m = 0, 1$, x un échantillon hôte et y le résultat de la quantification de x.

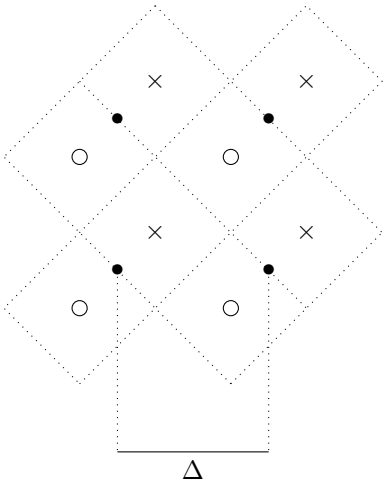


Figure 1 – *Représentation de l'espace de quantification (ou réseau euclidien) en dimension $L = 2$. Les croix représentent la quantification des vecteurs portant le bit 1 et les cercles pour le bit 0.*

Nous nous basons sur la figure 1 pour visualiser un exemple d'espace de quantification en dimension 2. Pour toute croix ou cercle y, les losanges en pointillés délimitent les frontières de chaque cellule de quantification. Quand un échantillon hôte x est quantifié, il est transformé en un y le plus proche (en fonction du bit d'information à insérer).
Pour l'étape de détection, on calcule lequel des deux cosets est le plus proche du vecteur reçu z :

$$\hat{m} = \underset{m\in\{0,1\}}{\arg\min}\, dist(z, \Lambda_m), \tag{2a}$$

$$dist(z, \Lambda) = \min_{y\in\Lambda} \|z - y\|_2 \tag{2b}$$

Ici, z est une version modifiée du vecteur quantifié y. Ainsi, z se décode en calculant le centre de cellule y dans laquelle z se trouve.

Dans la prochaine section, nous proposons d'observer la structure des erreur produites par la modification de luminance sur une marque issue de LQIM combinée avec un mot de code en métrique rang.

4 Étude de l'attaque luminance

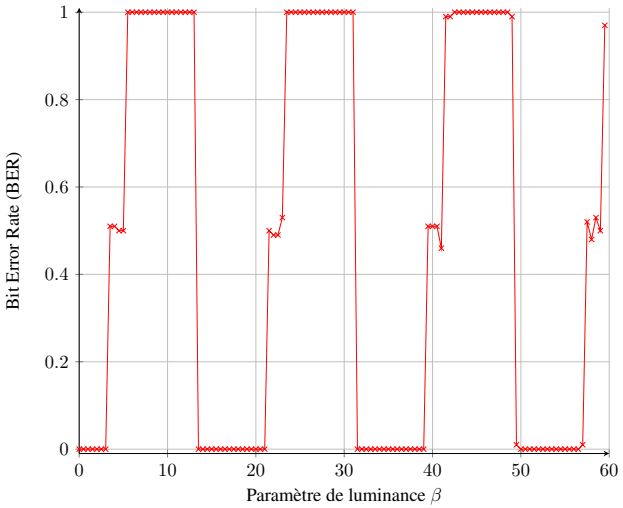


Figure 2 – *Taux d'erreur binaire de la méthode LQIM dans le domaine spatial en fonction d'une modification additive de luminance de paramètre β. Ici, nous avons des valeurs positives de β mais la courbe se comporte de la même manière pour des valeurs négatives.*

4.1 Protocole

Les expériences ont été faites sur la base d'image Corel où 1000 images ont été choisies équiprobablement dans les 10000 images disponibles.
Chaque message inséré est de longueur 49 et est généré aléatoirement, la dimension de l'espace de quantification est de $L = 6$ et le pas de quantification est $\Delta = 16$. Ces paramètres ont été fixé ainsi pour maintenir une qualité d'image fixe (DWR $\geq 30db$).
Pour la première expérience, nous montrons que les taux d'erreur binaire (BER) entre le message original (qui représente un mot de code en métrique rang sur $GF(2^7)$ et de paramètres $[7, 3, 5]$) et le message décodé en fonction du paramètre β de l'attaque de luminance.
L'équation suivante modélise les effets d'une modification de luminance sur un vecteur y :

$$z = y + \beta \tag{3}$$

avec z une version modifiée de y et β le paramètre de l'attaque de luminance. Quand β augmente, le vecteur z sature de plus en plus. D'un point de vue géométrique (en 2D pour la figure 1 par exemple), z se déplace d'une cellule de quantification vers une autre alternant le bit inséré a chaque changement de cellule.

4.2 Analyse de la détection après attaque

Dans la figure 2, nous observons une courbe de taux d'erreur ressemblant à une sinusoïde carré. On distingue trois cas : $\beta = 0, 0.5, 1$.
Dans le premier cas, nous n'avons pas d'erreur au décodage sur certains intervalles (par exemple $\beta \in [14, 21[$). Ensuite, le troisième cas est similaire au premier c'est à dire qu'un taux d'erreur égal à 1 se produit aussi par intervale (par exemple $\beta \in [5, 14[$), sauf que le taux d'erreur indique que tous les bits du message original ont été inversés.
Dans le second cas, le taux d'erreur indique que l'on a décodé des séquences aléatoires. Contrairement aux autres cas, celui-ci ne se produit que ponctuellement autour d'une valeur (comme $\beta = 4, 21, 40, ...$).
Cette courbe montre clairement l'existence d'une structure partielle de l'erreur et que les codes en métrique rang sont adaptées pour gérer ce type d'erreur. Dans la sous-section suivante, nous expliquons la structure de l'erreur et pourquoi elle se produit.

4.3 Structure de l'erreur

Quand une image subit une modification de luminance, tous les vecteurs quantifiés y subissent la même distorsion β de la forme :

$$\beta = \beta' \times (1, ..., 1) \in \mathbb{R}^L, \beta' \in \mathbb{R}. \quad (4)$$

En d'autre terme, si nous imaginons l'espace de quantification comme dans la figure 1, tous les vecteurs modifiés z vont voyager à travers les cellules de quantification de la même façon. Quand le taux d'erreur est nul, tous les vecteurs modifiés z se trouvent dans une cellule représentant le bit associé original (pas nécessairement la cellule originale). Le raisonnement est le même quand le taux d'erreur est égal à 1. Pour le cas où le taux d'erreur est à 0.5, chaque vecteur z se trouve d'un côté ou de l'autre d'une frontière de cellule.
Nous allons montrer dans la sous-section suivante comment un code en métrique rang permet de supprimer une grande majorité des erreurs.

4.4 Application des codes en métriques rang

Dans cette seconde expérience, nous avons combiné la méthode LQIM avec un code correcteur en métrique rang en insérant un mot de code en tant que signature. Les paramètres du code sont $[7, 3, 5]$ et celui-ci peut corriger des erreurs de rang 2. Cette fois, nous mesurons des taux d'erreur image c'est à dire le ratio d'images pour lesquelles on ne décode pas le bon message (i.e. le rang de l'erreur est strictement supérieur à 2).
Dans la figure 3, nous pouvons voir que le code en métrique rang est très efficace car le taux d'erreur image est nul presque partout. Pour certaine valeur de beta, aux valeurs où le taux d'erreur binaire est égal à 1, aucune marque correcte ne peut être extraite car l'erreur n'est plus structurée (on obtient une erreur de rang plein).

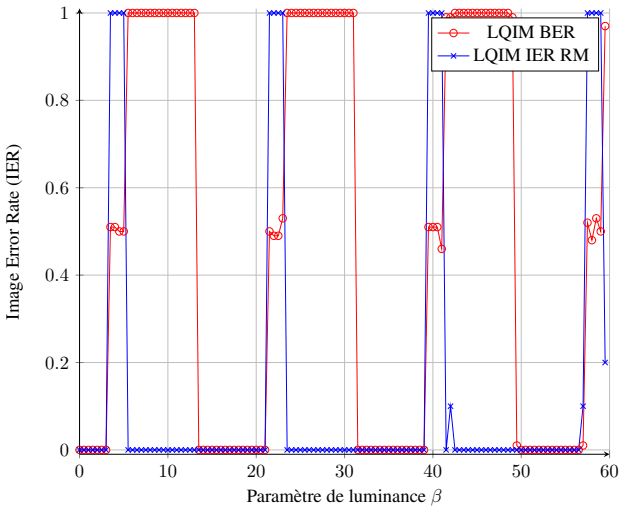


Figure 3 – *Taux d'erreur image de la méthode LQIM combinée avec une code correcteur en métrique rang dans le domaine spatial en fonction du paramètre de luminance β. Chaque point de la courbe bleue représente le taux d'image dont le décodage a échoué (rang de l'erreur supérieur à 2). Nous pouvons voir que lorsque le taux d'erreur binaire est égal à 0.5, le taux d'erreur image est égal 1 ce qui illustre bien les capacités de correction du code en métrique rang.*

Au décodage, on ne peut pas deviner la valeur de β en pratique ce qui peut mener à des problèmes de décodage. La probabilité de tomber sur une valeur de β pour laquelle l'erreur n'est pas structurée pour la métrique rang dépends du pas de quantification choisi Δ : plus il est petit, plus il est probable de ne pas pouvoir décoder.
De plus, on a pu voir que la courbe de taux d'erreur binaire n'a pas toujours une forme de sinusoïde carrée (par exemple pour $\beta = 43$). Ceci est du à la nature aléatoire des valeurs de pixel que contiennent les images c'est à dire qu'ils arrivent que certaines images saturent pour de petites valeurs de β. Un code capable de corriger avec des erreurs de rang 2 a été choisi mais, en théorie, un code corrigeant des erreurs de rang 1 suffit.
Dans le cas d'une modification de luminance d'une image, les codes en métrique rang permettent de corriger presque parfaitement les erreurs produites cette attaque. Dans la sous-section suivante, nous proposons une amélioration du décodeur LQIM en métrique rang pour éliminer les erreurs restantes.

4.5 Optimisation du décodeur LQIM + métrique rang

Un changement de luminance est paramétré par la constante β. Supposons que l'on ait récupéré une image marquée et endommagée par une modification de luminance. Au décodage, nous avons :

$$z = y + \beta \quad (5)$$

avec z les versions modifiées des vecteurs quantifiés y. En partant de l'équation du dessus, il est possible d'optimiser les performances de décodage en modifiant de manière controlée la luminance de l'image.
Tout d'abord, nous pouvons remarquer que les valeurs de β pour lesquelles nous ne pouvons pas décoder se trouvent à intervalle régulier. Ces valeurs de β représentent les transitions des vecteurs z d'une cellule à une autre. D'après la section 3, la construction des cosets permet d'en déduire que ces valeurs sont en fait des multiples de $\sqrt{2}\Delta/4$ (moitié de la distance entre un cercle et une croix les plus proches).

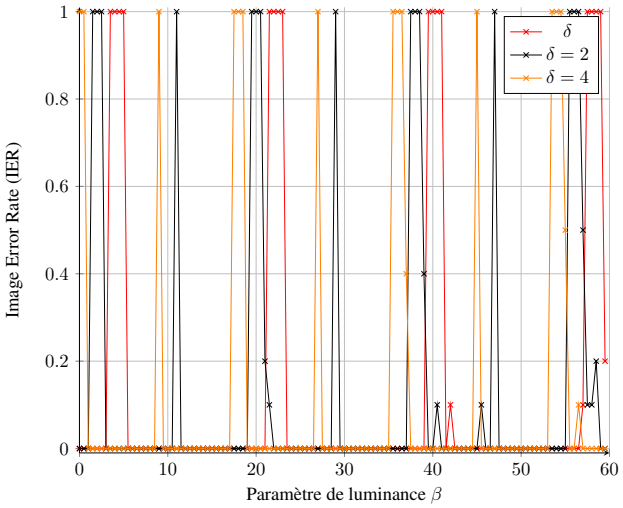


Figure 4 – *Taux d'erreur image de la méthode LQIM combinée avec une code correcteur en métrique rang dans le domaine spatial en fonction du paramètre de luminance β pour 3 valeurs de δ $(0, 2, 4)$. Nous pouvons voir que les distortions controllées (2 et 4) permettent de décaler les piques d'erreur.*

Soient $\delta_1, ..., \delta_n$ des entiers positifs plus petit que $\sqrt{2}\Delta/4$. D'après l'observation des différentes courbes de taux d'erreur image (figure 4) et les taux d'erreur binaire associés(figure 4), on voit qu'en dégradant volontairement l'image, les courbes d'erreur dont décalées. Nous pouvons donc en déduire la propriété suivante :

$\exists i$ unique tel que l'image modifiée $z + \delta_i$ ne peut être décodée avec le décodeur LQIM + métrique rang

et

$\forall j \neq i$, l'image modifiée $z + \delta_j$ est parfaitement décodée avec le décodeur LQIM + métrique rang

En modifiant de manière controllée la luminance de l'image reçue avec de petites valeurs δ, nous pouvons garantir que pour une valeur de β fixée et un vecteur z fixé, la majorité des $z+\delta_j$ (z est inclu dans cette liste) auront un décodage correct c'est à dire que la majorité des bits décodés seront identiques qu'à ceux insérés au départ. Avec cette

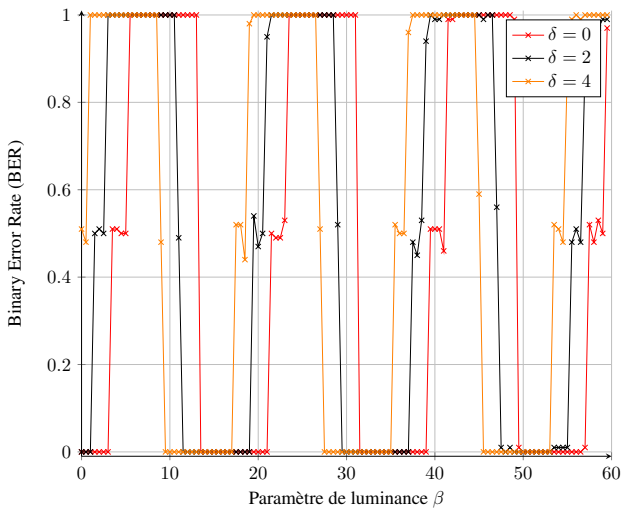


Figure 5 – *Taux d'erreur binaire du décodeur LQIM + métrique rang amélioré dans le domaine spatial en fonction du paramètre de luminance β.*

stratégie, nous pouvons donc décoder sans erreur (suppression des piques) au prix d'un temps de décodage n fois plus long. Une valeur de $n = 3$ suffit pour décoder sans erreur avec :

$$\begin{cases} d = \frac{\sqrt{2}}{4}\Delta \\ \delta_1 = 0 \\ \delta_2 = \frac{1}{3}d \\ \delta_3 = \frac{2}{3}d \end{cases}$$

Exemple de décodage : On a les valeurs suivantes :

$$\begin{cases} d \simeq 6 \\ \delta_1 = 0 \\ \delta_2 \simeq 2 \\ \delta_3 \simeq 4 \end{cases}$$

qui représentent les versions modifiées de l'image transmise z. On extrait alors 3 estimations du message original m_1, m_2 et m_3. D'après la propriété, deux des trois messages extraits sont corrects pour β fixé.
Le résultat du décodage amélioré est illustré par l'expérience de la figure 6. Nous observons que les taux d'erreur sont nuls quasiment partout. Cependant, cela ne montre pas que la méthode est inefficace pour certaines valeurs de β car il arrive que certaines images marquées perdent de l'information suite à la modification de luminance.

Pour résumer, nous avons proposé une stratégie pour améliorer le décodeur LQIM couplé avec le décodage d'un mot de code en métrique rang. Le cas où le taux d'erreur binaire est égal à 0.5 peut être éviter en prennant une estimation moyenne du décodage de plusieurs images dont la luminance a été volontairement modifiées à partir de l'image

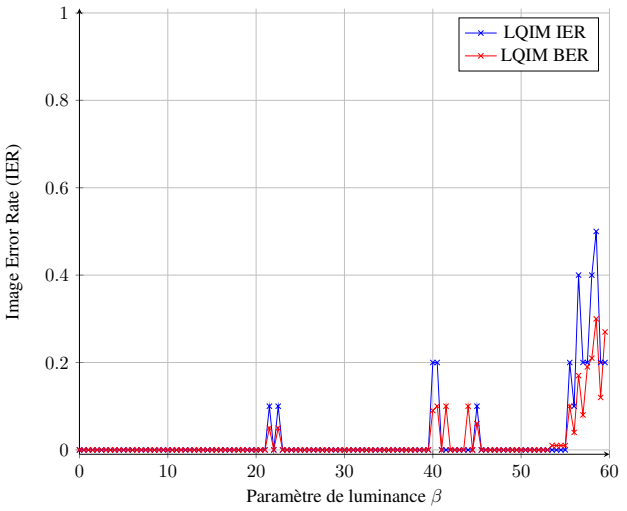


Figure 6 – *Taux d'erreur binaire et taux d'erreur image associé du décodeur LQIM + métrique rang amélioré dans le domaine spatial en fonction du paramètre de luminance β.*

transmise (les piques observables sur les courbes de taux d'erreur image ont été supprimées).

5 Conclusion

Nous avons présenté un nouveau type de codes correcteurs d'erreur utilisant la métrique rang au lieu de la métrique de Hamming. Ensuite, nous avons étudié la robustesse de l'association de la méthode de tatouage LQIM avec les codes en métrique rang face aux modifications de luminance. Nous avons vu que les erreurs produites suite à cette attaque admettent une structure d'erreur partielle qui est très bien gérée par les codes en métrique rang.

Grâce à notre analyse sur l'attaque luminance, nous avons montré que les codes en métrique rang peuvent avoir un grand pouvoir de correction si la structure de l'erreur est adaptée. Une première application de ces codes avec la méthode LQIM a donné de bons résultats de robustesse mais il reste tout de même quelques erreurs à corriger (apparition de piques IER). C'est pourquoi nous avons proposé une amélioration de la méthode de décodage de la méthode LQIM et qui a permis de pouvoir décoder une signature sans erreur.

Dans cet article, nous avons montré que les codes en métrique peuvent être appliqué au tatouage numérique. Bien sûr ce résultat se limite à contrer l'attaque de luminance à cause de la structure d'erreur produite. En terme de perspective, nous comptons approfondir le sujet pour découvrir comment les codes en métrique rang peuvent devenir efficace contre d'autres d'attaques.

Remerciements

Les auteurs remercient la Délégation Générale de l'Armement pour son financement.

Références

[1] Matt L Miller, Ingemar J Cox, Jean-Paul MG Linnartz, et Ton Kalker. A review of watermarking principles and practices. *Digital signal processing in multimedia systems*, pages 461–485, 1999.

[2] F. Cayre, C. Fontaine, et T. Furon. Watermarking security : theory and practice. *IEEE Transactions on Signal Processing*, 53(10) :3976–3987, Oct 2005.

[3] Wadood Abdul, Philippe Carré, et Philippe Gaborit. Error correcting codes for robust color wavelet watermarking. *EURASIP Journal on Information Security*, 2013(1) :1, Feb 2013.

[4] Danilo Silva et Frank R Kschischang. On metrics for error correction in network coding. *IEEE Transactions on Information Theory*, 55(12) :5479–5490, 2009.

[5] Philippe Gaborit, Olivier Ruatta, Julien Schrek, et Gilles Zémor. New results for rank-based cryptography. Dans *International Conference on Cryptology in Africa*, pages 1–12. Springer, 2014.

[6] Ernest Mukhamedovich Gabidulin. Theory of codes with maximum rank distance. *Problemy Peredachi Informatsii*, 21(1) :3–16, 1985.

[7] Ernst M. Gabidulin. *A fast matrix decoding algorithm for rank-error-correcting codes*, pages 126–133. Springer Berlin Heidelberg, Berlin, Heidelberg, 1992.

[8] Pierre Loidreau. *A Welch–Berlekamp Like Algorithm for Decoding Gabidulin Codes*, pages 36–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[9] Brian Chen et Gregory W. Wornell. Quantization index modulation : A class of provably good methods for digital watermarking and information embedding. *IEEE TRANS. ON INFORMATION THEORY*, 47(4) :1423–1443, 1999.

[10] P. Moulin et R. Koetter. Data-hiding codes. *Proceedings of the IEEE*, 93(12) :2083–2126, Dec 2005.

Anti-spoofing en reconnaissance faciale avec la mesure de qualité des images

Emna Fourati, Wael Elloumi
Worldline
19 Rue de la Vallée Maillard, Blois, France
{emna.fourati, wael.elloumi}@worldline.com

Aladine Chetouani
Laboratoire PRISME, Université d'Orléans
12 rue de Blois, Orléans, France
aladine.chetouani@univ-orleans.fr

Résumé

Les techniques d'authentification se basant sur la reconnaissance faciale peuvent être facilement contournées via différents types d'attaques. Les contre-mesures efficaces doivent répondre à certaines exigences en termes de robustesse et de rapidité. Dans ce papier, nous envisageons de trouver le meilleur compromis entre ces deux critères, en exploitant les Mesures de Qualité des Images (MQI) pour différencier entre les apparences réelles et reproduites du visage. Testée sur une base de données publique, notre solution montre une amélioration des performances par rapport aux approches de référence.

Mots clefs

Mesures de qualité des images, anti-spoofing, authentification biométrique faciale.

1 Introduction

La reconnaissance faciale est une modalité biométrique qui a récemment suscité beaucoup d'intérêt pour des objectifs d'identification et d'authentification. Grâce à sa commodité, cette technologie est de plus en plus intégrée dans la vie quotidienne des utilisateurs. En outre, vu qu'elle n'exige qu'une caméra frontale, elle a préparé le terrain pour l'authentification mobile basée sur les selfies, pour être adoptée ainsi par plusieurs applications critiques, notamment les services bancaires mobiles, les applications de paiement, et le contrôle aux frontières. Toutefois, la sécurité des systèmes de reconnaissance faciale représente le souci majeur des utilisateurs. La détection des attaques de spoofing (ou attaques de présentation) reste un défi stimulant. Les attaques de spoofing contournent le capteur biométrique en présentant de fausses données biométriques d'un utilisateur valide. En reconnaissance faciale, on utilise souvent les photos imprimées, des vidéos, ou des masques 3D dont la reproduction est devenue accessibles et peu coûteuse.

Etant donné tous ces risques de reconnaissance faciale, ce travail étudie la détection de propriétés de vivacité du visage en se basant sur les Mesures de Qualité des Images (MQI). Nos contributions adressent deux volets principaux : la sélection des métriques adéquates de MQI et l'approche utilisée pour les calculer.

Le reste de ce papier est organisé comme suit : la section 2 étudie les travaux connexes. La section 3 présente un aperçu sur les différentes phases de notre approche. Les résultats expérimentaux sont détaillés dans la section 4, avant de conclure à la section 5.

2 Travaux connexes

Les méthodes biométriques d'anti-spoofing peuvent être classifiées en trois catégories, selon le niveau du système biométrique : Niveau de capteurs (hardware), niveau de descripteurs (software) ou niveau de score (software et hardware) [1]. Dans cette étude, nous nous focalisons uniquement sur les méthodes software, intégrées dans le module d'extraction des descripteurs après l'acquisition de la donnée biométrique. Elles englobent des contre-mesures dynamiques, statiques, et multimodales.

Les contre-mesures dynamiques analysent le mouvement d'une séquence vidéo d'un visage afin de détecter des signes de vivacité physiologiques comme le clignement des yeux [2], le mouvement des yeux [3] ou le mouvement des lèvres [4]. Des indicateurs de mouvement peuvent aussi se baser sur des mouvements subtils entre les parties du visage [5], la corrélation globale entre les régions du visage et du fond [6], ou l'estimation de mouvement pour détecter les attaques planaires tels que des imprimés ou des écrans [7]. D'autres approches dynamiques analysent la structure 3D du visage [8,9], ou demandent à l'utilisateur d'effectuer une action spécifique telle qu'une expression faciale [10], une rotation de la tête [9] ou le mouvement de la bouche [4]. Cependant, l'exigence de la collaboration de l'utilisateur n'est pas adaptée aux tendances non-intrusives des systèmes d'authentification biométrique.

Les méthodes statiques se focalisent sur une seule instance du visage au lieu des données vidéo. Ces techniques sont

généralement plus rapides et moins intrusives que les approches dynamiques, elles n'exigent pas la collaboration de l'utilisateur. L'idée clé est qu'un faux visage est susceptible d'avoir une qualité inférieure à celle d'un visage réel. Ainsi, la majorité des méthodes statiques se basent sur l'analyse de l'apparence faciale, ou méthodes d'analyse de texture, en utilisant les informations de texture ou des MQI. La première peut se baser sur des descripteurs comme les motifs binaires locaux (Local-binary patterns : LBP, la transformé de Fourier, les dérivées de Gaussiennes (derivatives of Gaussians : DoG), et l'histogramme des gradients orientés (Histograms of oriented Gradients : HoG). D'autres descripteurs ont été exploités dans [11] pour analyser l'information mutuelle entre la couleur et la texture dans différents espaces de couleurs. La deuxième approche, basée sur les MQI, peut être divisée en trois catégories : Sans Référence (No-Reference : NR), avec référence (Full-Reference : FR) et à référence réduite (Reduced Reference : RR). Les MQI NR calculent une évaluation générale sur une image, en se référant à des connaissances a priori comme l'analyse spatiale ou des modèles statistiques de qualité des images. Par contre, les métriques FR comparent l'image donnée à une image de référence « idéale ». Cette comparaison peut être reliée à la sensibilité à l'erreur, la similarité structurelle, ou l'information mutuelle entre les deux images. Les métriques RR n'utilisent qu'une partie des attributs de l'image de référence. Vu la faible précision des méthodes statiques, certains travaux ont combinés plusieurs MQI pour détecter les attaques de spoofing : Utilisant une combinaison de 25 métriques, Galbally et al [12] ont proposé un système de classification binaire pour détecter les attaques de spoofing pour 3 modalité biométriques (Iris, empreinte, et reconnaissance faciale). Costa-Pazo et al [13] ont ensuite étudié et comparé deux méthodes de détection des attaques de présentation : la première emploie un sous ensemble de 18 métriques [12], tandis que la seconde est basée sur la texture en utilisant Gabor-Jets.

Les techniques multimodales ont été aussi étudiées comme étant des contre-mesures pour les attaques de spoofing. Elles combinent certaines modalités biométriques pour une authentification plus forte, comme la combinaison des reconnaissances faciale et vocale [14], la corrélation entre le visage, le mouvement des lèvres [4,15,16] ou la combinaison du visage, de la voix, et de l'iris [17,18]. Cependant, ces solutions sont souvent compliquées à mettre en œuvre (incommodité, coût, ou nuisance à l'expérience utilisateur).

Dans ce travail, nous nous concentrons sur les méthodes basées sur les MQI. Plusieurs bases de données publiques peuvent être utilisées pour l'évaluation. Nous avons choisi Replay Mobile [13], qui fournit des scénarii réalistes pour la détection des attaques de présentation sur les appareils mobiles.

3 Méthode proposée

Notre modèle passe par trois phases majeures : Etant donné une vidéo de l'utilisateur, nous commençons par extraire des frames spécifiques en utilisant des indicateurs de mouvement. Les MQI sont ensuite calculées sur ces frames. Les descripteurs obtenus constituent l'entrée du classifieur. Dans ce chapitre, nous expliquons les différents aspects considérés pour chacune de ces phases.

3.1 Extraction des frames

Le mouvement relatif du visage par rapport au fond peut être un indicateur révélateur pour distinguer entre un visage réel et une reproduction. Dans notre approche, nous considérons cet aspect en sélectionnant une frame donnée (I2) seulement si elle engendre un minimum de mouvements du visage par rapport au fond (Face-VS-Background Motion : FBM) par rapport à la dernière frame extraite (I1). La première frame de la vidéo est considérée comme étant la première référence de comparaison. Pour le travail présent, FBMmin=1.01. Le calcul de FBM est illustré par l'équation (1), où th est un seuil utilisé dans l'expression du mouvement calculé en (2). Le mouvement en une région d'intérêt donnée (Region of Interest: RoI) est mesuré par le nombre moyen des pixels de cette région où la différence d'intensité entre les deux images dépasse un seuil th=15. δ est la distribution de Dirac, D est la différence de pixels entre les deux frames au niveau de la RoI considérée, et SD est le nombre de pixels de cette dernière. Pour extraire la région du visage, nous avons utilisé CascadeObjectDetector fourni par la toolbox vision de Matlab, qui se base sur l'algorithme de Viola-Jones pour la détection de visage [19]. Cependant, nous avons étendu la région du visage dans les deux directions verticales pour inclure le cou et les cheveux comme des éventuels indicateurs de mouvements liés à la vivacité. Pour obtenir le fond, on soustrait la région du visage de l'image originale. Il est à noter que les valeurs des seuils th et FBM_{min} ont été choisies suite à des tests empiriques.

$$\mathrm{FBM} = \frac{\mathrm{Mouvement\ (Face, I1, I2, th)}}{\mathrm{Mouvement\ (Background, I1, I2, th)}} \tag{1}$$

$$\mathrm{Motion\ (RoI, I1, I2, th)} = \frac{\sum_{x,y} \delta(D(x,y) - th)}{SD} \tag{2}$$

3.2 Calcul des MQI

Pour les métriques FR, la disponibilité de l'image de référence est requise. Ceci représente un défi pour l'authentification biométrique, où on n'autorise que l'enregistrement d'un modèle de la donnée biométrique originale.

Dans [12], les auteurs calculent la qualité entre l'image de référence et sa version lissée, générée en utilisant un filtre gaussien. Dans notre modèle, nous calculons les métriques FR entre les frames de la vidéo, pour prendre en compte leurs différences reliées au mouvement. Les métriques NR sont ensuite calculées directement les frames sélectionnées. La combinaison entre toutes les mesures de qualité fournit le vecteur de descripteurs final. Il est à noter que lors du test de notre méthode, nous avons considéré l'extraction de toutes les frames ainsi que l'approche proposée basée sur la sélection de frames à partir de l'analyse du mouvement. Dans les deux cas, l'approche de calcul mentionnée précédemment est appliquée à l'ensemble de frames considéré.

3.3 Mesures de Qualité des Images (MQI)

L'étude de Galbally et al [12] propose un ensemble de 25 métriques pour détecter les attaques de spoofing. Dans [13], les auteurs améliorent les résultats de classification en utilisant 18 de ces métriques, notamment MSE, PSNR, AD, SC, NK, MD, LMSE, NAE, SNRv, RAMDv, MAS, MAMS, SME, GME, GPE, SSIM, VIF, et HLFI dont la dernière est une métrique NR. Dans notre étude, nous avons utilisé ces 18 métriques comme point de départ, et nous avons ajouté d'autres MQI NR citées dans le tableau 1. Parmi ces 6 métriques, nous avons sélectionné BIQI, GM-LOG-BIQA, et BRISQUE (sélection que nous discuterons dans la section 4). Dans ce qui suit, nous nous concentrons sur les trois métriques que nous avons gardées.

- BIQI [20,21] : La mesure finale de qualité est calculée suite à l'identification de la distorsion susceptible d'avoir affecté l'image, en se basant sur des modèles statistiques de coefficients d'ondelette. Dans notre implémentation, nous n'avons considéré que les valeurs de la première étape : chaque image est caractérisée par le modèle statistique (18 coefficients d'ondelette).
- GM-LOG-BIQA [22] : Comme indiqué par son nom, cette MQI exploite des descripteurs liés à l'amplitude du gradient et le Laplacien du Gaussien. En particulier, la prédiction de qualité est basée sur les statistiques communes de ces deux aspects, suite à une procédure de normalisation adaptive. Il est à noter que cette métrique engendre 40 mesures pour chaque image, contribuant au modèle de classification.
- BRISQUE [23,24] : De même, cette métrique exploite des descripteurs bas niveau, en calculant des coefficients centrés et normalisés de contraste. L'implémentation de BRISQUE inclut aussi un système d'évaluation de qualité appris à partir de ces descripteurs en utilisant un module de régression. Cependant, nous n'avons considéré que les 18 valeurs de descripteurs fournis par la première étape.

2	***Nom***	***Acronyme***	***Ref***
1	Blind/Referenceless Image Spatial QUality Evaluator	BRISQUE	[23, 24]
2	Gradient-Magnitude map and Laplacian-of-Gaussian based Blind Image Quality Assessment	GM-LOG-BIQA	[22]
3	Blind Image Quality Index	BIQI	[20, 21]
4	Naturalness Image Quality Estimator	NIQE	[25, 26]
5	Robust BRISQUE index	Robustbrisque	[27]
6	HDR Image GRADient based Evaluator - 1	HIGRADE-1	[28]

Tab. 1: *Liste des MQI utilisées dans nos tests en plus des 18 MQI utilisées en [13]*

4 Résultats expérimentaux

Cette section est dédiée à l'évaluation de la robustesse de notre méthode sur la base de données Replay Mobile, en utilisant le protocole d'évaluation décrit ci-dessous.

4.1 Protocole d'évaluation

Nous avons mené des simulations préliminaires sur les 6 MQI citées dans le tableau 1. La Fig.1 illustre une comparaison de leurs temps de calcul pour une image aussi bien que leurs performances individuelles si on considère 10 frames équidistantes pour chaque vidéo des ensembles d'apprentissage et de test. Le temps de calcul est mesuré sur un PC Windows 7 (64-bits) avec un processeur 2.50 GHz et 16 GB de mémoire RAM, en utilisant Matlab R2017a. Les trois métriques les plus rapides assurent une performance acceptable. HIGRADE-1 et Robustbrisque engendrent des taux d'erreurs plus faibles, mais elles sont relativement lentes. NIQE donne la plus mauvaise précision parmi les 6 métriques. Afin d'assurer un compromis entre la rapidité et la précision, nous avons sélectionné BRISQUE, GM-LOG-BIQA, et BIQI, en plus de l'ensemble initial de 18 MQI [13].

La base d'évaluation Replay Mobile comporte 1190 vidéos d'apparences réelles et d'attaques de présentation, enregistrées avec des appareils mobiles, sous différentes conditions de lumière. Les attaques incluent des enregistrements de vidéos et photos de durée minimale de 10 secondes. La base est divisée en sous-ensembles d'apprentissage et de test sans chevauchement que nous avons utilisés pour entrainer le modèle et prévoir les classes respectivement.

Pour l'évaluation des résultats de classification, nous calculons les faux négatifs (False Genuine Rate : FGR), les faux positifs (False Fake Rate : FFR) et la moyenne entre eux (Half Total Error Rate : HTER).

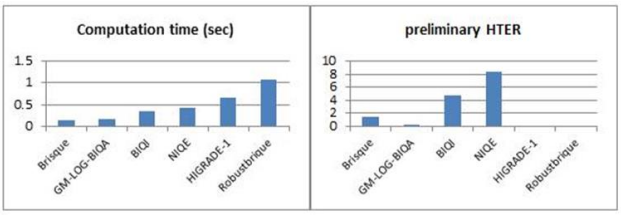


Fig. 1: *Comparaison des 6 IQMs en termes de temps de calcul (en sec) sur une image et les erreurs de classification LDA sur un sous ensemble de Replay Mobile, en utilisant 10 frames par vidéo*

4.2 Résultats sur Replay Mobile

Les erreurs de classification sur la base Replay Mobile sont détaillées dans le tableau 2, où nous considérons deux classifieurs : La pré-implémentation sous Matlab de l'Analyse Discriminante Linéaire (Linear Discriminant Analysis : LDA) avec un type de discrimination pseudo-linéaire, et la librairie LIBSVM [29] des Machine à vecteurs de support (Support Vector machine: SVM) avec un noyau à base radiale et un paramètre $\gamma = 1.5$.
Les notations « Tous » et « Mvt » dans le tableau 2 désignent l'utilisation de toutes les frames et la restriction à la sélection basée sur le mouvement respectivement. Les taux d'erreurs sont calculés par frame, sauf les deux dernières colonnes (HTER-v) où on attribue un résultat à chaque vidéo par une règle de majorité sur les prédictions faites sur l'ensemble de frames considéré, et ce pour assurer une comparaison légitime entre les protocoles « Tous » et « Mvt ». HTER-f désigne le HTER obtenu en considérant chaque frame comme étant un échantillon indépendant, tandis que HTER-v désigne le HTER résultant sur les vidéos. En outre, nous avons testé l'effet de la normalisation par la méthode Z-score sur la performance de classification.

MQI	Frames	Norm.	FGR		FFR		HTER-f		HTER-v	
			LDA	*SVM*	*LDA*	*SVM*	*LDA*	*SVM*	*LDA*	*SVM*
18 MQI initiales	Tous	sans	3.26	0.00	8.15	36.23	5.71	18.11	4.13	18.21
		Z-score	3.04	3.47	8.25	2.69	5.65	3.08	4.13	0.99
	Mvt	sans	1.25	0.00	9.71	31.85	5.48	15.92	**5.75**	**18.05**
		Z-score	0.52	1.63	11.29	4.73	5.91	3.18	**7.34**	**2.18**
21 MQI finales	Tous	sans	0.05	0.00	3.65	36.23	1.85	18.11	1.65	18.21
		Z-score	2.50	41.07	1.46	0.00	1.98	20.53	1.49	20.03
	Mvt	sans	0.07	0.00	1.66	31.85	0.86	15.92	**0.79**	**18.05**
		Z-score	1.32	0.00	4.02	28.59	2.67	14.29	**2.38**	**16.07**

Tab. 2: *Résultats de LDA et SVM en pourcentage sur Replay Mobile*

L'observation la plus intéressante que l'on peut tirer de ces résultats est la chute du taux d'erreurs pour la classification de LDA de plus que 4% à moins de 1% en utilisant notre sélection finale de 21 métriques par rapport à l'ensemble initial de 18 MQI. Par contre, le SVM est moins adapté aux MQI ajoutées, vu qu'il n'assure une bonne performance qu'avec la sélection initiale de 18 métriques, et ce avec la normalisation. Par ailleurs, celle-ci a un impact important sur la performance de SVM, contrairement au LDA.

Enfin, on remarque que le taux d'erreur le plus faible 0.79% est obtenu en appliquant notre approche sur la sélection de frames. Ceci confirme la pertinence de l'exploitation des mouvements au niveau de l'extraction des frames et du calcul des MQI. En outre, notre méthode permet de réduire le temps de calcul grâce au module de sélection de frames présenté dans la partie 3.1. En effet, en moyenne sur toutes les vidéos d'apprentissage et de test de Replay Mobile, 172 frames sont extraites sur un total approximatif de 300 frames par vidéo.

5 Conclusion

Une solution d'anti-spoofing non-intrusive et relativement rapide basée sur les MQI a été proposée et testée sur une base de données publique.
La première contribution est le choix des 21 MQI implémentées dans le système de classification binaire, parmi lesquelles 18 ont été déjà combinés dans des travaux antérieurs. Nous avons aussi exploité des indicateurs de mouvement d'une vidéo donnée afin d'extraire les frames les plus significatives, sur lesquelles les MQI sont calculées suivant une approche novatrice. Les résultats montrent que notre méthode assure une meilleure performance que les solutions de référence, tout en exploitant la rapidité relative du LDA pour la classification. De plus, le temps de calcul est réduit grâce à l'extraction sélective de frames.

S'il est vrai que ce travail a atteint son objectif, il n'en reste pas moins vrai qu'il présente certaines limitations à considérer. Les paramètres utilisés dans le module d'extraction des frames sont configurées empiriquement sur Replay Mobile, vu sa corrélation avec notre contexte d'application. Pour les travaux futurs, nous envisageons de tester notre méthode sur d'autres bases fréquemment utilisées dans la littérature pour les applications d'anti-spoofing. En outre, nous proposons de mener une étude plus approfondie des contributions individuelles des descripteurs implémentés, vu l'impact important du choix des MQI sur la précision de prédiction.

Références

[1] A.Anjos,J.Komulainen, S.Marcel, A.Hadid and M.Pi etikainen: Face Anti-spoofing: Visual Approach, Handbook of Biometric Anti-Spoofing, pages 65-82, Springer-Verlag, 2014.

[2] G.Pan, Z.Wu and L.Sun,: Liveness detection for face recognition, Recent Advances in Face Recognition, 109–124. InTech, 2008.

[3] H.K. Jee, S. U.Jung and , J. H.Yoo.: Liveness detection for embedded face recognition system, International Journal of Biological and Medical Sciences, vol. 1(4), pp. 235-238, 2006.

[4] K.Kollreider, H.Fronthaler and M.I.Faraj: Real-time face detection and motion analysis with application in "liveness" assessment, IEEE Transactions on Information Forensics and Security, 2(3-2):548–558, 2007.

[5] K.Kollreider, H.Fronthaler and J.Bigun: Non-intrusive liveness detection by face images. Image and Vision Computing 27: 233–244, 2009.

[6] A.Anjos and S.Marcel: Counter-measures to photo attacks in face recognition: a public database and a baseline. Proc. Proceedings of IAPR IEEE International Joint Conference on Biometrics (IJCB), 2011.

[7] W. Bao, H.Li and N.Li: A liveness detection method for face recognition based on optical flow field. Proc. 2009 International Conference on Image Analysis and Signal Processing, IEEE, 233–236, 2009.

[8] A. Lagorio, M.Tistarelli and M.Cadoni: Liveness Detection based on 3D Face Shape Analysis, Biometrics and Forensics (IWBF), 2013 International Workshop, 1-4, 2013.

[9] T.Wang, J.Yand and Z.Lei: Face Liveness Detection Using 3D Structure Recovered from a Single Camera, International Conference on Biometrics, Madrid, Spain, 2013.

[10] E.S. Ng and A.Y.S.Chia, Face verification using temporal affective cues. Proc. International Conference on Pattern Recognition (ICPR), 1249–1252, 2012.

[11] Z.Boulkenafet, J.Komulainen and A.Hadid: Face anti-spoofing based on color texture analysis, IEEE International Conference on Image Processing (ICIP), 2636-2640, 2015.

[12] J.Galbally, S.Marcel and J.Fierrez,: "Image quality assessment for fake biometric detection: Application to iris, fingerprint and face recognition," IEEE Trans. on Image Processing, vol. 23, no. 2, pp. 710–724, 2014.

[13] A.Costa-Pazo, S.Bhattacharjee and E.V. Fernandez: "The replay-mobile face presentation-attack database," International Conference on Biometrics Special Interests Group (BioSIG), 2016.

[14] H. T. Cheng, Y.H. Chao and S.L.Yeh: An efficient approach to multimodal person identity verification by fusing face and voice information. IEEE International Conference on Multimedia and Expo, pages 542–545, 2005.

[15] J. Komulainen, I. Anina and J. Holappa: "On the robustness of audiovisual liveness detection to visual speech animation," IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), Niagara Falls, NY, USA, 2016.

[16] A.Melnikov, R. Akhunzyanov, O. Kudashev and E. Luckyanets: "Audiovisual Liveness Detection," pages 643–652. International Conference on Image Analysis and Processing (ICIAP), Springer, 2015.

[17] P.H. Lee, L.J.Chu and Y.P.Hung: Cascading Multimodal Verification using Face, Voice and Iris Information. IEEE International Conference on Multimedia and Expo., Beijing, China, pp 847-850, 2007.

[18] T. Barbu, A.Ciobanu and M.Luca, :Multimodal biometric authentication based on voice, face and iris. E-Health and Bioengineering Conference (EHB), Iasi, Romania, 2015.

[19] P.Viola and M.J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume: 1, pp.511–518, 2001.

[20] A.K.Moorthy and A.C.Bovik,: "A Modular Framework for Constructing Blind Universal Quality Indices", submitted to IEEE Signal Processing Letters, 2009.

[21] A.K.Moorthy and A.C.Bovik,:"BIQI Software Release", URL: http://live.ece.utexas.edu/research/quality/biqi.zip, 2009.

[22] W. Xue, X.Mou and L.Zhang: "Blind Image Quality Prediction Using Joint Statistics of Gradient Magnitude and Laplacian Features", Trans. on Image Processing, IEEE selben Jahr. Format-Verlag, 2014.

[23] A.Mittal, A.K.Moorthy and A.C.Bovik: "BRISQUE Software Release", http://live.ece.utexas.edu/research/quality/BRISQUE_release.zip, 2011.

[24] A.Mittal, A.K.Moorthy and A.C.Bovik.:" No Reference Image Quality Assessment in the Spatial Domain"

[25] A.Mittal, R.Soundararajan and A.C.Bovik: "NIQE Software Release", http://live.ece.utexas.edu/research/quality/niqe.zip, 2012

[26] A.Mittal, R.Soundararajan and A.C.Bovik: "Making a Completely Blind Image Quality Analyzer", submitted to IEEE Signal Processing Letters, 2012

[27] A. Mittal, A. K.Moorthy and A. C Bovik:"Making image quality assessment robust", Forty-Sixth Annual Asilomar Conf. on Signals, Systems, and Computers, Monterey, California, http://live.ece.utexas.edu/research/quality/robustbrisque_release.zip, 2012.

[28] D.Kundu, D. Ghadiyaram, A.C.Bovik and B.L.Evans: "No-Reference Quality Assessment of High Dynamic Range Images," IEEE Transactions on Image Processing, 2016. http://users.ece.utexas.edu/~bevans/papers/2017/crowdsourced/index.html

[29] C.C.Chang and C.J.Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Analyse d'image, de la vidéo, et des données 3D

Session 2

Reconnaissance d'expressions corporelles à l'aide d'un mouvement neutre synthétisé

A. Crenn[1] H. Konik[2] A. Meyer[1] S. Bouakaz[1]

[1] Université de Lyon 1, LIRIS, UMR5205, F-69622, France

[2] Université de Saint-Etienne, LHC, UMR5516, F–42000, France

{arthur.crenn, alexandre.meyer, saida.bouakaz}@liris.univ-lyon1.fr, hubert.konik@univ-st-etienne.fr

7 octobre 2017

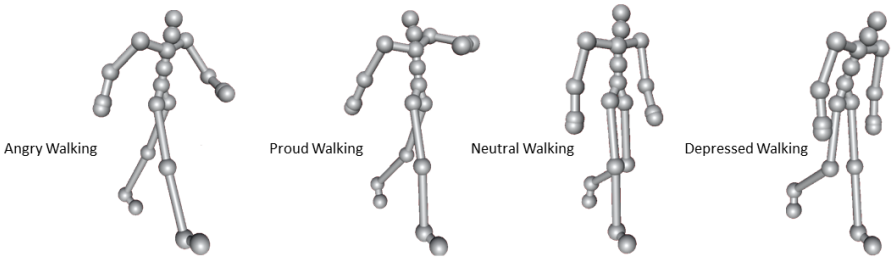


FIGURE 1 – *Notre méthode reconnaît l'expression corporelle d'un squelette 3D humain.*

Résumé

Nous présentons une approche pour la reconnaissance d'expressions corporelles qui accompagnent le mouvement d'un personnage réel. L'idée développée dans cet article s'inspire de travaux en synthèses d'animations. Notre système se base sur l'analyse de la différence obtenue dans le domaine fréquentiel entre un mouvement expressif et un mouvement neutre. Dans ce papier, nous introduisons la notion de mouvement neutre et proposons un algorithme qui permet de l'obtenir par la minimisation d'une fonction de coût global. Le coût global tient compte de la distance parcourue par les différentes articulations et l'accélération de chaque articulation au cours du mouvement. Nous avons évalué notre méthode sur différentes bases de données contenant des mouvements et expressions corporelles hétérogènes. Les résultats obtenus sont encourageants et confirment la validité de notre démarche.

Mots clefs

Vision par ordinateur, Expression Corporelle, Reconnaissance Automatique, Squelette 3D, Classification.

1 Introduction

Les humains expriment l'information émotionnelle à travers différents canaux (visage, corps, voix, etc.). Plusieurs études montrent que les expressions corporelles expriment autant que les expressions faciales l'intensité d'une émotion[1]. Si la reconnaissance d'expressions faciales a largement été étudiée [2, 3, 4], le domaine de la reconnaissance d'expressions corporelles est encore assez jeune. Les expressions corporelles tout comme les expressions faciales sont extrêmement difficile à formaliser. Les premières méthodes pour l'analyse d'un mouvement corporel se sont principalement focalisées sur un type de mouvements ou d'expressions [5, 6]. La généralisation des dispositifs permettant l'obtention du squelette 3D ont entraîné une demande croissante de leur usage. Cette utilisation a alors motivé le besoin de reconnaître automatiquement aussi bien les mouvements que les expressions corporelles.

Nous nous intéressons plus particulièrement aux expressions corporelles qui portent des émotions : fatigue, colère, joie, etc. Ces expressions peuvent s'exprimer en même temps que des mouvements ordinaires tels que la marche, le saut, un coup de pied, etc. Cet article présente une méthode de détection et de classification d'expressions à travers une séquence de poses d'un squelette 3D en s'appuyant sur l'analyse de la différence obtenu entre le mouvement original et le mouvement neutre synthétisé par notre méthode. La synthèse d'une animation neutre permet d'avoir une méthode de reconnaissance d'expressions invariante au mouvement corporel exécuté.

Les chercheurs en psychologie ont été les premiers à s'intéresser aux expressions corporelles selon la posture et le mouvement du corps. Ces derniers soutiennent l'idée que les expressions corporelles "parlent" plus que les expressions faciales [7]. Les différentes tentatives de formalisation ont dégagé deux niveaux de descripteurs : les descripteurs de haut niveau et ceux de bas niveau. Bien que diffé-

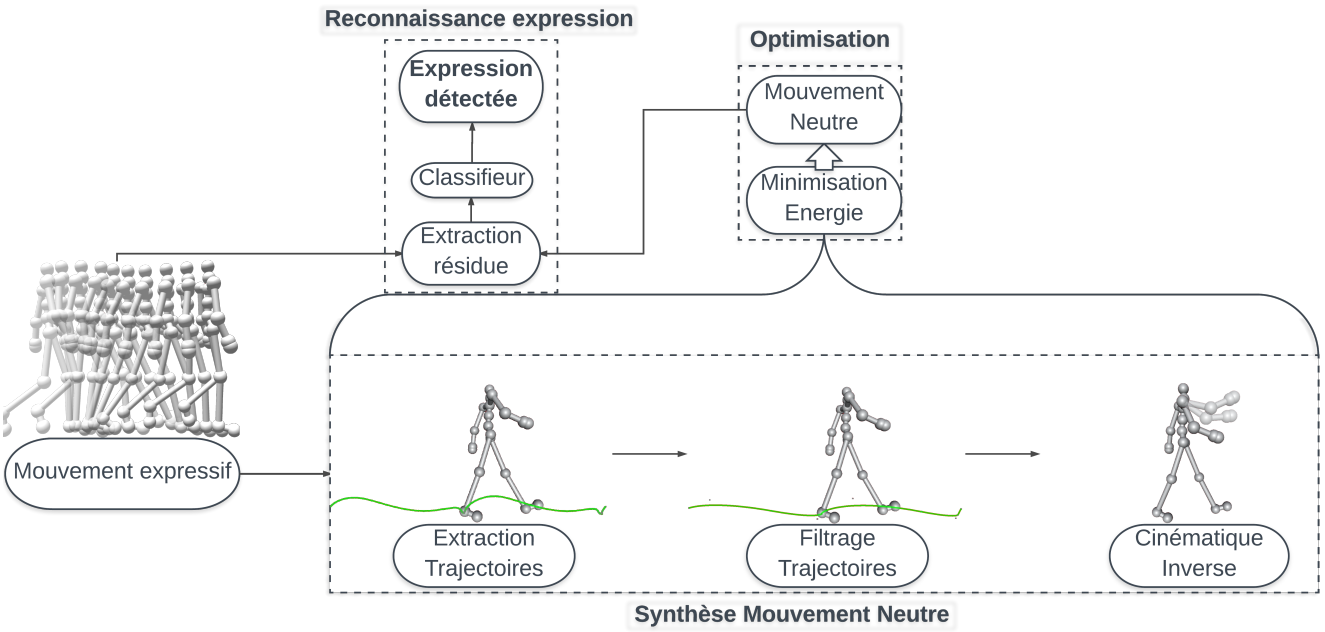


Figure 2 – *Schéma général de notre méthode.*

rents termes ont été adoptés pour ces descripteurs, on peut retrouver les mêmes notions suivant l'expression décrite. De façon formelle, ces descripteurs se réfèrent aux articulations : angle de rotation, position 3D, vitesse, accélération, distance entre articulations, etc. Les descripteurs de haut niveau dépendent du contexte de l'action et sont liés au mouvement. Le formalisme le plus connu est certainement celui proposé par Laban et Ullmann [8] pour l'analyse de mouvements de danse. Ce modèle caractérise un mouvement en cinq notions : le corps, l'espace, l'effort, la forme et la relation. Ceci permet de décrire et d'interpréter un mouvement avec un ensemble réduit de paramètres.

Dans la littérature, peu de travaux proposent une reconnaissance d'expressions corporelles indépendamment du mouvement exécuté. Diverses approches se sont focalisées sur l'étude d'une action spécifique comme la marche [5, 9, 10], l'action de frapper à une porte [11, 12], la danse [13] ou encore un scénario de discussion [14].

Wang et al. [15] propose un système temps-réel pour la reconnaissance d'expressions corporelles qui utilise une combinaison de descripteurs 3D bas niveaux et de descripteurs géométriques. Ils obtiennent un taux de classification de 78% avec un classifieur de type forêt d'arbres décisionnels sur la base UCLIC proposée par Kleinsmith et al. [16]. Truong et al. [17] ont proposé un nouvel ensemble de descripteurs 3D basé sur la généralisation du modèle Laban. Ces descripteurs ont été appliqués pour la reconnaissance de gestes sur la base Microsoft Research Cambridge-12 et ont obtenu un taux de reconnaissance de 97%. Cette même méthode a été étendue pour la reconnaissance d'expressions corporelles. Ils ont aussi testé leur approche pour la reconnaissance d'expressions corporelles sur leur base de données contenants 882 gestes et ont obtenu un F-score de 56.9%. Finalement, Crenn et al. [18] ont proposé un ensemble de descripteurs génériques empruntés au domaine de la psychologie. L'avantage de ces descripteurs est d'être générique et de s'adapter à différents types de mouvements. Cette méthode a été évaluée sur différentes bases de données et produit de bons résultats.

Dans le domaine de la synthèse d'animations, la génération d'une animation est obtenue par la modification des paramètres de chaque articulation comme la vitesse, la position, l'amplitude spatiale, etc [19, 20, 21]. Le mouvement d'une animation correspond à un signal sur lequel on peut appliquer la transformée de Fourier. Ceci permet d'éditer un mouvement ou de transférer un style, concept incluant l'expression corporelle [22, 23]. Plus récemment, Yumer et Mitra [24] ont proposé une méthode pour le transfert de style basée sur le résidu calculé comme une différence, dans le domaine fréquentiel, entre une animation stylisée et une animation neutre. Sur le principe que le résidu peut quantifier l'information relative au style, nous utilisons la notion de mouvement neutre pour extraire l'information décrivant l'expression corporelle. Néanmoins, à la différence de Yumer et Mitra, nous ne disposons pas de mouvement neutre. La première étape consiste donc à obtenir un mouvement neutre à partir d'un mouvement expressif. Le résidu calculé est utilisé par un classifieur afin de reconnaître l'expression corporelle. Nous avons évalué notre méthode sur différentes bases de données de mouvements et d'expressions variés. Dans la suite de notre article, nous détaillons la méthode proposée dans la Section 2. Dans la Section 3, nous analysons les résultats obtenus sur les différentes bases de données en les comparant avec l'état de l'art. Dans la Section 4, nous donnons enfin les perspectives à ces travaux.

2 Méthode proposée

Le principal objectif de ce travail est la reconnaissance d'expressions corporelles indépendamment du mouvement exécuté. Dans nos travaux, nous considérons que le résidu obtenu entre une animation stylisée et une animation neutre contient l'information relative au style. Le résidu est calculé comme une différence, dans le domaine fréquentiel, sur chaque degré de liberté entre une animation stylisée et une animation neutre. Cette idée a déjà été exploitée en synthèse d'animations par Yumer et Mitra [24]. A la différence du domaine d'animations, nous ne disposons pas du mouvement neutre. Le premier verrou consiste donc à obtenir une animation neutre à partir d'un mouvement expressif. Pour cela, nous proposons une fonction de coût qui caractérise un mouvement neutre. Cette fonction est basée sur des caractéristiques cinématiques (distance, vitesse, accélération) calculées pour chaque articulation durant un mouvement. La figure 2 donne le schéma général de notre méthode.

2.1 Synthèse de mouvement neutre

La notion de mouvement neutre versus expressif est toujours liée à un contexte. Souvent compris comme étant une suite d'actions sans marque émotionnelle, un mouvement neutre peut éventuellement être confondu avec un mouvement robotique. Notre approche se base sur un filtrage des trajectoires de chaque articulation afin de réduire les oscillations dans le mouvement expressif. La seconde étape fait appel à la cinématique inverse pour produire sans expression : mouvement robotique. Finalement, on introduit une fonction de coût qui va réaliser un compromis entre le filtrage des articulations et la cinématique inverse.

Filtrage de la trajectoire d'une articulation. Un mouvement est représenté par des échantillons temporels Correspondant au angles de chaque articulation de notre squelette. Dans la première nous représentons la trajectoire 3D de chaque articulation par une B-spline. Afin d'atténuer les oscillations dans la trajectoire initiale, nous procédons à un "filtrage" de cette trajectoire. Pour cela, nous réduisons chaque B-spline en retirant un point de contrôle à chaque courbe de Bézier composant cette B-spline. (Figure 3. La trajectoire (B-spline) est ensuite reconstruite avec les points de contrôle restants.
L'étape de filtrage permet de réduire les oscillations contenus dans le mouvement stylisé de façon incrémentale. En analysant différents mouvements expressifs contenus dans différentes bases de données, nous avons catégorisé deux styles de mouvements expressifs : les styles énergiques (joie, fierté, colère, etc.) et les styles modérés (déprimé, âgé, etc.). Par ailleurs, nous avons également constaté qu'un mouvement neutre est plus "plat" qu'un mouvement énergique, c'est-à-dire qu'il présente une trajectoire avec des oscillations modérées. Notre étape de filtrage des trajectoires permet de générer des mouvements neutres dans ce cas-là. Le filtrage de la trajectoire aura peu d'effet sur les mouvements modérés. Pour cela, le recours à la cinématique inverse combinée à la fonction de coût, permet de synthétiser des animations neutres pour les mouvements modérées.

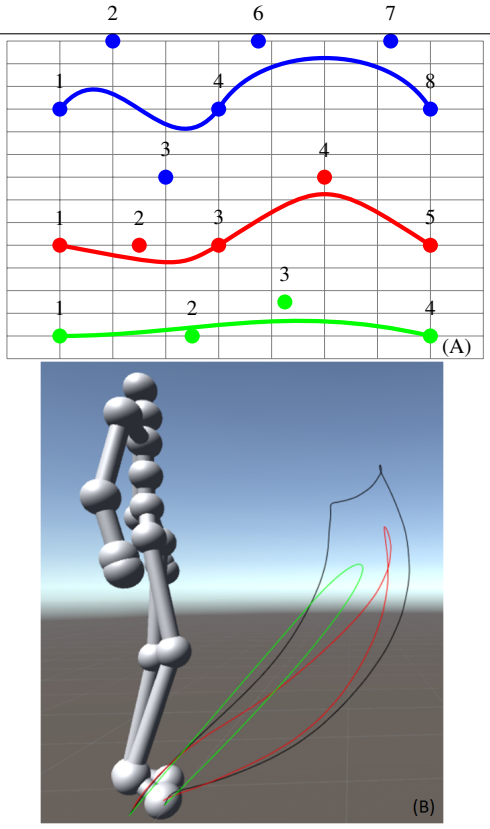


Figure 3 – *Figure A, réduction de la trajectoire par filtrage de B-spline à partir du mouvement initial (en bleu) : en rouge (resp.) après une (resp.deux) itération. Figure B, comparaison de différentes trajectoires lors d'un mouvement de coup de pied. Le mouvement original est en noir (colère). La vérité-terrain est représentée en rouge. La trajectoire "neutre" générée par notre méthode après optimisation est en vert.*

Cinématique Inverse. Le filtrage des trajectoires de chaque articulation n'est pas suffisant pour produire une animation neutre. D'une part, le mouvement généré par l'étape de lissage ne respecte pas différentes contraintes (longueur des articulations, pieds qui glissent). D'autre part, comme il a été signalé dans la section précédente, un mouvement robotique est une bonne référence pour un mouvement neutre et la cinématique inverse permet de produire ce type de mouvement. C'est pourquoi nous appliquons une étape de cinématique inverse (IK) pour déterminer un ensemble de poses, pour chaque articulation, permettant de déplacer les effecteurs finaux à une position optimale en terme de distance et de temps de calcul par rapport à une position désirée. Dans notre cas, nous utilisons l'algorithme Fabrik proposé par Aristidou et Lasenby [25]. Il possède l'avantage de produire des résultats visuels réalistes en peu d'itérations et d'accepter l'ajout de

contraintes.
L'algorithme 1 présente la génération d'un mouvement à partir de n'importe quel mouvement expressif. Cette méthode peut produire des mouvements peu naturels. Cependant, comme le montre le taux de reconnaissance obtenu dans la section 3, elle reste néanmoins efficace. Notre algorithme est générique. Il fait intervenir différents paramètres qui le rendent adaptables à tous types de mouvements. Nous présentons dans la section suivante la fonction de coût utilisée engendrant les paramètres optimaux pour un mouvement donné.

Algorithm 1: Algorithme pour la synthèse de mouvement neutre

Input: $M_i i$: Mouvement original
Data: joints : tableau des articulations
trajectories : tableau des trajectoires de chaque articulation
trajectoriesSmooth : tableau des trajectoires filtrées pour chaque articulation
Result: M_n : Mouvement Neutre
Parameters: samplingValue : paramètre temporel afin de déterminer les postures clés pour la phase de cinématique inverse temporal
weightTargets : tableau de poids permettant d'équilibrer la posture finale obtenu par la cinématique inverse (1 = sur la cible et 0 = squelette au repos).
weightHints : tableau de poids permettant d'équilibrer les indices de la posture finale obtenu par la cinématique inverse (1 = sur la cible et 0 = squelette au repos).

```
7 foreach joint, j_i, in joints do
    trajectories_i ← computeTrajectory(j_i);
    trajectoriesSmooth_i ← smoothTrajectory(trajectories_i);
end
foreach EndEffector, end_i, in joints do
    indiceHint ← end_i.getParent();
    for i =0; i < endTime; i += samplingValue do
        target_{end_i} ← trajectoriesSmooth_i ;
        hint_{end_i} ← trajectoriesSmooth_{indiceHint} ;
        IK_Step( target_{end_i}, hint_{end_i}, weightTargets, weightHints);
    end
end
```

Fonction de coût pour la synthèse d'un mouvement neutre. Nous proposons une fonction de coût qui caractérise un mouvement neutre. Cette dernière est relativement simple et produit des animations neutres pertinentes pour la classification. Cette fonction est basée sur la distance parcourue par chaque articulation et son accélération durant un mouvement. Nous assumons qu'un mouvement neutre correspond à une dépense d'énergie minimale lors d'un mouvement. En effet, une personne cherchant à exécuter un mouvement cherchera à économiser son énergie. En posant $D_s(j)$ (respectivement $D_o(j)$) la distance parcourue par l'articulation j lors du mouvement neutre synthétisé (respectivement le mouvement original). De même, $\mathbf{A_s}(j)$ (respectivement $\mathbf{A_o}(j)$) l'accélération de l'articulation j durant le mouvement neutre synthétisé (respectivement le mouvement original). La fonction de coût est définie comme la somme des différences entre la distance et l'accélération originale pour chaque articulation et celle synthétisée. La minimisation de cette fonction de coût fournit un mouvement neutre grossier utilisé dans la section 12 pour calculer le résidu obtenu entre le mouvement original et le mouvement synthétisé.

$$Cout = \sum_{j\in\theta} |(1-\lambda)(D_s(j) - D_o(j)) + \lambda(\mathbf{A_s}(j) - \mathbf{A_o}(j))|^2 \quad (1)$$

avec j qui représente une articulation, θ est l'ensemble des articulations du squelette et $\lambda \in [0, 1]$ est un paramètre de poids. L'influence du facteur λ pour le taux de classification sur la base SIGGRAPH est présentée dans la figure 4. Les distances $D_s(j)$ et $D_o(j)$ parcourues par une articulation j lors d'un mouvement sont données par la longueur de nos B-spline. Les accélérations $\mathbf{A_s}(j)$ et $\mathbf{A_o}(j)$ fournissent de l'information sur l'énergie dépensée par une articulation j lors d'un mouvement. Afin de trouver ces accélérations, nous calculons la dérivée seconde de chaque B-spline. La fonction de coût est minimisée itérativement en utilisant l'algorithme d'optimisation par essaim de particules (PSO). PSO possède l'avantage de ne n'imposer que peu de contraintes voire aucune sur la fonction à optimiser.

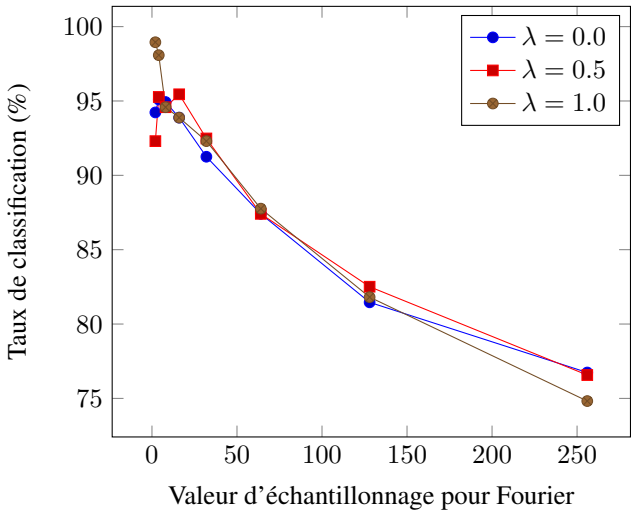


Figure 4 – *Comparaison de taux de classification avec différentes valeurs de λ et de ré-échantillonnage pour la base SIGGRAPH[21].*

Résidu entre le mouvement neutre et expressive. À ce stade, nous avons le mouvement original expressif et le mouvement neutre obtenu en minimisant la fonction de coût. Nous calculons la représentation fréquentielle via la transformée de Fourier de ces deux mouvements. Dans la

représentation fréquentielle, la magnitude contient l'information de mouvement ainsi que de l'expression pour une animation donnée. En calculant la différence de magnitude pour chaque degré de liberté de chaque articulation, nous obtenons le résidu entre l'animation neutre et l'animation expressive. Le résidu forme le vecteur de descripteurs utilisé en entrée pour la classification afin de détecter l'expression corporelle. Dans notre approche, nous avons un paramètre à définir concernant le nombre d'échantillon de notre signal d'entrée. La figure 4 présente la variation du taux de classification en fonction de différentes valeurs de ré-échantillonnage. Le taux de classification diminue lorsqu'on augmente le nombre d'échantillons. Cette observation est liée au fait que lorsqu'on augmente la taille de notre signal d'entrée, les valeurs du résidu obtenues sont très proches de zéro dans les basses fréquences créant du bruit pour notre classifieur.

3 Résultats et conclusion

Nous avons évalué notre méthode sur quatre bases de données présentées dans le tableau 1. Trois de ces bases ont été réalisées par motion capture d'acteurs jouant diverses actions comme marcher, frapper, porter, lancer, etc. La dernière base, nommée SIGGRAPH, est utilisée en synthèse d'animation [21].

Base de donnée	Nb mouvements	Nb expressions
UCLIC [16]	183	4
Biological [26]	1356	4
MPI [14]	1443	11
SIGGRAPH [21]	572	4 et 4 styles

BDD	**Résultats état de l'art**	**Nos résultats**
UCLIC	78% [15]	**83%**
Biological	50% à 80% [11]	**57%**
MPI	–	**50%**
SIGGRAPH	93% [18]	**98%**

Tableau 1 – *Présentation des bases de données utilisées ainsi que la comparaison de notre méthode par rapport aux méthodes de l'état de l'art.*

La figure 5 montre l'influence de la taille de l'ensemble d'apprentissage sur la performance des trois classifieurs utilisés dans notre méthode. Nous avons comparé la performance de notre méthode en utilisant différents algorithmes de classifications : SVM avec un noyau χ^2, Random Forest avec 100 arbres et 2-Nearest neighbor basé sur la distance Euclidienne. La figure 5 a été produite sur la base SIGGRAPH. Pour chaque classifieur, nous avons calculé le taux de classification en utilisant différentes valeurs d'échantillons par la méthode de validation croisée *k*-fold. La performance de notre système a été évaluée en utilisant l'algorithme Random Forest avec 10 échantillons et 100 arbres qui fournit le meilleur taux de classification. Le fait que de bons résultats soient obtenus avec cet algorithme montre que notre espace de descripteurs est bien discriminant.

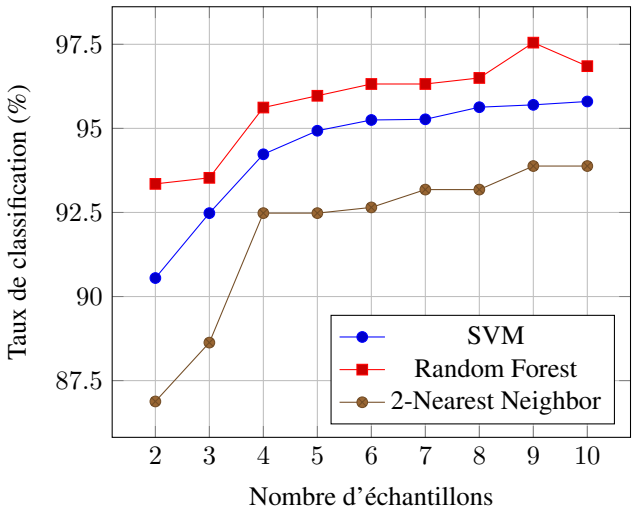


Figure 5 – *Évolution du taux de classification pour la base SIGGRAPH en faisant varier le nombre d'échantillons pour la validation croisée par k-fold*

Le tableau 1 compare notre méthode avec les méthodes de l'état de l'art sur les mêmes bases de données. Il démontre que notre approche dépasse les autres approches de l'état de l'art en terme de taux de reconnaissance d'expressions corporelles. De plus, notre méthode est générique alors que l'état de l'art comporte des méthodes spécifiques à certains gestes. Notre taux de reconnaissance d'expressions corporelles est meilleur que l'état de l'art pour les bases SIGGRAPH et UCLIC. Sur la base Biological Motion, la méthode [11] suppose que tous les gestes sont du même type pour calculer un mouvement moyen, ici tous les mouvements répondent à l'action de frapper à une porte. Cette hypothèse empêche de généraliser leur approche à des gestes dont le type n'est pas connu à l'avance, contrairement à notre approche. Nous pensons donc que pour comparer notre approche à la leur il est cohérent d'utiliser leur taux de reconnaissance générique de 50% alors que la notre est de 57%. Qui plus est, dans la littérature, aucune méthode de reconnaissance d'expressions corporelles n'a utilisé cette base MPI. Notre méthode obtient un taux de reconnaissance de 50%, ce qui est fort louable avec cette base extrêmement difficile car elle contient de nombreuses expressions avec un nombre d'exemples très variable selon l'expression. En utilisant un filtre de ré-échantillonnage classique afin de gérer des bases de données déséquilibrées, nous obtenons même un taux de classification de 67%.

4 Conclusion et perspectives

Nous avons présenté une approche pour la reconnaissance automatique d'expressions corporelles à partir d'un squelette 3D obtenu par motion capture. Nous soutenons l'idée que l'expression corporelle peut être détectée de manière robuste en analysant la différence obtenue entre un mouvement neutre et un mouvement expressif. Nous avons pro-

posé un algorithme qui permet de synthétiser un mouvement neutre à partir d'un mouvement expressif. À partir du mouvement neutre synthétisé, notre méthode classifie l'expression du mouvement original en calculant la différence dans le domaine fréquentiel entre l'animation neutre synthétisée et le mouvement expressif. Les résultats obtenus sur différents bases de données sont très prometteurs. Ainsi, notre approche ouvre de nouvelles possibilités pour de nombreuses applications : interaction homme-machine, jeux vidéo, réalité virtuelle, psychologie, etc. L'une des suites de nos travaux sera de travailler sur le réalisme de l'animation neutre produite notamment en améliorant la phase de la synthèse du mouvement neutre. Actuellement nous traitons les degrés de liberté de chaque articulation de manière indépendante. Aussi, nous nous orientons vers une approche s'appuyant sur la transformée de Fourier quaternionique afin d'unifier le traitement des différents degrés de liberté de chaque articulation.

Références

[1] Andrea Kleinsmith et Nadia Bianchi-Berthouze. Affective body expression perception and recognition : A survey. *Affective Computing, IEEE Transactions on*, 4(1) :15–33, 2013.

[2] Vinay Bettadapura. Face expression recognition and analysis : the state of the art. *arXiv preprint arXiv :1203.6722*, 2012.

[3] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik, et Saida Bouakaz. Framework for reliable, real-time facial expression recognition for low resolution images. *Pattern Recognition Letters*, 34(10) :1159 – 1168, 2013.

[4] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, et Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. Dans *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[5] Michelle Karg, Kolja Kühnlenz, et Martin Buss. Recognition of Affect Based on Gait Patterns. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4) :1050–1061, Août 2010.

[6] Andrea Kleinsmith, Tsuyoshi Fushimi, et Nadia Bianchi-Berthouze. An incremental and interactive affective posture recognition system. Dans *International Workshop on Adapting the Interaction Style to Affective Factors*, pages 378–387, 2005.

[7] Albert Mehrabian et John T Friar. Encoding of attitude by a seated communicator via posture and position cues. *Journal of Consulting and Clinical Psychology*, 33(3) :330, 1969.

[8] R. von Laban et L. Ullmann. *The mastery of movement*. Numéro vol. 1971,ptie. 1 dans The Mastery of Movement. Macdonald & Evans, 1971.

[9] Avi Barliya, Lars Omlor, Martin A. Giese, Alain Berthoz, et Tamar Flash. Expression of emotion in the kinematics of locomotion. *Experimental brain research*, 225(2) :159–176, 2013.

[10] C.L. Roether, Lars Omlor, Andrea Christensen, et Martin A. Giese. Critical features for the perception of emotion from gait. *Journal of Vision*, 9(6) :1–32, 06 2009.

[11] Daniel Bernhardt et Peter Robinson. Detecting affect from non-stylised body motions. Dans *Affective Computing and Intelligent Interaction*, pages 59–70. Springer, 2007.

[12] M. Melissa Gross, Elizabeth A. Crane, et Barbara L. Fredrickson. Methodology for Assessing Bodily Expression of Emotion. *Journal of Nonverbal Behavior*, 34(4) :223–248, Décembre 2010.

[13] Simon Senecal, Louis Cuel, Andreas Aristidou, et Nadia Magnenat-Thalmann. Continuous body emotion recognition system during theater performances : Continuous body emotion recognition. *Computer Animation and Virtual Worlds*, 27(3-4) :311–320, Mai 2016.

[14] Ekaterina Volkova, Stephan de la Rosa, Heinrich H. Bülthoff, et Betty Mohler. The MPI Emotional Body Expressions Database for Narrative Scenarios. *PLoS ONE*, 9(12) :e113647, Décembre 2014.

[15] Weiyi Wang, Valentin Enescu, et Hichem Sahli. Adaptive Real-Time Emotion Recognition from Body Movements. *ACM Transactions on Interactive Intelligent Systems*, 5(4) :1–21, Décembre 2015.

[16] Andrea Kleinsmith, P. Ravindra De Silva, et Nadia Bianchi-Berthouze. Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers*, 18(6) :1371–1389, 2006.

[17] Arthur Truong, Hugo Boujut, et Titus Zaharia. Laban descriptors for gesture recognition and emotional analysis. *The Visual Computer*, 32(1) :83–98, Janvier 2016.

[18] Arthur Crenn, Rizwan Ahmed Khan, Alexandre Meyer, et Saida Bouakaz. Body expression recognition from animated 3d skeleton. pages 1–7. IEEE, Décembre 2016.

[19] Kenji Amaya, Armin Bruderlin, et Tom Calvert. Emotion from motion. Dans *Graphics interface*, volume 96, pages 222–229. Toronto, Canada, 1996.

[20] Eugene Hsu, Kari Pulli, et Jovan Popović. Style translation for human motion. *ACM Transactions on Graphics (TOG)*, 24(3) :1082–1089, 2005.

[21] Shihong Xia, Congyi Wang, Jinxiang Chai, et Jessica Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)*, 34(4) :119, 2015.

[22] Armin Bruderlin et Lance Williams. Motion signal processing. Dans *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 97–104. ACM, 1995.

[23] Munetoshi Unuma, Ken Anjyo, et Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. Dans *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 91–96. ACM, 1995.

[24] M. Ersin Yumer et Niloy J. Mitra. Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics*, 35(4) :1–8, Juillet 2016.

[25] Andreas Aristidou et Joan Lasenby. FABRIK : A fast, iterative solver for the Inverse Kinematics problem. *Graphical Models*, 73(5) :243–260, Septembre 2011.

[26] Yingliang Ma, Helena M. Paterson, et Frank E. Pollick. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior research methods*, 38(1) :134–141, 2006.

Performance Study of View Synthesis with Small Baseline for Free Navigation

P. Nikitin[1,2] J. Jung[1] M. Cagnazzo[2] B. Pesquet[2]

[1] Orange Labs ,[2] LTCI, Télécom ParisTech, Université Paris-Saclay

{pavel.nikitin, joel.jung}@orange.com, {marco.cognazzo, beatrice.pesquet}@telecom-paristech.fr

Abstract

In a typical Free Navigation service, view synthesis is expected to provide virtual views between the real captured views, in order to improve the smoothness of the navigation. Practical constraints prevent from capturing views with a very small baseline, so view synthesis is required. One way of synthesizing views is to use texture and depth information. It is of interest to understand how much current view synthesis technology is able to provide acceptable quality for synthesized views, in the framework of Free Navigation.

A new super multi-view content has been recently provided by the University of Brussels. This high-density content has the characteristic to have a very small baseline of 1mm and is particularly adapted for this study.

In this study, some experiments of view synthesis with small baseline were performed. Experimental results are reported to understand how far view synthesis can be used, both from an objective and from a subjective point of view. It was shown that according to subjective point of view, more views can be synthesized while maintaining acceptable quality.

Key words

Free Navigation, View synthesis, Super multiview

1 Introduction

In a future video services user should have a possibility to freely navigate within the scene. In order to achieve this a huge number of views should be captured, which is not possible, because of the physical constraints. In addition to physical limitations, it is currently not possible to handle thousands of cameras and distribute such content. As a consequence, view synthesis is required.

In the most ambitious scenario, a user will be able to move freely and stop on any view, no matter if it is a synthesized view or a captured view. In a shorter term scenario, we might consider that a user is able to move freely, but to stop only on captured views. In this case, view synthesis can be seen as a means to make the navigation smoother.

One way of synthesizing views is to use texture and depth information. For instance, the MPEG reference software VSRS 4.1 [1] is well known for its ability to synthesize views out of a pair of views and corresponding depths. Similarly, during the 3D-HEVC standardisation process, VSRS1d-fast [2] has been extensively used.

Recent FTV Call for Evidence [3] has been too optimistic, trying to synthesize too many views in between two captured views (seen differently, considering a distance between two cameras too huge), so that even the anchor had unacceptable visual quality. Today, it is of interest to understand how much current view synthesis technology is able to provide acceptable quality for synthesized views.

A new super multi-view content has been recently provided by the University of Brussels [4]. This content has the characteristic to have a very small baseline of 1mm. Although the content is static (single frame), it is a first step that allows to test very different configurations of view synthesis, while always having a reference that has been captured, to compare with.

In this study, we perform some tests of view synthesis with small baseline. Experimental results are reported to understand how far view synthesis can be used. It is very important to understand that any result that claims a given distance between two cameras is tight to the sequence itself, because the distance between the objects of the scene and the cameras also needs to be taken into account. Initial conclusions are mostly derived from visual quality inspection by a group of four experts. Section 2 briefly describes the new test set provided by the ULB. Section 3 gives some preliminary results on how many views can be synthesized from an objective point of view, while Section 4 gives some preliminary clues from a subjective point of view.

2 Description of the content

The ULB test set [4] is a high density LightField content captured with a 2D rail robotic system. The resulting scene is static and was captured using Kinect2 RGB sensor (1920x1080@24bits) and a Kinect2 depth sensor (512x424@16bits).

2.1 Scene description

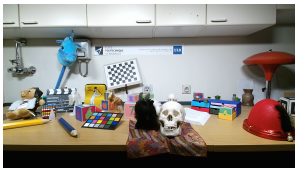
The whole scene is 2.3m wide and composed by conventional objects, semi-transparent ones and objects with fur. There is also a rotated checkerboard and color chart. The closest object of scene is 0.6m and the farthest is 1.6m from the sensor. The platform with camera is moved millimeter by millimeter, which provides a very high density of views. The 1mm precision for positioning the camera vertically

and horizontally was obtained by several motors and a rail system composed of ball bearings.

2.2 Pre-processing of the content

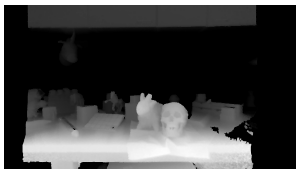
The Kinect contains two distinct cameras for the color and depth images, consequently the extrinsic and intrinsic parameters for these images are different. Because of this, the depth images had to be reprojected onto the coordinate system of the color images. The resulting depth map needs to indicate for each pixel of the color image, an orthogonal distance to the color image's camera plane. The resulting most left and right views are shown in Figure 1 with their corresponding depth maps.

(a) *Leftmost texture*



(b) *Rightmost texture*

(c) *Leftmost depth*



(d) *Rightmost depth*

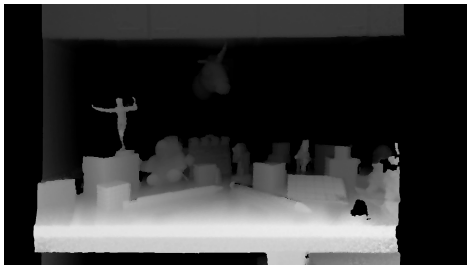
Figure 1: *Leftmost and rightmost views of the ULB test set, with corresponding depth maps.*

Due to the fact that RGB and depth sensors have different field of view, the depth information is missing on the left and right sides of the views, as shown in figure 1. When depths are used for synthesis, the resulting synthesized view exhibits severe artifacts on the borders too, as shown on figure 2. These artifacts prevent from drawing conclusions from the tests:

1.Objective results obtained from synthesis are largely biased by these left and right areas.

2.Subjective tests are biased (viewers assess the whole view).

3.Some bits are used to encode these useless areas of the depths, once compression is involved.



(a) *Depth*

(b) *Synthesized view*

Figure 2: *Uncropped depth and resulting synthesized view.*

We consequently have cropped the original data. This cropping is a simple manual post-processing of the depth. The number of removed columns compared to the original is 192 on the left side, 320 on the right side, yielding to a resolution of 1408x1080. The texture input has been cropped similarly. In preliminary tests it was shown that quality of synthesis is better from cropped content, so only cropped version will be used for objective results and subjective quality evaluation.

The main advantage of this new content provided by the ULB is that it gives the ability to synthesize large number of views between two views, while still having a reference anchor for objective tests, even if the distance between the two cameras is small. So far, this could only be obtained with computer generated content.

3 View synthesis – Objective results

For view synthesis, we have used VSRS1D-fast [2], which has been extensively used during the standardization process of 3D-HEVC. It was shown that VSRS1D-fast can provide better results compared to VSRS4.1 for 1D linear content. Comparison of these two synthesis tools is shown in Figure 3.

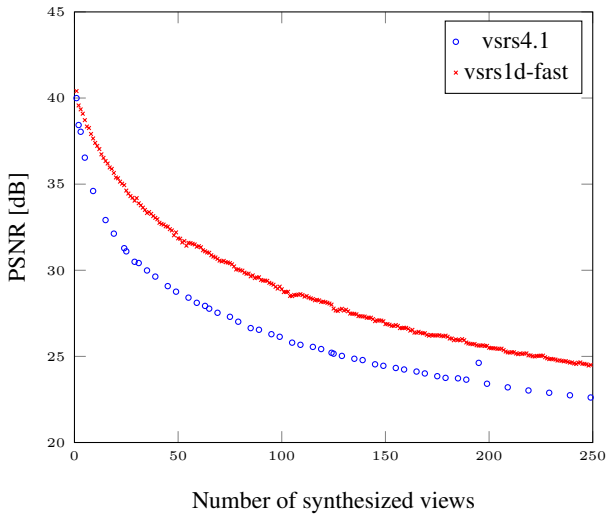
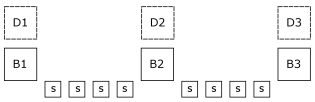
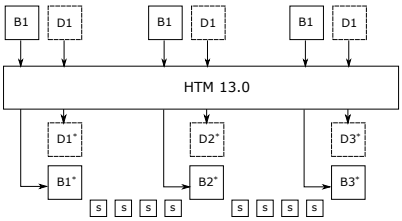


Figure 3: *Average PSNR for original data comparing VSRS4.1 and VSRS1D-fast.*



(a) *Synthesis from uncompressed base views*



(b) *Compression of base views before synthesis.*

Figure 4: *Overall scheme of study*

In the following experiments, intermediate views are synthesized between two captured (reference) views. The scheme of the study is shown in Figure 4. We consider two different scenarios. In the first one synthesis out of original uncompressed data is done, figure 4a. In the second case all the base views are compressed using HTM13.0 reference software, and the synthesis is done from decoded views. The synthesis out of compresses views is shown in Figure 4b, where views with 'star' denotes decoded texture or depth.

To present results in this paper the following notation is used: Synth-n,b means that n views are synthesized between two captured views distant from b millimeters.

3.1 Synthesis out of uncompressed views

The first part of study is related to the synthesis of intermediate views from original uncompressed data, as it is shown in Figure 4(a).

Experiment 1. The distance b between two reference views is progressively increased, and the corresponding PSNR loss is provided. Figure 3 shows the results for all configurations from Synth-b,b+1 for b=1...249. Among the 851 views available, only 751 views were used for PSNR calculation (view 1 to view 751). The figure 3 can be read this way: if for Synth-b,b+1 the PSNR is p dB, it means that when b views are synthesized between two cameras distant by b+1 millimeters, and the average PSNR of the n synthesized views is p dB. If we consider a Free Navigation application where the user can only watch real captured view, and move from one to the other, and synthesis is used to smooth the navigation effect, we believe this representation, computing the average of views, as represented in figure 3 is representative.

We can observe that the average PSNR drops from 40.4dB for Synth-1,2 configuration to 24.5dB for Synth-249,250. An average 35dB quality is achieved when synthesizing about 23 consecutive views, which correspond to 2.4cm distance between two cameras. An average 30 dB quality is achieved when synthesizing about 78 consecutive views, which corresponds to 7.9cm distance between two cameras.

Experiment 2. Between two reference views, the PSNR of the synthesized views is not constant: it depends on the distance with the closest reference view. This is depicted in figure 5 that show the PSNR variation for different Synth-b,b+1 configurations. We can observe that for all configurations, the PSNR decreases when the synthesized view is far from one of the two reference views. The worst quality is close to the middle point between the two base views. The PSNR is inconsistent from one view to another. This can be seen easily on configuration Synth-1,2 but the issue holds for other configurations. Some views are particularly degraded as for instance view 550. One of the possible sources of such errors can be erroneous camera parameters. In order to explain the issue, figure 6 represents the square error difference image between original and synthesized images for view 550 and for view 580. Blue color on this heat map represents the minimum error, and red the largest. While the difference for view 580 is small, all the contours are visible in the difference image of view 550, which tends to confirm that camera parameters for view 550 is erroneous or there is misalignment between depth and texture.

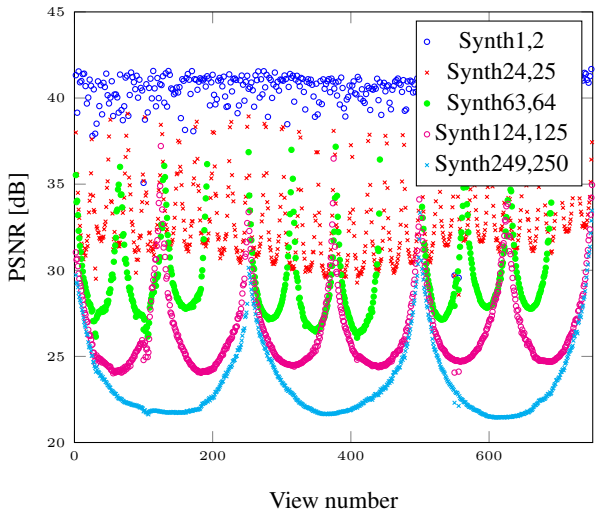
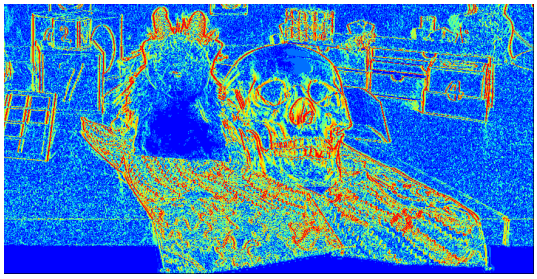
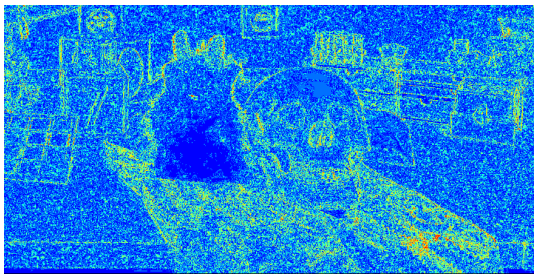


Figure 5: *PSNR evolution for different Synth-n,b configurations.*



(a) *view 550*



(b) *view 580*

Figure 6: *Square error heat map for Synth-1,2.*

Experiment 3. figure 7 represents the evolution of the quality for Synth-1,b configuration, with b varying from 2 to 848. This means that one single view is synthesized for two reference cameras distant by b millimeters. Said differently, we represent the PSNR of the worst point of figure 5. If we consider a Free Navigation application where the user can stop on any view (synthesized or captured), we believe the min value as reported in figure 7 is representative: it represents the worst case.

In this case, we observe that a quality of 35dB is achieved by skipping 15 views, this leads to a possible distance of 16mm, with the ULB content, while a quality of 30dB leads to a distance of 51mm.

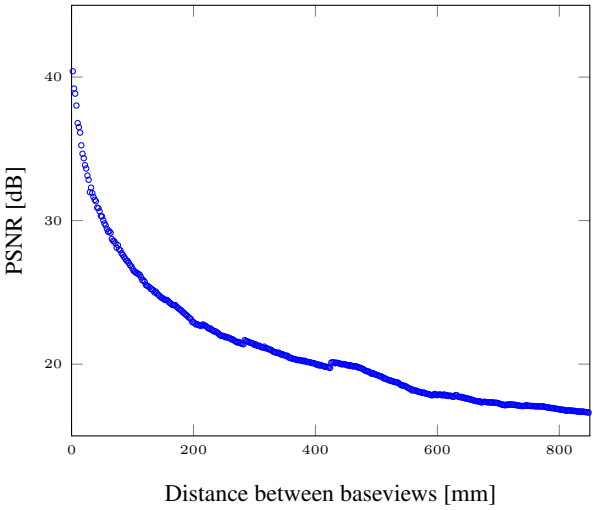


Figure 7: *Min PSNR for configuration from Synth-1,2 to Synth-1,848*

It has been demonstrated many times that the PSNR is very sensitive to view synthesis artifacts [5], such as shifts, while human quality evaluation makes abstraction of this parameter. So most of the time it reflects lower quality than the perceived one. In the next section, subjective quality is assessed.

3.2 Synthesis out of compressed views

For this study base views were compressed using HTM13.0 reference software, simulating the scenario where base views are available on decoder side and intermediate views can be synthesized to improve the smoothness of transition between base views, which allows the user to have better immersive experience.

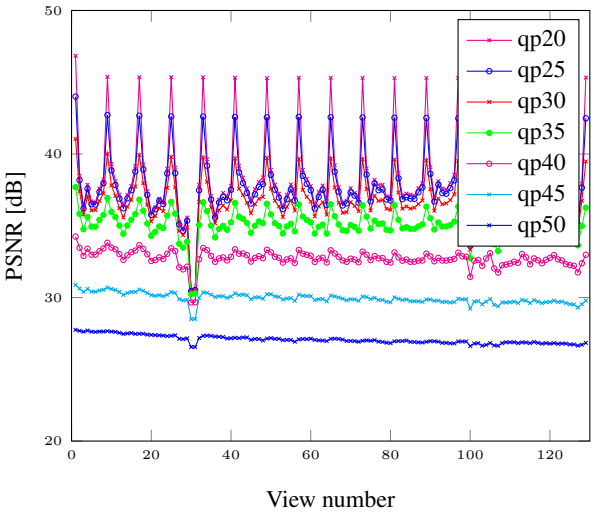


Figure 8: *Synth-7,8 with compressed baseviews*

Figures 8, 9 and 10 represent the PSNR of synthesized views for different quantization parameter's (QP) values.

In Figure 8 it is shown that for the synthesized views, the average PSNR, as well as the min value of the PSNR, are decreasing when the QP increases. The average PSNR of synthesized views for QPs equal to 20 and 50 respectively is 38.2dB and 26.5dB. This difference is huge, meaning that the synthesis is significantly impacted by the compression.
Increasing the distance between base views mainly influences on the synthesized views as it is shown in Figure 9. The coding does not suffer from increasing the distance to the inter-view prediction reference picture, as the content is very dense. The PSNR values of synthesized views become closer to each other. In Figure 10 all synthesized views for different QPs have practically the same quality in terms of PSNR. So if the synthesis is erroneous due to the large distance between base views the quality of base views does not play an important role.

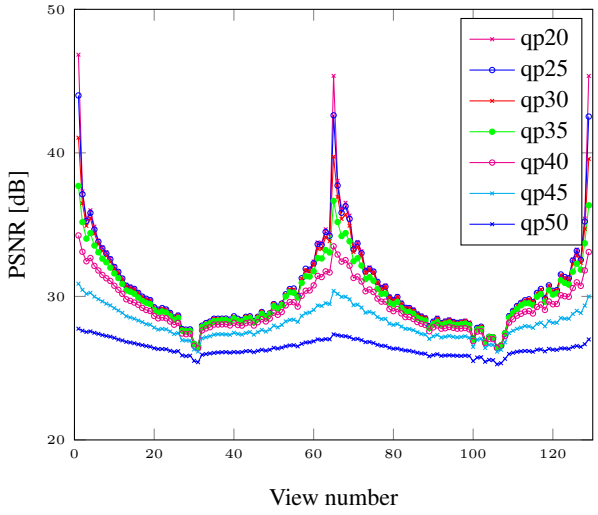


Figure 9: *Synth-63,64 with compressed baseviews*

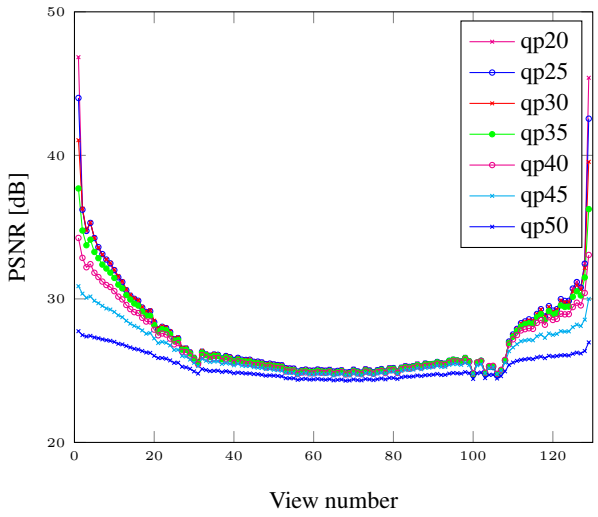


Figure 10: *Synth-127,128 with compressed baseviews*

4 View synthesis – Subjective results

According to the objective results, an acceptable quality is obtained only for a very small baseline. However, it has been shown that PSNR is more sensitive to artifacts from synthesis than the human eye is [5]. The goal of this section is to verify from a subjective point of view if it is possible to synthesize views with more distant cameras.
We have not performed subjective tests, as there is no standardized procedure to perform tests for this kind of scenario. The reminder should more be seen as a visual quality inspection, where four different expert viewers have shared their opinions.

4.1 Subjective results - Uncompressed data

For this quality inspection, viewers have basically expressed their opinion with the following words, to rate the quality of the transition between the two reference views separated by d millimeters:
1: Artifacts are not noticeable.
2: Artifacts are noticeable but not annoying.
3: Artifacts are noticeable and annoying.
4: Artifacts are too annoying (I would prefer to switch directly from one view to another without smoothness/synthesis.)

To simulate a scenario, when the user swaps the views on the tablet or smart phone, we have built the following sequence:
1.The viewer watches the same view (reference view) for 1.5 seconds
2.The viewer moves to the right, watches synthesized and reference views, and stops on a reference view located d millimeters away from the initial one.
3.The viewer watches this reference view for 1.5 seconds.
For this experiment the views in the range between 251 and 751 are used. The sequence is generated at 60fps, because in the preliminary tests this configuration provided for most of the Synth-n,b configurations better impression of the smooth transition between base views.
Several Synth-d,d+1 configurations have been presented to the viewers, starting from Synth-249,250. As reported in Table 1, viewers have generally agreed that the quality is unacceptable. As a consequence, d has been decreased progressively, etc.
From this table, we can conclude that:

- for d < 31, barely no artifact is observed.
- for 31< d < 124, some artifacts are noticeable but qualified as non-annoying (except for one view-er).
- for 124 < d < 249, artifacts are noticeable and qualified as annoying.
- for d > 249, the level of artifacts is too annoying, and synthesis it not accepted as a feature to make the transition smoother.

Configuration	PSNR [dB]	V1	V2	V3	V4	Avg
Synth-249,250	24.5	4	4	4	4	4.00
Synth-124,125	28.0	2	2	3	2	2.25
Synth-99,100	29.1	2	2	3	2	2.25
Synth-63,64	31.1	1	2	3	1	1.75
Synth-31,32	33.9	1	1	1	1	1.00
Orig, no synth	-	1	1	1	1	1.00

Table 1: *Average PSNR of each synthesized view for different Synth-d,d+1 configuration and the corresponding subjective scores. Tests performed on the cropped data set.*

Synth-n,n+1	QP35-QP25		QP50-QP35	
	psnr, dB	subjective	psnr, dB	subjective
31	-1.4	-0.3	-6.2	2.5
63	-0.8	0.5	-4.4	2.8
127	-0.4	0.3	-2.4	3
249	-0.2	0.4	-1.0	1.9

Table 2: *Comparison of deferences for different QPs for subjective scores and PSNR .*

4.2 Subjective results - Compressed data

For the synthesis out of compressed base views the following view inspection was done: the process of sweeping between two compressed base views is emulated.
The way, how the sequence is generated, is similar to the case with uncompressed data. But for this inspection viewers watch two sequences one after another. First is synthesis out of uncompressed data and second is synthesis from compressed base views. The viewers give theirs opinion on quality of the second sequence compared to the first using the continuous scale from -3 to +3, where +3 means second sequence is much better than the first one. Sequences appear randomly, so the viewer does not know which one is synthesized from the compressed data.
As it is shown in the Table 2, the difference between QP35 and QP25 subjectively was not visible by most of the expert viewers. For some experiments, for example, for Synth249,250 and QP 35 participants considered better quality for the configuration with compression rather than synthesis from uncompressed data.

5 Conclusion and future work

Several view synthesis experiments have been performed, using the new high density ULB content. From these experiments, some conclusion and suggestions can be made:
According to the PSNR, the acceptable distance between two cameras is very low.
The result of the visual quality inspection, performed by four expert viewers, confirms that from a subjective point of view, more views can be synthesized while maintaining acceptable quality. It could be suggested that objective (PSNR based) results can only be considered to rank two algorithms applied to the same configuration. PSNR gives an idea of the ranking (comparison), it evolves correctly (higher PSNR usually fits with higher subjective quality). But it does not reflect the overall quality. Even when considering subjective quality, the distance between two cameras that yields to acceptable quality for virtual views remains low.
The maximal distances, allowing acceptable quality for synthesized views, found by those experiments are quite small. However, the quality of synthesis depends not only on baseline between two cameras, but also the distance to the closest and farthest objects in the scene. With the ULB test set, the objects are very close to the scene, and this drastically reduces the allowed distance between two cameras.
In experiments with compression it was shown that the gap between compressed base views and synthesized intermediate views becomes smaller when the quality of base views becomes lower(QPs 45-50). So the transition between base views becomes smoother. Subjective inspection results showed a correlation with objective tests.
For the future perspectives improving the quality of the depths will drastically increase view synthesis quality and there is a room for further improvements of view synthesis algorithms in the near future. It is also important to mention, that depending on the application, constraints related to view synthesis will be different. The method described in [6] can be used in the future for comparing view synthesis algorithms.

References

[1] Wegner, Stankiewicz, Tanimoto, et Domanski. Enhanced view synthesis reference software (VSRS) for free-viewpoint television. Dans *ISO/IEC JTC1/SC29/WG11*. m31520, October 2013.

[2] Zhang, Tech, Wegner, et Yea. 3D-HEVC test model 5. Dans *ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*. JCT3V-E1005, July 2013.

[3] Lafruit, Wegner, et Tanimoto. Draft call for evidence on FTV. Dans *ISO/IEC JTC1/SC29/WG11*. m40293, February 2015.

[4] Bonatto, Lenertz, Li, Schenkel, et Lafruit. [MPEG-I-visual/apps] ULB high density 2D camera array data set, version 1. Dans *ISO/IEC JTC1/SC29/WG11*. m40293, April 2017.

[5] Dricot, Jung, Cagnazzo, Pesquet, Dufaux, Kovacs, et Kiran Adhikarla. Subjective evaluation of super multiview compressed content on high end light field 3D display. *Elsevier Signal Processing: Image Communication*, 39:369–385, November 2015.

[6] Purica, Valenzise, Cagnazzo, Pesquet-Popescu, et Dufaux. Using region-of-interest for quality evaluation of DIBR-based view synthesis methods. September 2016.

Vers un alignement spatio-temporel du visage en conditions non contrôlées

R. Belmonte[1,3,*] N. Ihaddadene[1,*] P. Tirilly[2,†] M. Bilasco[3,†] C. Djeraba[2,†]

[1] ISEN Lille, Yncréa Hauts-de-France, France
[*]{romain.belmonte, nacim.ihaddadene}@yncrea.fr

[2] Univ. Lille, CNRS, Centrale Lille, IMT Lille Douai, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

[3] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
[†]{pierre.tirilly, marius.bilasco, chaabane.djeraba}@univ-lille1.fr

Résumé

L'alignement du visage est une tâche essentielle pour un grand nombre d'applications. Elle a pour objectif de localiser des points caractéristiques sur le visage, dans le but d'identifier sa structure géométrique. En conditions non contrôlées, les différentes variations pouvant intervenir dans le contexte visuel, associées à l'instabilité de la détection du visage, en font un problème difficile à résoudre. Bien que de nombreuses méthodes aient été proposées, leurs performances en présence de ces contraintes ne sont toujours pas satisfaisantes. Dans cet article, nous invitons à étudier l'alignement du visage à l'aide de séquences d'images et non d'images fixes, comme cela a pu être réalisé jusqu'à présent, et montrons l'importance de la prise en compte de l'information temporelle en conditions non contrôlées.

Mots clefs

Alignement du visage, approche temporelle, points caractéristiques, séquences d'images, conditions non contrôlées.

1 Introduction

Le problème de l'alignement du visage, aussi appelé localisation des points caractéristiques du visage, a suscité beaucoup d'intérêt et connu de rapide progrès ces dernières années [1]. Etant donné la position et la taille d'un visage, l'alignement, illustré en figure 1, consiste à déterminer la géométrie des composantes du visage contenant le plus d'information sémantique (pour ex., les yeux, le nez, la bouche). Cette capacité à modéliser les structures non rigides du visage est aujourd'hui exploitée dans divers domaines tels que l'analyse faciale (pour ex. identification, expressions), l'interaction homme-machine, ou le multimédia (pour ex. recherche, indexation). Toutefois, malgré le nombre important de méthodes présentes dans la littérature, les performances de l'alignement du visage en conditions non contrôlées restent limitées [2].

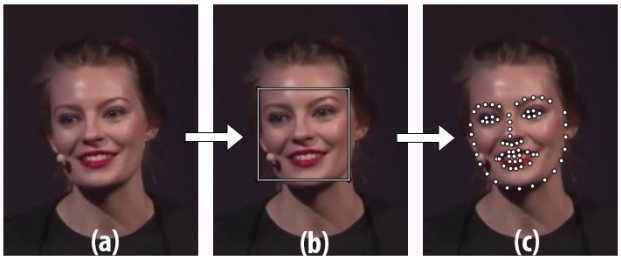


Figure 1 – *Processus d'alignement du visage : (a) image originale, (b) détection du visage, (c) alignement du visage. Images issues de 300VW [3].*

Encore aujourd'hui, ce problème continue d'être étudié à l'aide d'images fixes [1]. Pourtant, du fait de l'omniprésence des capteurs vidéos, la grande majorité des applications reposent sur des séquences d'images. De plus, bon nombre de tâches liées à l'analyse faciale ou plus largement, à l'analyse du comportement humain, ont su tirer profit de l'information temporelle [4, 5]. Les premières synthèses sur l'alignement du visage ont déjà suggéré d'étudier le problème à l'aide de séquences d'images, mais sans toutefois avancer de réels arguments [6]. Notre motivation à travers cet article est de montrer que la prise en compte de l'information temporelle pour ce problème pourrait grandement contribuer à l'amélioration des performances en conditions non contrôlées.

Cet article est structuré de la manière suivante : dans la section 2, nous décrivons les raisons qui nous ont conduit à nous orienter vers une approche spatio-temporelle pour l'alignement du visage. Nous passons notamment en revue les solutions existantes et mettons en perspective les performances des méthodes les plus récentes. Nous consacrons ensuite la section 3 aux bénéfices de l'information temporelle en conditions non contrôlées. Enfin, nous concluons avec la section 4.

2 État des lieux

2.1 Jeux de données et métriques d'évaluation

Ces dernières années, de nombreux corpus pour l'alignement du visage ont été mis à la disposition de la communauté scientifique (c.f. Tableau 1). Les images incluses dans ces corpus ont été collectées sur des réseaux sociaux tels que Google, Flickr, ou Facebook, apportant ainsi plus de réalisme à l'étude du problème. L'annotation a été réalisée soit manuellement, soit de manière semi-automatique, avec parfois l'aide de la plate-forme Amazon Mechanical Turk. La qualité des annotations peut toutefois être assez variable [2, 7, 6]. Le schéma d'annotation utilisé (c.-à-d., position et nombre de points) peut également différer d'un jeu de données à un autre. À l'heure actuelle, c'est le schéma composé de 68 points [8], illustré en Figure 1, qui est le plus largement utilisé. Ce schéma ne permettant pas de conserver une correspondance pour la totalité des points lors de poses extrêmes, un schéma basé sur 39 points a récemment été proposé pour les visages de profil [9].

Type	Base de données	#Images	#Points
Statique	AFLW [10]	25.993	21
	AFW [11]	250	6
	HELEN [12]	2330	194
	iBug [13]	135	68
	LFPW [14]	1432	35
	COFW [15]	1007	29
	300W [2]	3837	68
	300W-LP [16]	61.225	68
	MENPO [9]	∼9000	68/39
Dynamique	300VW [3]	218.595	68

Tableau 1 – *Jeux de données capturés en conditions non contrôlées.*

L'évaluation d'une prédiction se fait généralement à l'aide de l'erreur quadratique moyenne normalisée par la distance interoculaire. Au delà de 8%, la prédiction est généralement considérée comme un échec. La normalisation par la distance interoculaire, bien que peu robuste aux poses extrêmes, est la plus répandue. D'autres normalisations sont parfois utilisées, comme notamment la diagonale de la fenêtre de détection. Sur un ensemble d'images, l'erreur moyenne est la métrique d'évaluation la plus simple et intuitive à calculer. Toutefois, elle peut être fortement impactée par la présence de quelques erreurs aberrantes. Une représentation graphique de la fonction de répartition de l'erreur est, de ce fait, de plus en plus souvent employée. Elle correspond à la proportion d'images pour laquelle l'erreur est inférieure ou égale à un certain seuil (par ex., 8%). L'aire sous la courbe et le taux d'échec, c'est-à-dire le pourcentage d'images pour lesquelles l'erreur est supérieure au seuil défini, sont parfois calculés à partir de cette représentation.

Aujourd'hui, du fait notamment de l'émergence des techniques d'apprentissage profond, il peut être nécessaire de disposer d'un grand nombre d'images annotées, ce que ne permettent pas forcément les jeux de données existants. Dans la littérature, différents procédés d'augmentation sont utilisés pour contourner ce problème. Certaines opérations peuvent être appliquées sur les images afin d'en générer de nouvelles (par ex., rotation, inversion horizontale, perturbation de la fenêtre de détection). D'autres procédés plus complexes tels que la génération d'images de synthèses commencent également à être employés [16].

Figure 2 – *Illustration des défis rencontrés en conditions non contrôlées : occultations (b), (d), (f), variations de pose (a), (e), éclairage (a), (b), expressions (c). Images issues de 300VW [3].*

Il est crucial d'avoir des données représentatives du problème afin de pouvoir y répondre. Les jeux de données statiques ne couvrent pas toutes les difficultés rencontrées par les applications, largement basées sur des séquences d'images. Les contraintes notamment liées au mouvement des personnes ou de la caméra ne sont actuellement pas considérées. Un corpus composé de séquences d'images capturées en conditions non contrôlées a toutefois récemment été publié par Shen et al. [3] (cf. Figure 2). Ces données, en plus d'être plus représentatives du problème, fournissent des éléments de réflexion en faveur de l'utilisation de l'information temporelle pour l'alignement du visage (cf. Section 3).

2.2 Synthèse des solutions

Dans la littérature se distinguent deux grandes catégories de méthodes pour localiser les points caractéristiques du visage. Il y a tout d'abord les méthodes dites génératives, qui s'appuient sur des modèles paramétriques conjoints d'apparence et de forme [17]. L'alignement est alors formulé comme un problème d'optimisation avec comme objectif de trouver les paramètres permettant de générer la meilleure instance possible du modèle pour un visage donné. L'apparence peut être représentée aussi bien de ma-

nière holistique que locale, à l'aide de régions d'intérêt centrées sur les points caractéristiques.

Il y a ensuite les méthodes dites discriminatives qui infèrent la position des points caractéristiques directement à partir de l'apparence du visage. Cela est rendu possible soit par l'apprentissage de détecteurs locaux indépendants ou de régresseurs pour chaque point caractéristique associés à un modèle de forme permettant de régulariser les prédictions [18], soit par l'apprentissage d'une ou plusieurs fonctions de régression vectorielles capablent d'inférer l'ensemble des points caractéristiques et de conserver implicitement une contrainte de forme [19, 20, 21, 22]. Dans cette catégorie, les méthodes basées sur des techniques d'apprentissage profond (par ex., réseaux de neurones convolutionnels, auto-encodeurs) ont récemment permis d'améliorer de manière conséquente les performances en conditions non contrôlées grâce notamment à leur capacité à modéliser la non-linéarité et à apprendre des caractéristiques spécifiques au problème [23, 24, 25, 26].

Bien que la plupart des méthodes s'attaquent de manière globale au problème, certaines se concentrent spécifiquement sur une difficulté [15, 27, 16]. Burgos-Artizzu et al. [15] se sont intéressés à modéliser explicitement les occultations et ont montré que cette information supplémentaire aidait à améliorer l'estimation de la position des points caractéristiques en conditions non contrôlées. L'apprentissage nécessite cependant un travail conséquent d'annotation des occultations. Zhu et al. [16] se sont eux focalisés sur les poses extrêmes et ont proposé d'inférer un modèle dense 3D plutôt qu'un modèle éparse 2D. Leur méthode est capable de gérer des variations de pose horizontale allant de -90°à 90°.

D'autres travaux suggèrent que l'alignement du visage ne doit pas être traité comme un problème indépendant et proposent d'apprendre conjointement différentes tâches connexes afin d'obtenir des gains de performances individuels [28, 29]. Dans les travaux de Zhang et al. [29], l'alignement n'est pas appris de manière isolé mais conjointement avec différentes tâches connexes telles que l'estimation de la pose, du genre, des expressions faciales et de l'apparence des attributs faciaux. Ce type d'approche peut toutefois rendre l'étape d'apprentissage beaucoup plus complexe en raison des taux de convergence pouvant être variables d'une tâche à une autre.

Les performances des méthodes d'alignement du visage les plus récentes sont référencées dans le Tableau 2. Ces méthodes ont été évaluées sur le jeu de données 300W, composé de catégories de difficultés variables. La catégorie 300-A correspond à des images ne présentant pas de fortes contraintes. La catégorie 300-B contient des images plus complexes présentant de fortes variations de pose et d'expression, ainsi que des occultations. On constate que pour la catégorie B l'erreur moyenne vaut plus du double de celle obtenue sur la catégorie A. Les résultats toutes catégories confondues ne sont pas forcément pertinents du fait du déséquilibre entre les deux catégories (554 images pour la catégorie A contre 135 pour la catégorie B).

Méthode	300W-A	300W-B	300W
FPLL [11]	8.22	18.33	10.20
RCPR [15]	6.18	17.26	8.35
DRMF [30]	6.65	19.79	9.22
SDM [19]	5.57	15.40	7.50
ESR [20]	5.28	17.00	7.58
CFAN [31]	5.50	16.78	7.69
LBF [21]	4.95	11.98	6.32
RCR [32]	4.83	12.02	6.24
CFSS [22]	4.73	9.98	5.76
CMC-CNN [23]	4.91	12.03	6.30
RPPE [27]	5.50	11.57	6.69
3DDFA [16]	6.15	10.59	7.01
TCDCN [29]	4.80	8.60	5.54
RAR [24]	4.12	**8.35**	**4.94**
RCFA [25]	**4.03**	9.85	5.32
R-DSSD [26]	4.16	9.20	5.59

Tableau 2 – *Performances des méthodes récentes de la littérature. L'erreur quadratique normalisée moyenne est reportée.*

Malgré la quantité de méthodes proposées dans la littérature et les avancées majeures récentes, nous pouvons voir à travers ces résultats que les problèmes rencontrées en conditions non contrôlées sont encore loin d'être résolus. Du fait de leur influence significative sur l'apparence du visage, les variations de pose et les occultations font partie des défis les plus difficiles à relever. Nous montrons dans la Section 3 comment l'approche temporelle pourrait contribuer à résoudre ces problèmes.

3 Les bénéfices de l'information temporelle

3.1 Suivi non rigide

La détection du visage en conditions non contrôlées est un problème complexe à résoudre [33]. Étant donné son rôle pour l'alignement du visage, Yang et al. [34] se sont intéressés à la dépendance susceptible de s'exercer entre ces deux tâches. Leur étude a pu mettre en évidence une forte sensibilité de l'alignement à la détection. Ainsi, au-delà d'une incapacité à détecter un visage, d'autres facteurs tels que les variations d'échelle et de position de la fenêtre de détection peuvent venir perturber l'alignement.

Une solution pour s'abstraire de la dépendance à la détection est d'effectuer un suivi non rigide du visage. Shen et al. [3] ont récemment proposé une analyse comparative des méthodes actuelles de suivi non rigide spécifique au visage. La stratégie la plus populaire reste le suivi par détection, c'est-à-dire la détection du visage et son alignement sur chaque image de manière indépendante, sans tirer profit

Méthode	Catégorie 1		Catégorie 2		Catégorie 3	
	AUC	FR(%)	AUC	FR(%)	AUC	FR(%)
HyperFace [28]	0.642	5.56	0.662	0.68	0.563	7.23
Yang et al. [35]	**0.791**	**2.400**	**0.788**	**0.322**	**0.710**	**4.461**
Uricar et Franc [40]	0.657	7.622	0.677	4.131	0.574	7.957
Xiao et al. [41]	0.760	5.899	0.782	3.845	0.695	7.379
Rajamanoharan et Cootes [42]	0.735	6.557	0.717	3.906	0.659	8.289
Wu et Ji [43]	0.674	13.925	0.732	5.601	0.602	13.161

Tableau 3 – *Comparaison des méthodes de l'analyse comparative de Shen et al. [3] avec Hyperface [28], méthode multitâche statique, sur les 3 catégories de 300VW. L'aire sous la courbe (AUC) et le taux d'échec (FR) sont reportés.*

des images adjacentes. Yang et al. [35] ont toutefois montré des résultats supérieurs à cette stratégie en proposant une régression en cascade spatio-temporelle. Leur méthode consiste à initialiser la forme sur l'image courante à partir des paramètres de similitude de l'image précédente. Elle intègre également un mécanisme de ré-initialisation basé sur la qualité de la prédiction afin d'éviter toute dérive de l'alignement. Comparée au suivi par détection, la régression en cascade spatio-temporelle permet de réduire considérablement le taux d'échec tout en améliorant les performances (cf. Tableau 3).

Une alternative au suivi par détection consiste à utiliser, en substitution à la détection, un algorithme de suivi générique (c.-à-d. rigide). L'un des avantages des algorithmes de suivi génériques est qu'ils sont capables de tenir compte de certaines variations d'apparence de l'objet cible durant le suivi [36]. Chrysos et al. [37] ont évalué cette stratégie et l'ont comparé au suivi par détection. De manière générale, le suivi générique permet d'être plus robuste aux contraintes rencontrées en conditions non contrôlées. Il est cependant probable que, tout comme avec la détection, l'alignement soit sensible aux variations de la fenêtre de suivi.

Tenir compte des variations d'apparence sans passer par un algorithme de suivi générique devient alors judicieux. Sanchez et al. [38] ont proposé une régression en cascade continue incrémentale. Contrairement à Yang et al. [35] qui conservent un modèle générique une fois l'apprentissage effectué, ici un modèle pré-entraîné est mis à jour en ligne afin de devenir spécifique à la personne au cours du suivi. Ce genre d'approche montre de meilleurs résultats qu'avec un modèle générique mais n'exploite cependant que les variations d'apparence. D'autres informations telles que la trajectoire des points caractéristiques à travers la séquence d'images semblent pertinentes à considérer [39].

3.2 Contraintes supplémentaires

Qu'elle soit explicite ou implicite, la contrainte de forme présente dans la plupart des méthodes d'alignement est un élément crucial pour obtenir de bonnes performances en conditions non contrôlées. Dans des séquences d'images, une contrainte supplémentaire peut être appliquée sur la trajectoire des points caractéristiques. Peng et al. [44] ont utilisé un auto-encodeur récurrent afin d'exploiter les caractéristiques dynamiques du visage. Ils ont comparé une version récurrente et une non récurrente de leur auto-encodeur et ont montré que l'apprentissage récurrent améliorait, d'une part, la stabilité des prédictions et, d'autre part, la robustesse aux occultations, variations de pose et d'expression. Ils supposent notamment que le réseau de neurones récurrent a permis l'apprentissage de motifs de trajectoires, mais ne l'ont cependant pas démontré.

Une autre option pour tenir compte de la dynamique faciale est l'utilisation de filtres bayésien tels que les filtres de Kalman ou les filtres particulaires. Gu et al. [45] ont cependant mis en évidence le gain marginal apporté par ces approches et on montré la supérioté des réseaux de neurones récurrents pour l'analyse dynamique du visage. Les performances de leur méthode sont référencées dans le Tableau 4. La prise en compte de contraintes supplémentaires montre une réduction notable de l'erreur moyenne de plus de 1% comparée à une méthode statique actuelle.

Les réseaux de neurones récurrents, bien qu'avantageux, ne sont capables de caractériser que le mouvement global. Dans d'autres tâches connexes, le mouvement local (c.-à-d. sur quelques trames) parfois associé au mouvement global a toutefois su montrer son intérêt [46, 5] et pourrait être tout aussi bénéfique à l'alignement du visage.

Approche	**Méthode**	**300VW**
Statique	TCDCN [29]	7.59
Dynamique	RED-NET [44]	6.25
	Gu et al. [45]	**6.16**

Tableau 4 – *Comparaison de TCDCN [29], méthode multitâche statique, avec 2 méthodes dynamiques basées sur des réseaux de neurones récurrents [44, 45]. L'erreur quadratique normalisée moyenne est reportée.*

4 Conclusion

Dans ce papier nous avons présenté une synthèse des travaux sur l'alignement du visage en conditions non contrôlées. Malgré des applications majoritairement basées sur des séquences d'images, ce problème est depuis plusieurs décennies étudié à partir d'images fixes. Un nombre impressionnant de méthodes sont proposées dans la litté-

rature. Toutefois leurs performances en conditions non contrôlées ne sont toujours pas satisfaisantes du fait notamment des nombreuses variations pouvant intervenir dans le contexte visuel. Nous avons montré qu'étudier le problème à l'aide de séquences d'images pouvait grandement contribuer à pallier ces difficultés en plus d'améliorer la cohérence avec les applications. Nous pensons que l'ajout de contraintes temporelles est une voie intéressante à suivre.

Références

[1] Xin Jin et Xiaoyang Tan. Face alignment in-the-wild : A survey. *arXiv preprint arXiv :1608.04188*, 2016.

[2] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, et Maja Pantic. 300 faces in-the-wild challenge : Database and results. *Image and Vision Computing*, 47 :3–18, 2016.

[3] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, et Maja Pantic. The first facial landmark tracking in-the-wild challenge : Benchmark and results. Dans *ICCV Workshops*, pages 1003–1011. IEEE, 2015.

[4] Karen Simonyan et Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. Dans *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[5] Yin Fan, Xiangju Lu, Dian Li, et Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. Dans *ICMI*, pages 445–450. ACM, 2016.

[6] Oya Çeliktutan, Sezer Ulukaya, et Bülent Sankur. A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1) :13, 2013.

[7] Adrian Bulat et Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem ?(and a dataset of 230,000 3d facial landmarks). *arXiv preprint arXiv :1703.07332*, 2017.

[8] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, et Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5) :807–813, 2010.

[9] S Zaferiou. The menpo facial landmark localisation challenge. Dans *CVPR Workshops*, volume 1, 2017.

[10] Martin Köstinger, Paul Wohlhart, Peter M Roth, et Horst Bischof. Annotated facial landmarks in the wild : A large-scale, real-world database for facial landmark localization. Dans *ICCV Workshops)*, pages 2144–2151. IEEE, 2011.

[11] Xiangxin Zhu et Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. Dans *CVPR*, pages 2879–2886. IEEE, 2012.

[12] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, et Thomas Huang. Interactive facial feature localization. *ECCV*, pages 679–692, 2012.

[13] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, et Maja Pantic. A semi-automatic methodology for facial landmark annotation. Dans *CVPR Workshops*, pages 896–903, 2013.

[14] Peter N Belhumeur, David W Jacobs, David J Kriegman, et Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 35(12) :2930–2940, 2013.

[15] Xavier P Burgos-Artizzu, Pietro Perona, et Piotr Dollár. Robust face landmark estimation under occlusion. Dans *ICCV*, pages 1513–1520, 2013.

[16] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, et Stan Z Li. Face alignment across large poses : A 3d solution. Dans *CVPR*, pages 146–155, 2016.

[17] Timothy F. Cootes, Gareth J. Edwards, et Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6) :681–685, 2001.

[18] Jason M Saragih, Simon Lucey, et Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2) :200–215, 2011.

[19] Xuehan Xiong et Fernando De la Torre. Supervised descent method and its applications to face alignment. Dans *CVPR*, pages 532–539, 2013.

[20] Xudong Cao, Yichen Wei, Fang Wen, et Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2) :177–190, 2014.

[21] Shaoqing Ren, Xudong Cao, Yichen Wei, et Jian Sun. Face alignment at 3000 fps via regressing local binary features. Dans *CVPR*, pages 1685–1692, 2014.

[22] Shizhan Zhu, Cheng Li, Chen Change Loy, et Xiaoou Tang. Face alignment by coarse-to-fine shape searching. Dans *CVPR*, pages 4998–5006, 2015.

[23] Qiqi Hou, Jinjun Wang, Lele Cheng, et Yihong Gong. Facial landmark detection via cascade multi-channel convolutional neural network. Dans *ICIP*, pages 1800–1804. IEEE, 2015.

[24] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, et Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. Dans *ECCV*, pages 57–72. Springer, 2016.

[25] Wei Wang, Sergey Tulyakov, et Nicu Sebe. Recurrent convolutional face alignment. Dans *ACCV*, pages 104–120. Springer, 2016.

[26] Hao Liu, Jiwen Lu, Jianjiang Feng, et Jie Zhou. Learning deep sharable and structural detectors for face alignment. *IEEE Transactions on Image Processing*, 26(4) :1666–1678, 2017.

[27] Heng Yang, Xuming He, Xuhui Jia, et Ioannis Patras. Robust face alignment under occlusion via regional

predictive power estimation. *IEEE Transactions on Image Processing*, 24(8) :2393–2403, 2015.

[28] Rajeev Ranjan, Vishal M Patel, et Rama Chellappa. Hyperface : A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv :1603.01249*, 2016.

[29] Zhanpeng Zhang, Ping Luo, Chen Change Loy, et Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5) :918–930, 2016.

[30] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, et Maja Pantic. Robust discriminative response map fitting with constrained local models. Dans *CVPR*, pages 3444–3451, 2013.

[31] Jie Zhang, Shiguang Shan, Meina Kan, et Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. Dans *ECCV*, pages 1–16. Springer, 2014.

[32] Chengchao Qu, Hua Gao, Eduardo Monari, Jurgen Beyerer, et Jean-Philippe Thiran. Towards robust cascaded regression for face alignment in the wild. Dans *CVPR Workshops*, pages 1–9, 2015.

[33] Stefanos Zafeiriou, Cha Zhang, et Zhengyou Zhang. A survey on face detection in the wild : past, present and future. *Computer Vision and Image Understanding*, 138 :1–24, 2015.

[34] Heng Yang, Xuhui Jia, Chen Change Loy, et Peter Robinson. An empirical study of recent face alignment methods. *arXiv preprint arXiv :1511.05049*, 2015.

[35] Jing Yang, Jiankang Deng, Kaihua Zhang, et Qingshan Liu. Facial shape tracking via spatio-temporal cascade shape regression. Dans *ICCV Workshops*, pages 41–49, 2015.

[36] Matej Kristan, Aleš Leonardis, Jiři Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin, Tomas Vojir, Gustav Häger, Alan Lukežič, et Gustavo Fernandez. *The Visual Object Tracking VOT2016 Challenge Results*, pages 777–823. Springer International Publishing, Cham, 2016.

[37] Grigorios G. Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, et Stefanos Zafeiriou. A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision*, pages 1–35, 2017.

[38] Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, et Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. Dans *ECCV*, pages 645–661. Springer, 2016.

[39] Ghassan Hamarneh et Tomas Gustavsson. Deformable spatio-temporal shape models : Extending asm to 2d+ time. Dans *BMVC*, pages 1–10, 2001.

[40] Michal Uricár, Vojtech Franc, et Václav Hlavác. Facial landmark tracking by tree-based deformable part model based detector. Dans *ICCV Workshops*, pages 10–17, 2015.

[41] Shengtao Xiao, Shuicheng Yan, et Ashraf A Kassim. Facial landmark detection via progressive initialization. Dans *ICCV Workshops*, pages 33–40, 2015.

[42] Georgia Rajamanoharan et Timothy F Cootes. Multi-view constrained local models for large head angle facial tracking. Dans *ICCV Workshops*, pages 18–25, 2015.

[43] Yue Wu et Qiang Ji. Shape augmented regression method for face alignment. Dans *ICCV Workshops*, pages 26–32, 2015.

[44] Xi Peng, Rogerio S Feris, Xiaoyu Wang, et Dimitris N Metaxas. A recurrent encoder-decoder network for sequential face alignment. Dans *ECCV*, pages 38–56. Springer, 2016.

[45] Jinwei Gu Xiaodong Yang Shalini De et Mello Jan Kautz. Dynamic facial analysis : From bayesian filtering to recurrent neural network.

[46] Behzad Hasani et Mohammad H Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. *arXiv preprint arXiv :1705.07871*, 2017.

Image coding using Leaky Integrate-and-Fire neurons

Melpomeni Dimopoulou, Effrosyni Doutsi, Marc Antonini
Université Côte d'Azur, CNRS, I3S, France

mel.dimopoulou@gmail.com
doutsi@i3s.unice.fr
am@i3s.unice.fr

Résumé

This paper aims to build an image coding system based on the model of the mammalian retina. The retina is the light-sensitive layer of tissue located on the inner coat of the eye and it is responsible for vision. Inspired by the way the retina handles and compresses the visual information and based on previous studies we aim to build and analytically study a retinal-inspired image quantizer, based on the Leaky Integrate-and-Fire (LIF) model, a neural model according to which function the ganglion cells of the Ganglionic retinal layer that is responsible for the visual data compression. In order to have a more concrete view of the encoder's behavior, in our experiments, we make use of the spatiotemporal decomposition layers provided by extensive previous studies on a previous retinal layer, the Outer Plexiform Layer (OPL). The decomposition layers produced by the OPL, are being encoded using our LIF image encoder and then, they are reconstructed to observe the encoder's efficiency.

Mots clefs

Retina, Ganglion cells, Leaky Integrate-and-Fire (LIF) model, neural coding, image coding,weighted difference of Gaussians.

1 Introduction

As technology advances, the need for finding new ways for the efficient transmission and storage of information augments dramatically. Living in the age of the social networks, the media to be stored and transmitted grows rapidly. However, despite the fact that during the past few decades compression standards kept evolving, the compression ratio does not evolve accordingly to the needs. Consequently, the urge for finding new means of compression remains to be of a high importance. With this paper, we aim to propose a different, bio-inspired, dynamic approach for the encoding of images.

Our work is being inspired by the mammalian visual system and more specifically, by the way the retina works for the perception and compression of natural images. The retina can be divided into three basic layers. The Outer Plexiform Layer (OPL) which acts as a spatiotemporal filter on images, the Inner Plexiform Layer (IPL) that performs a non-linear rectification, and the Ganglionic Layer which is responsible for the encoding of the data. The Ganglionic layer consists of the ganglion cells, a type of neuron which compresses visual information, in response to the brightness of light. The Ganglion cells function according to the Leaky Integrate-and-Fire (LIF) neural model which encodes intensity values into spikes. Under the main belief that nature performs in an optimal way, and based on previous works on the OPL filtering in [1], we built a quantization system making use of the LIF properties to be applied on the compression of images already filtered by the OPL. This quantization scheme, unlike the already existing static encoding algorithms, encodes images in a dynamic way and then using an inverse function the encoded information provides an estimation of the initial image.

In section 2, we are going to provide the theoretical background for the LIF, explaining the physical and biological function of the LIF neural model. We also describe the LIF encoding and decoding process performed by the LIF quantizer. In section 3, we discuss about the OPL and the way it acts as a spatiotemporal filter on input images producing decomposition layers. Furthermore we analyze the procedure for the reconstruction. Finally, in section 4, we present our experiments on the extended encoding system for the case of uniform selection of OPL subbands and a non-uniform one which emphasizes on the most informative layers.

2 The LIF

2.1 Background

As described in [2], the LIF is a neural model which is described by the circuit shown in Figure 1. The input current $I(t), t \in \mathbb{R}^+$ is being divided in the current I_R, which passes through the resistor and the current I_C which charges the capacitor. Given the Ohm's law for I_R and the definition of capacity as $C = q/u$ (where q is the charge and u the voltage) the total current can be written as :

$$I(t) = I_R + I_C = \frac{u(t)}{R} + C\frac{du}{dt}. \quad (1)$$

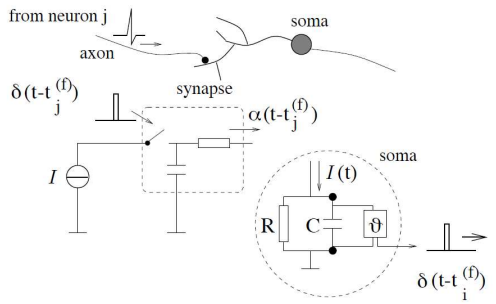


Figure 1 – *The LIF neuron circuit which consists of a resistance R in parallel with a capacitor C (Figure taken from [2]).*

By multiplying eq. (1) by R and by introducing a time constant $\tau_m = RC$ the equation becomes :

$$\tau_m \frac{du}{dt} = -u(t) + RI(t). \tag{2}$$

In the integrate-and-fire model, the form of an action potential is not described explicitly. Spikes are generated at a firing time $t^{(f)}$. This firing time is defined by the following threshold criterion :

$$t^{(f)} : u(t^{(f)}) = \theta. \tag{3}$$

Immediately after $t^{(f)}$ the potential is set to a new value $u_r < \theta$,

$$\lim_{t \to t^{(f)}; t > t^{(f)}} u(t) = u_r. \tag{4}$$

While $t < t^{(f)}$ the dynamics is given by eq. (2) until the next threshold crossing occurs. The LIF neuron may also incorporate a refractory period. In this case, if u reaches the threshold at time $t = t^{(f)}$, the dynamics is interrupted during an absolute refractory time Δ^{abs} and the integration restarts at time $t^{(f)} + \Delta^{abs}$ with a new initial condition.
Let's consider the simple case of a constant input current stimulus $I(t) = I_0$. For the sake of simplicity we will assume a reset potential $u_r = 0$. Assuming that the k^{th} spike has occurred at time $t = t^k$ when the trajectory of the membrane potential is given by integrating eq. (2) with the initial condition $u(t) = u_r = 0$. The solution is given by the relation :

$$u_k(t) = RI_0 \left[1 - exp\left(-\frac{t - t^k}{\tau_m}\right)\right]. \tag{5}$$

After each spike, the potential is reset to the value $u_r = 0$ and the integration process starts again. The condition $u(t) = \theta$ is satisfied for $t = t^{k+1}$, where t^{k+1} denotes the time when the next spike occurs. Then, eq. (3) can be written as following :

$$u(t^{k+1}) = \theta = RI_0 \left[1 - \exp\left(-\frac{t^{k+1} - t^k}{\tau_m}\right)\right]. \tag{6}$$

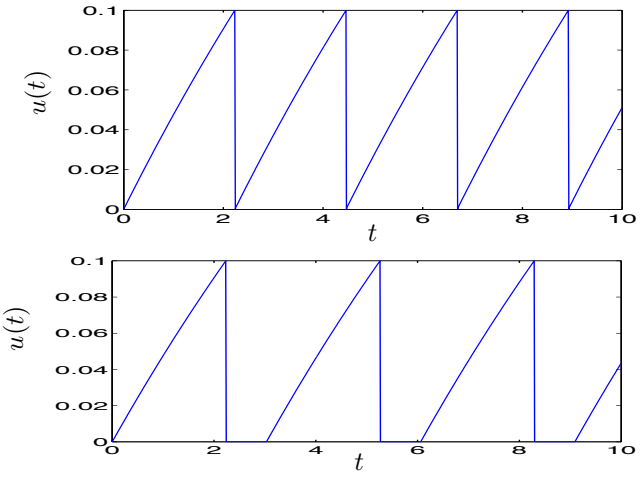


Figure 2 – *The LIF firing process - Top : model with no refractory period. Bottom : model with a constant refractory period $\Delta^{abs} = 2$ ms. In the model without a refractory period the next integration starts exactly after the spike emission while in the presence of a constant refractory period the new integration begins delayed by a time period equal to Δ^{abs} after the latest spike emission time.*

We assume $d(RI_0) = t^{k+1} - t^k$, the inter-spike delay of an integrate-and-fire neuron with no refractory period, which depends on the input current I_0. For example, the higher the amplitude of the current is, the smaller the delay.
Consequently, solving (6) for the delay $d(RI_0)$ and simplifying the notation setting $u = u(I_0) = RI_0$ yields :

$$d(u) = \begin{cases} \infty, & u < \theta \\ h(u;\theta) = \tau_m \ln(\frac{u}{u-\theta}), & u \geq \theta, \end{cases} \tag{7}$$

The firing rate of the LIF neuron, is then given by the relation $\nu = 1/d(u)$.
At this point, it is important to denote that for the case of a neuron with an absolute refractory period, the occurrence of the next spike will be delayed by the duration of the refractory period Δ^{abs}. So, in this case, the inter-spike delay $d'(u)$ is given by :

$$d'(u) = d(u) + \Delta^{abs} = t^{k+1} - t^k + \Delta^{abs}, \tag{8}$$

where $t = t^{k+1} - t^k + \Delta^{abs}$ is the time instance when the next integration will start after the emission of the $(k+1)^{th}$ spike. The firing process of a LIF neuron is described by Figure 2.

2.2 The LIF Quantizer

The LIF quantizer, which has analytically been studied and explained in [3], uses the LIF properties described in the previous section to encode input intensities into numbers of spikes. More specifically, the LIF Quantizer works according to the following procedure. In the encoder, according to Ohm's law, we compute the action potential of the

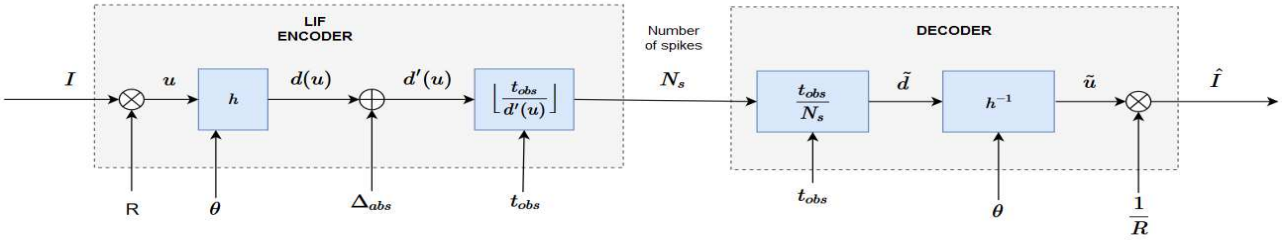


Figure 3 – *The LIF encoding and decoding process*

LIF neuron from the relation $u = RI$. Then using eq. 7 we compute the integration delay $d(u)$. By adding the refractory period we get the interspike delay $d'(u)$.
In the decoder, the floor value of the division of the parameter t_{obs} by the interspike delay $d'(u)$ gives the number of spikes. Following the inverse procedure, dividing the parameter t_{obs} by the encoded number of spikes and using the estimated delay for the computation of the membrane potential estimation, we finally get the quantized values of the input current $\hat{I}$ by dividing by the resistance R. The process of the quantization using the LIF is shown in Figure 3.
There were some previous attempts to use similar bio-inspired coding schemes [4, 5]. However, both these works tried to approximate the LIF model by conventional compression tools. The originality of this work with respect to the other ones is related to the implementation of a real LIF quantizer with refractory period.

2.3 Refractory period

As described in section 2.1, after the production of a spike in the LIF model there can exist a refractory period. For our Quantizer this refractory period works as an additive noise. In this paper we are going to use a half-Gaussian refractory period. This means that after each spike we generate a random positive refractory period of a specified variance. More specifically the refractory period of our model behaves according to the following distribution :

$$P(x) = A \exp\left(-B(x^2)\right), \tag{9}$$

where $A = \frac{\sqrt{2}}{\sigma\sqrt{\pi}}$ and $B = \frac{1}{2\sigma^2}$ for a given standard deviation σ.

3 The Outer Plexiform Layer (OPL)

3.1 Background

The Outer Plexiform Layer (OPL) is the first layer of the retina which consists of a dense network of synapses between the photoreceptor, horizontal and bipolar cells (see Fig. 4) [6]. The input of the OPL cells is the visual stimulus $f(x,t), x \in \mathbb{R}^2, t \in \mathbb{R}^+$ which is spatiotemporally transformed into the equivalent electrical signal. This electrical signal is dynamically encoded by the neurons into a code of spikes which is propagated to the analysis center; the visual cortex. The neural code is not used to reconstruct

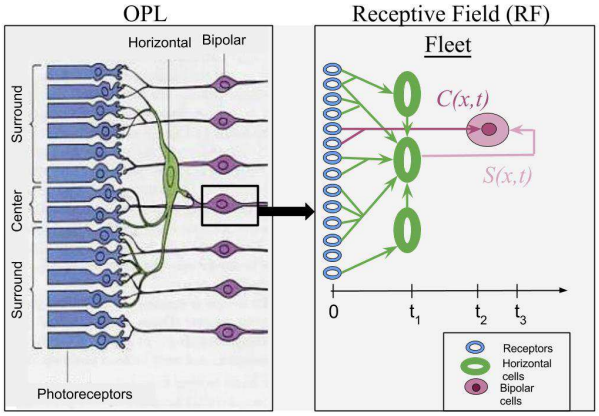


Figure 4 – *The OPL structure (Figure taken from [7]).*

the input signal but to learn and take decisions. However, our goal is to build a bio-inspired coding/decoding system for images. Figure 5 shows the proposed coding/decoding architecture we aim to build.

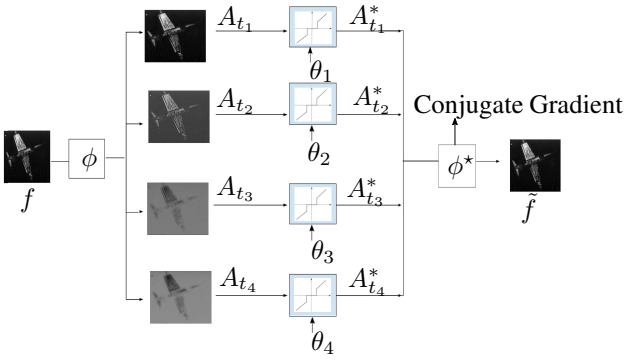


Figure 5 – *Retina-inspired image codec. The figure describes the decomposition of the input image f using the retina-inspired filter ϕ into several layers A_{t_j} each one of which is sparsified by the LIF quantizer Q. Based on the quantized layers $A^*_{t_j}$, one is able to reconstruct $\tilde{f}$ and compute the distortion comparing to f.*

The dynamic OPL transformation which is caused to the visual stimulus while it is captured and propagated through the OPL cells has been approximated by the the retina-

inspired filter [1, 7]. The retina-inspired filter $\phi(x,t)$ is a novel Weighted Difference of Gaussian (WDoG) [7] which models the center-surround structure of the receptive field of the bipolar cells :

$$\phi(x,t) = a(t)G_{\sigma_C}(x) - b(t)G_{\sigma_S}(x), \qquad (10)$$

where $a(t)$ and $b(t)$ are two time-varying weights which tune the shape of the DoG, σ_c and σ_s are the standard deviations of the center and the surround Gaussians respectively with $\sigma_c < \sigma_s$.
The retina-inspired filtering, which is a frame, is applied to temporally constant input signals $f(x,t) = f(x)\mathbf{1}_{[0\leq t\leq T]}(t)$ resulting in high redundancy :

$$A(x,t) = \phi(x,t) \overset{x}{*} f(x), \qquad (11)$$

where $\overset{x}{*}$ is a spatial convolution. Let $t_1, \ldots, t_m$ some temporal samples. For each time instant $t_j, j = 1, \ldots, m$ there is a different decomposition layer $A_{t_j} = A(x,t_j)$. This redundancy is sufficient to perfectly reconstruct the input signal $\tilde{f}$.

3.2 Reconstruction

It is proven in [1] that the retina-inspired filter is a frame hence, the filter is invertible meaning that it is possible to reconstruct the input image. In practice, one needs to solve the linear system $A = \Phi f$ and reconstruct $\tilde{f}$. At time $t = T$, the exact estimation of $\tilde{f} = f$ according to :

$$\tilde{f} = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top A, \qquad (12)$$

where Φ^{-1} denotes the inverse of a matrix Φ and $\Phi^\top$ denotes its transpose. The retina-inspired filter Φ is a frame, as a result, we can define as $\Phi^T\Phi$ its frame operator. However, in practice, due to the large size of matrix Φ and in order to avoid a time consuming and resource demanding reconstruction processing, we used the conjugate gradient descent which is one of the most efficient iterative methods [8]. We are interested in reducing this redundancy and discard all the coefficients of low energy keeping only the most informative ones for the reconstruction.

4 Experiments

4.1 Results on one subband

In our experiments we first tested the LIF quantizer on a single subband, in order to understand and evaluate the quantizer's behavior. Let $x_1, \ldots, x_n$ some spatial samples such that $A_{t_j} = \big(A(x_1,t_j), \ldots, A(x_n,t_j)\big), j = 1, \ldots, m$ a discrete decomposition layer. The LIF quantizer is applied to every single spatiotemporal sample $A(x_k,t_j)$ where $k = 1, \ldots, n$. For the experiment, we have chosen grayscale images of the size $n = 512 \times 512$ pixels taken from USC-SIPI database [9]. As described in section 2.3, in our tests, we are going to use a random refractory period which follows a half-Gaussian distribution.

Figure 6, shows the evolution of the Mean Squared Error (MSE) between the original image and the decoded one using the LIF, in function of the value of θ for different values of observation time t_{obs}. It is clear that, the refractory period introduces overload noise on the input image which yields the existence of an optimal threshold value θ that minimizes the MSE. This optimal θ value is different according to the value of the observation time t_{obs}.

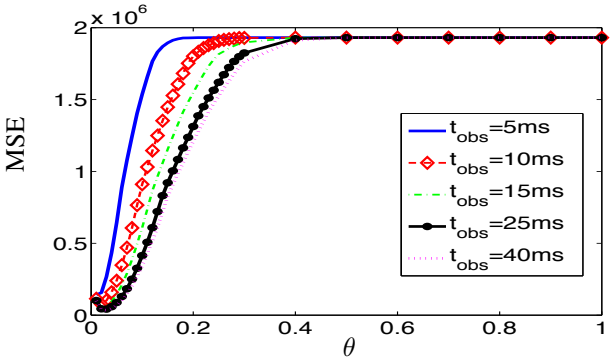


Figure 6 – *The MSE curve in function of the threshold parameter θ for different observation times.*

In order for our quantizer to be adaptive to the needs of the quantization process, we are going to select for each realisation the appropriate value of theta for the LIF.

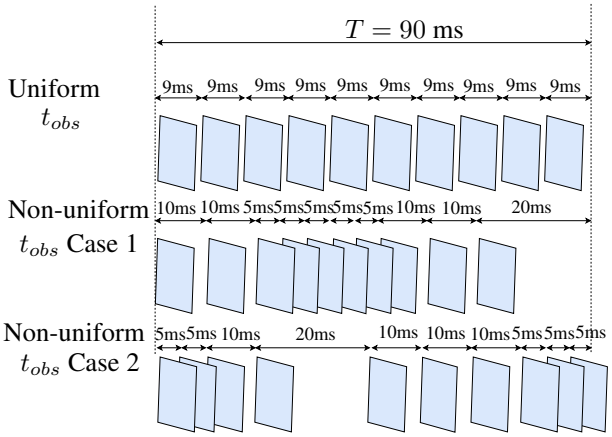


Figure 7 – *Subband generation rate. Case 1, corresponds to a subband generation rate with dense middle bandpass frames, corresponds to a subband generation rate with sparse middle bandpass frames.*

4.2 Subband Generation using the OPL

The purpose of this paper is to experiment on the application of the LIF quantizer on each of the subbands produced by the retina-inspired filter and evaluate the quality and the efficiency of the extended system depicted in Figure 5. Extended studies in [7] have shown that the amount of information on the subbands produced by the OPL decomposition varies while time evolves. More specifically, according

$t_1 = 5ms$ $t_{obs} = 10ms$ | $t_2 = 15ms$ $t_{obs} = 10ms$ | $t_3 = 25ms$ $t_{obs} = 5ms$ | $t_4 = 30ms$ $t_{obs} = 5ms$ | $t_5 = 35ms$ $\Delta t_{obs} = 5ms$

$t_1 = 40ms$ $t_{obs} = 5ms$ | $t_2 = 45ms$ $t_{obs} = 5ms$ | $t_3 = 50ms$ $t_{obs} = 10ms$ | $t_4 = 60ms$ $t_{obs} = 10ms$ | $t_5 = 70ms$ $t_{obs} = 20ms$

Figure 8 – *The non-uniform subband generation using the OPL filtering and the corresponding time of appearance* t_1, t_2, t_3, t_4, t_5 *and observation time* t_{obs} *respectively.*

to the bio-plausible filtering parameters given in [7], in the very first subbands the range of the intensity values is very small while in the last subbands (i.e. $t \geq 120ms$) there is no big change in the subbands' content. Consequently, in order to achieve a sparse reconstruction and reduce the redundancy of the latest subbands, in our experiments we are going to generate 10 subbands in the range $0 \leq t \leq 90ms$. As a first step, we tested the generation of 10 subbands uniformly distributed in the total filtering range, observing each of the produced layers for a $t_{obs} = 9ms$ as described in Figure 7.

Moving on, we experimented on the non-uniform case, trying the two different non-uniform schemes shown in Figure 7. The first one, corresponds to an attempt to keep most of the middle and most informative subbands in the bandpass range $25ms \leq t \leq 50ms$. The subbands produced by this scheme are visually presented in Figure 8. In this case, although we keep most of the middle informative subbands, we observe each layer for a shorter observation time t_{obs}. Then we also tried the second non-uniform scheme, depicted in Figure 7, which corresponds to a subband generation with sparser layers in the bandpass range of observation times. In this second case, while we keep less of the informative subbands, they are better encoded, as we observe them for a longer observation time. At this point, we should mention the fact that this is only a first experimental attempt to apply the LIF to the layers produced by the OPL filter, in order to evaluate and better understand the properties of our proposed encoder. As a result, our subband selection for the non-uniform sampling cases has been experimentally achieved, without using some specific function.

For our experiments, we have used two different images. After the subband generation we apply the LIF quantization on each of the generated subbands, we reconstruct the encoded layers and evaluate the quality of the reconstructed image compared to the original one as described in Figure 5. In Figure 9 we present the visual results of our experiments showing also the values of the Entropy, the Peak Signal to Noise Ratio (PSNR), and the Structural Similarity Index (SSIM)[10].

We observe that for the first image, the non-Uniform subband generation with the denser subbands in the bandpass area provides a better value of PSNR and SSIM compared to the uniform case, while the entropy is being slightly reduced. On the contrary, the nonuniform generation with sparser subbands in the middle observation times behaves poorly in comparison to the uniform generation. For the second image though, depicted in the lower part of Figure 9, we observe that both non-uniform cases of subband generation provides better results of PSNR and SSIM than the uniform case, with the denser middle subband generation behaving slightly better than the sparser middle subbands case.

Consequently, we can assume that the selection of the good θ value according to the observation time as well as the good rate of subband generation in the OPL filtering, can provide very promising results and significantly improve the rate-distortion trade off. In addition to this, we conclude that the good rate of subband generation varies according to the image characteristics (statistics, content).

4.3 Conclusions

In this work we have implemented an extended retina-inspired compression system. This is an innovative approach which uses a dynamic way of quantization adapted to the needs of the encoding process unlike the existing encoding algorithms. Our study, reveals the fact that this bio-inspired dynamic encoding process can provide very promising results. The good choice of layers produced by the OPL filter, plays an important role to the quality of the image reconstruction and gives a strong motive to further

Original Image

PSNR $= 17.0814$ dB
SSIM $= 0.5204$
$H = 3.316$ bpp

PSNR $= 15.1268$ dB
SSIM $= 0.4635$
$H = 4.704$ bpp

PSNR $= 24.7936$ dB
SSIM $= 0.8187$
$H = 3.1$ bpp

Original Image

PSNR $= 14.7250$ dB
SSIM $= 0.4843$
$H = 4.769$ bpp

PSNR $= 19.8719$ dB
SSIM $= 0.7204$
$H = 4.592$ bpp

PSNR $= 20.4562$ dB
SSIM $= 0.7384$
$H = 6.611$ bpp

Figure 9 – *The comparison of the visual results and quality metrics of the PSNR, SSIM and Entropy for the a)original image (first image on the left) b)the uniform subband generation(second image from left to right) c) the non-Uniform subband generation with a sparser middle (second image from left to right) and d)the non-uniform scheme with a denser middle sumbbands (first image on the right)*

study the behavior of the model according to the total observation time. Since this is a very first attempt to apply this extended encoding system on images, we should underline the significance of improving these results by further experimenting and studying the system's behavior. Furthermore, since in our experiments we used an experimental way of non-uniform subband generation, the use of a particular function that will be able to minimize the rate distortion trade-off according to the image characteristics, is a very important future step that should be studied and implemented.

Références

[1] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, "Retina-inspired filtering for dynamic image coding," *IEEE International Conference in Image Processing (ICIP)*, pp. 3505–3509, 2015.

[2] W. Gerstner and W. Kistler, *Spiking neuron models : Single Neurons, Populations, Plasitcity*, Cambridge University Press, 2002.

[3] M. Dimopoulou and M. Antonini, "Signal Quantization using a Leaky Integrate-and-Fire neuron," in *GRETSI*, 2017.

[4] K Masmoudi, M. Antonini, and P. Kornprobst, "Streaming an image through the eye : The retina seen as a dithered scalable image coder," *Signal Processing : Image Communication*, vol. 28, no. 8, pp. 856–869, 2013.

[5] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, "Bio-inspired Sparse Representation of Images," in *Gretsi*, 2017, number 1, pp. 2–5.

[6] H. Kolb, "How the Retina Works," *American Scientist the magazine of Sigma Xi, The Scientific Research Society*, vol. 91, pp. 28–35, 2004.

[7] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, "Retina-inspired Filtering," *hal-01350686*, 2016.

[8] J R Shewchuk, "An Introduction to the Conjugate Gradient Method Without the Agonizing Pain," *Science*, vol. 49, no. CS-94-125, pp. 64, 1994.

[9] A. Weber, "The USC-SIPI Image Database," 1977.

[10] A Horé and D Ziou, "Image quality metrics : PSNR vs. SSIM," *20th International Conference on Pattern Recognition, ICPR*, pp. 2366–2369, 2010.

Biométrie, forensics et protection du contenu

Sessions 2 et 3

Study on color space for the performance of visible wavelength iris recognition

Xinwei Liu[1,2] Christophe Charrier[1] Marius Pedersen[2] Patrick Bours[2]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
[2] NTNU - Norwegian University of Science and Technology, Gjøvik, Norway
xinwei.liu@unicaen.fr

Abstract

With the introduction of visible wavelength iris recognition, color iris images are captured during the acquisition process. Compared to the traditional near infrared iris images, the representation of color information must be considered in visible wavelength iris recognition. In this paper, we propose to study different color space components for representing an image for the visible wavelength iris recognition. Additionally, image-based degradations are introduced to iris images for the analysis of how different color space components influence the performance of iris recognition on distorted images.

Key words

Iris recognition, visible wavelength, color space, image-based degradations, performance.

1 Introduction

Biometrics are more and more popular in recent years. Among all the existing biometric modalities, iris recognition is one of the well known technologies. Thanks to the development of color imaging technology and biometric recognition application, iris images captured under visible wavelength by classical devices such as smartphone, webcams or compact camera can be used for iris recognition [1]. Unlike the traditional near infrared iris images, visible wavelength iris samples are color images that are represented in the most commonly used RGB color space. By including color information in iris recognition, the unique characteristics that are exclusive to color iris image have become relevant to recognition performance [2]. However, many of existing iris recognition approaches only using gray-scale images that converted from RGB images by taking the average of three color components. The drawback of this method is that we do not consider the impact of each color component in a RGB image, and, moreover, we ignore the influence of other color spaces than RGB space. It has been investigated that the recognition performance varies by using different components in RGB color space [3]. The impact of color components on recognition performance has lead us to further analyze the representation of color through various color spaces to try to explore additional information from the color image that can be used to increase recognition performance. Another factor that influence the performance of iris recognition is the quality of iris sample images. The visible wavelength iris images can be captured under unconstrained environment conditions. Therefore, some image-based degradations could be introduced during acquisition process. Using such iris images for recognition is a more challenging issue compared to the traditional near infrared iris biometrics. In order to investigate how above mentioned two factors affect the performance of iris recognition, the goal of this paper is to determine whether an optimal color space can gives better performance than traditional RGB space on distorted iris images. This paper is organized as follows. Section 2 presents state-of-the-art concerning the studies of color space and sample quality for iris recognition. In Section 3, we introduce the experiment setup. Section 4 illustrates the experimental results, and we conclude this work in Section 5.

2 State-of-the-art

There is not many study on iris color space analysis in the literature. Boyce *et al.* [4] and Monaco [3] investigated the influence of different color space components on iris recognition performance. In order to convert between the original RGB color space and one of the other color spaces, they first segmented and normalized the iris samples. Thus, a RGB iris template is used as the baseline image on which all color space transforms are employed. Since they only care about the transformation of the visible wavelength iris images, the near infrared component is omitted. The normalized template is then subjected to a color space transform. Once the transform is complete, comparison is performed on each channel of the color image independently. In addition to the RGB color space, CIE Lab, YCbCr, HSV, and CMYK color spaces are used in [3, 4]. The findings in their papers are : there is no single color space transformation was found to increase matching performance across all components. However, individual components of certain color spaces showed potential as ideal candidates for iris recognition. In color spaces such as CIE Lab and YCbCr, the luminosity functions showed consistently high performance across all eye color classes. When the chromaticity is completely segmented from the luminosity, such as in the CIE Lab and YCbCr color spaces, the chromaticity shows

to have high correlations between its two components indicating that their performance and viability as templates are similar. Overall, the traditional RGB color space showed strong performance.

Recent research by Liu *et al.* [5] investigated how the image-based degradations influence the biometric quality assessment and the performance of visible wavelength iris recognition. A new multi-modality biometric database, which includes visible wavelength iris images was created. Unlike the other databases, there is no modality-based degradations in the database (e.g. occlusion). This database is used for investigating how image-based degradations affect iris recognition performance.

3 Experiment setup

3.1 Visible wavelength iris image database - GC^2 Multi-modality Biometric Database

If we want to investigate how image-based distortions affect the performance of visible wavelength iris recognition by taking into account color spaces, it is recommended to use a database without modality-based distortions. Most of the existing visible wavelength iris databases contain both image-based and modality-based distortions. The 'GC^2 Multi-modality Biometric Database' is a new database without modality-based distortions. For the visible wavelength iris sub-dataset in this database, three cameras were used to capture iris images : 1) a Lytro [6] first generation Light Field Camera (LFC) (11 Megapixels), 2) a Google Nexus 5 embedded camera (8 Megapixels), and 3) a Canon D700 with Canon EF 100mm f/2.8L Macro Lens (18 Megapixels). There are 50 subjects in the dataset and 15 samples images are taken for each eye (left and right) per subject. Totally, 4500 iris images are obtained in the iris dataset.

In order to obtain image-based distortions, we follow the protocol introduced in [5, 7] : five image-based attributes are used to degrade iris images. We used Matlab R2016a to conduct the experiment as following :

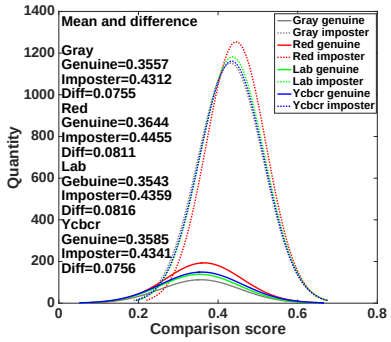
- **Contrast distortions**. There are two kinds of contrast distortions : too low and too high contrast. We use Matlab function '$J = imadjust(I, [low_{in}; high_{in}], [low_{out}; high_{out}])$', which maps the values in I (original iris image) to new values in J (degraded iris image) such that values between low_{in} and $high_{in}$ map to values between low_{out} and $high_{out}$. For low contrast, the low_{in} and $high_{in}$ values are set to 0 and 0.4, low_{out} and $high_{out}$ values are set to 0 and 1. For high contrast, the low_{in} and $high_{in}$ values are set to 0.6 and 1, low_{out} and $high_{out}$ values are set to 0 and 1.
- **Sharpness distortions**. We generate two sharpness distortions : motion blur and Gaussian blur. For motion blur we use Matlab function '$h = fspecial('motion', len, theta)$', which returns a filter to the linear motion of a camera by len pixels, with an angle of $theta$ degrees in a counterclockwise direction. The len value is set to 30 and the $theta$ is set to 45. For Gaussian blur we use function '$h = fspecial('gaussian', hsize, sigma)$', which returns a rotationally symmetric Gaussian lowpass filter of size $hsize$ with standard deviation $sigma$ (positive). The $hsize$ value is set to [3 3] and the $sigma$ is set to 0.5.
- **Luminance distortion**. There are two kinds of luminance distortions : too low and too high luminance. We use Matlab function '$J = imadjust(I, [low_{in}; high_{in}], [low_{out}; high_{out}])$', again to simulate luminance distortions. For low luminance, the low_{in} and $high_{in}$ values are set to 0 and 1, low_{out} and $high_{out}$ values are set to 0 and 0.2. For high luminance, the low_{in} and $high_{in}$ values are set to 0 and 1, low_{out} and $high_{out}$ values are set to 0.8 and 1.
- **Artifacts**. We introduce two artifacts to iris images : poisson noise and JPEG compression artifacts. We use Matlab function '$J = imnoise(I,' poisson')$' to add poisson noise and the JPEG compression ratio is 0.1.
- **Color distortions**. Since we need to investigate how different color spaces affect the performance of iris recognition. The iris images might be recognized in an other color space by mistake. Here we convert RGB iris images to HSV and YCbCr color spaces to simulate this situation.

3.2 Color space

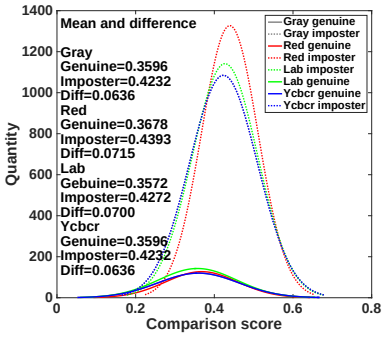
According to the analysis results from [3], three components from three color spaces other then grayscale have better performance then the rest components : Red channel from RGB color space, L channel from CIE-Lab color space, and Y channel from YCbCr channel. Therefore, we evaluate the performance of iris recognition system by representing iris image in four color components : grayscale, Red, L, and Y. The color space transformations have been done in Matlab R2016a by using default scripts.

3.3 Iris recognition system

The iris recognition system used in this paper is proposed by Masek [8]. It is an open-source iris recognition system that verify both the uniqueness of the human iris and also its performance as a biometric. This iris recognition system consists of an automatic segmentation system that is based on the Hough transform, and is able to localize the circular iris and pupil region, occluding eyelids and eyelashes, and reflections. The extracted iris region is then normalized into a rectangular block with constant dimensions to account for imaging inconsistencies. Finally, the phase data from 1D Log-Gabor filters is extracted and quantized to four levels to encode the unique pattern of the iris into a bit-wise biometric template. The Hamming distance is

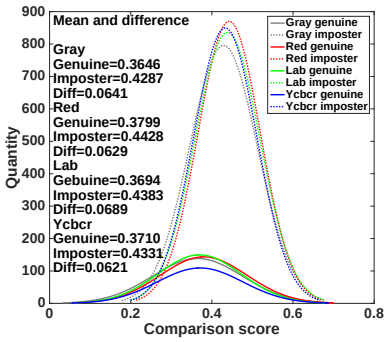


(a) Left eye

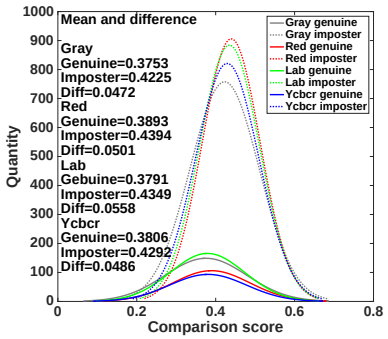


(b) Right eye

Figure 1 – *Comparison score and the difference of their mean values from left and right eye for LFC.*



(a) Left eye



(b) Right eye

Figure 2 – *Comparison score and the difference of their mean values from left and right eye for smartphone.*

employed for classification of iris templates, and two templates were found to match if a test of statistical independence was failed.

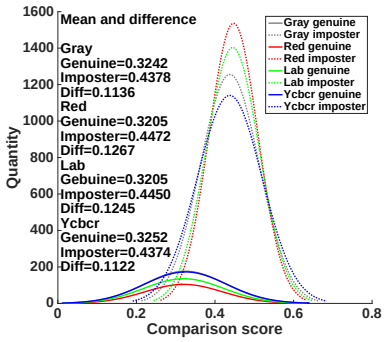
3.4 Methods for the evaluation the performance of iris recognition system

The histograms of comparison scores are obtained from the genuine (comparison between samples from the same subject) and imposter (comparison between samples from different subjects) comparisons for all image samples. In general, high quality biometric samples could generate relatively 'good' genuine comparison scores (in our case, a score closer to 0 the more similar the two iris samples), which are well separated from imposter comparison scores [9]. Equal Error Rate (EER) (when FMR and False FNMR are equal) is another most commonly used method to evaluate the performance of biometric system. We use it as one of the indicators to represent the performance in this paper.

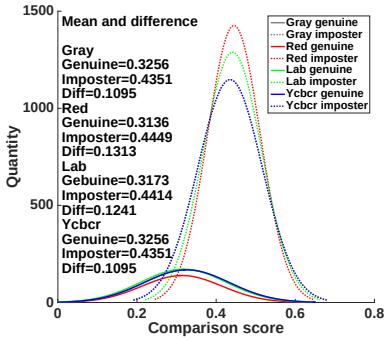
4 Experimental results

4.1 Histogram of the comparison scores and the difference of their mean values

In order to evaluate the performance of iris recognition system on degraded iris images when taking into account color space, we first plot the fitted histogram of the comparison score and the difference of their mean values in Figure 1, 2, and 3. The x-axis represents the score and the y-axis



(a) Left eye



(b) Right eye

Figure 3 – *Comparison score and the difference of their mean values from left and right eye for reflex camera.*

Table 1 – *EER obtained from iris recognition by using different color components*

Color component	Gray	Red	L	Y
LFC				
Left	0.5409	0.5419	0.5413	0.5413
Right	0.5396	0.5417	0.5408	0.5405
Smartphone				
Left	0.5489	0.5664	0.5779	0.5409
Right	0.5258	0.5377	0.5645	0.5154
Reflex				
Left	0.6482	0.6890	0.6639	0.6466
Right	0.6462	0.7001	0.6793	0.6266

represents the quantity of the comparison. The line plots (continued line for genuine comparison and dot line for imposter comparison) is the fitted line for the histogram of the comparison score. The mean values and their difference when using different color components to represent the iris images are also given in the Figures. The gray color represents the comparison score from grayscale iris images, the red color represents the red channel, the green color represents the L channel from CIE Lab color space, and blue color represents the Y channel from YCbCr color space.
From these three Figures we can see that, when the color space changed there is not big difference between the mean of the comparison score. It means that using different components from selected color spaces cannot significantly affect the performance of the iris recognition system. However, the influence of color spaces is different in three cameras. From Figure 1 we can see that, the difference mean value between genuine score and imposter score by using L channel is larger than the other three color components for left eye (red channel for right eye). According to the rule proposed in [9], using L channel representing iris images has better recognition performance than using the other three color components for left eye from LFC (red channel is better for right eye). We can also find out that, the influence is different even for different eyes by using the same camera. From Figure 2 we observe that, for both left and right eyes, the L channel always has better performance because the difference of mean comparison score is greater than the others. In Figure 3, the better color component becomes red channel.
From the analysis above we can conclude that, there is not one color component can increase the performance of iris recognition system better then the others. However, for iris images from different eyes and different cameras, we can use alternative color component to represent iris images in order to obtain a better system performance.

4.2 EER

As mentioned before, we also use EER as an indicator to examine the performance of the iris recognition system when using different color component representing iris images. The lower EER the better system performance. In Table 1 we can discover similar findings than comparison score : by using different color components to represent iris image, the iris recognition performance is slightly affected. For iris images taken by LFC, EER obtained from gray components is lower than the other components (see values in red color in Table 1). For iris images taken by smartphone and reflex camera, the lower EER is always from Y component from YCbCr color space. Unlike the results from comparison score, the better color component is always the same for both eyes from EER. Similar conclusion can be drawn here : there is not a single color component can better increase the iris recognition performance than the others. However, Y channel from YCbCr color space gives lower EER for two cameras.

5 Conclusions

In this paper, we investigate how different color space components affect the recognition system performance on degraded visible wavelength iris samples images. Four color components are selected : gray, red channel from RGB color space, L channel from CIE Lab color space, and Y channel from YCbCr color space. We can conclude from experimental results that, there is not a single color component that can significantly increase the performance of iris recognition system. However, depending on different applications, the iris recognition performance can be improved by using appropriate color component representing visible wavelength iris images.

6 Acknowledgment

This research is supported by the Conseil Regional Basse-Normandie Grant 14P02048 and Research Council of Norway through project number 221073 : HyPerCept - Color and Quality in Higher Dimensions.

Références

[1] Hugo Proenca. Iris recognition : On the segmentation of degraded images acquired in the visible wavelength. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8) :1502–1516, 2010.

[2] Emine Krichen, Mohamed Chenafa, Sonia Garcia-Salicetti, et Bernadette Dorizzi. Color-based iris verification. *Advances in Biometrics*, pages 997–1005, 2007.

[3] Matthew K Monaco. *Color space analysis for iris recognition*. West Virginia University, 2007.

[4] Christopher Boyce, Arun Ross, Matthew Monaco, Lawrence Hornak, et Xin Li. Multispectral iris analysis : A preliminary study51. Dans *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 51–51. IEEE, 2006.

[5] Xinwei Liu, Marius Pedersen, Christophe Charrier, et Patrick Bours. Can no-reference image quality metrics assess visible wavelength iris sample quality ? Dans

IEEE International Conference on Image Processing, 2017. ICIP 2017. 17-20 September, Beijing, China., page 5. Accepted.

[6] Lytro. Lytro, inc. Dans *https ://www.lytro.com/about*. Visited on 16/06/2017.

[7] Xinwei Liu, Marius Pedersen, et Christophe Charrier. Image-based attributes of multi-modality image quality for contactless biometric samples. Dans *Signal Processing and Integrated Networks (SPIN), 2016 3rd International Conference on*, pages 106–111. IEEE, 2016.

[8] Libor Masek et al. Recognition of human iris patterns for biometric identification. 2003.

[9] Patrick Grother et Elham Tabassi. Performance of biometric quality measures. *IEEE transactions on pattern analysis and machine intelligence*, 29(4), 2007.

Etude d'algorithmes d'authentification pour petits capteurs d'empreinte digitale

Mathilde Bourjot Régis Perrier Jean-François Mainguet

CEA Leti - Grenoble *
Département Systèmes
{prenom.nom}@cea.fr

Résumé

Les nouveaux capteurs d'empreintes digitales intégrés dans les smartphones ont déclenché une petite révolution dans le domaine de la biométrie. Ils permettent aux utilisateurs de s'authentifier sans avoir à retenir de code, tout en garantissant une sécurisation des données annoncée comme très bonne par les fabricants. Cependant, la taille de ces capteurs de l'ordre de quelques millimètres carrés proscrit l'usage d'algorithmes d'authentification à minuties. A notre connaissance, peu d'études se sont penchées sur ce problème, malgré l'existence de nombreux algorithmes de reconnaissance de formes et de solutions commerciales supposées robustes. Cette étude cherche à apporter une première réponse en analysant les performances de trois algorithmes pour cette application.

Mots clefs

empreinte digitale, petits capteurs, authentification, descripteurs, points saillants, corrélation.

1 Introduction

L'intérêt des empreintes digitales pour l'identification humaine n'est plus à démontrer depuis longtemps. Si la police les utilise depuis des décennies, leur usage dans les produits commerciaux tels que les ordinateurs et les PDA (*Personal Digital Assistant*) a commencé vers 2004 où le nombre de capteurs vendus se comptaient en millions.

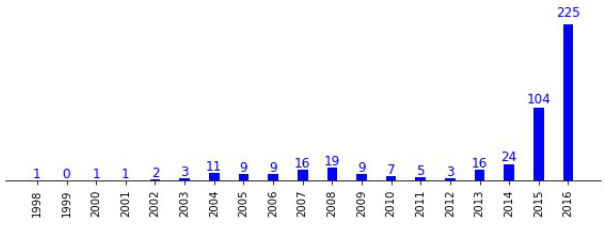


Figure 1 – *Nouveaux modèles de téléphone portable avec un capteur d'empreinte (source : mainguet.org).*

Les smartphones ont emballé le mouvement en 2013 lorsqu'Apple a intégré la reconnaissance d'empreinte dans son iPhone. Aujourd'hui la production de capteur atteint le milliard par an.
Si les applications gouvernementales s'accommodent de grands capteurs optiques, ce n'est pas le cas des smartphones où ces capteurs pour des raisons de coût et d'intégration sont de plus en plus petits, ce qui n'est pas sans poser de problèmes. En effet, l'écart de taille entre les capteurs gouvernementaux de plusieurs centimètres carrés enregistrant le doigt entier comparée à celle des capteurs intégrés aux smartphones des quelques millimètres carrés [1] pose question quant à la fiablitité de ces derniers.
Les applications gouvernementales utilisent systématiquement des algorithmes à minuties ; c'est un standard international. Il a déjà été démontré que ces algorithmes avaient une limite sur le recouvrement minimal entre deux images se situant aux alentours de 7mm pour un capteur à balayage [1]. Une étude plus récente [2] démontre également que plus la surface du capteur est importante, meilleurs sont les taux de vrai refusés (FRR) à un taux de faux accepté (FAR) fixe de 1 pour 10000 : le FRR passe de l'ordre de 1% pour un grand capteur à 20% pour un capteur de 8x8mm^2.

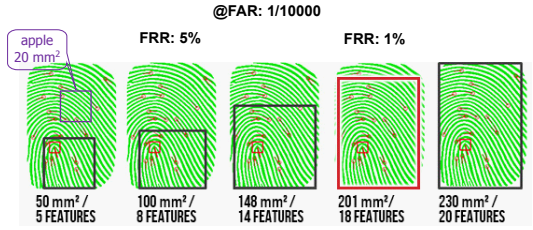


Figure 2 – *Taux de reconnaissance suivant la taille du capteur pour un algorithme à minutie [2].*

Cependant, ces études sont fondées sur des algorithmes « classiques » dits à minuties. Ces dernières sont des caractéristiques d'une empreinte digitale dites de niveau 2 ; le niveau 1 étant la forme générale de l'empreinte. Etant donnée les taux de reconnaissance annoncés, et a priori jamais testés ou publiés, par les vendeurs de smartphones [2], il parait certain que les algorithmes utilisés pour des petits capteurs utilisent d'autres caractéristiques, comme celles dites de niveaux 3 (pores, variations d'épaisseur des lignes, ...).
Dans cet article, nous comparons trois algorithmes extraits de la littérature en vision par ordinateur pour cette application d'authentification d'empreinte digitale acquise par un petit capteur : une méthode dite directe et deux approches

*Cette étude a été financée par un projet Carnot interne au CEA Leti

1. l'iPhone5 avait un capteur de 4.5x4.5mm^2
2. par exemple Apple annonce un FAR de 1/50000 sans FRR fixé

par descripteurs. A notre connaissance, peu de travaux ce sont intéréssés à l'utilisation de méthodes alternatives aux minuties pour la reconnaissance d'empreinte digitale, à des exceptions près comme [3, 4] qui utilisent des bancs de filtres de Gabor et dont les résultats s'approchant des méthodes à minuties, et très récemment [5] qui effectue une comparaison de nombreuses méthodes à descripteurs sur des images d'empreintes de mauvaise qualité et dans laquelle SIFT se démarque. Les méthodes à minuties ont largement satisfait la demande en algorithmes robustes et rapides depuis les années 90, mais l'introduction de petits capteurs a changé la donne. Plusieurs solutions industrielles existent sans qu'il y ait eu d'analyse précise dans la littérature sur les moyens de parvenir à un algorithme de reconnaissance satisfaisant dans ce contexte. Notre étude cherche a apporter une première réponse dans ce sens à l'aide d'algorithmes classiques et eprouvés.

2 Vocabulaire en biométrie

Les systèmes biométriques ont deux phases [6] :

- Une première, l'**enrôlement**, capture les caractéristiques biologiques de l'utilisateur, ici l'image de l'empreinte digitale. Un traitement réduit l'image initiale à des caractéristiques qui sont enregistrées en mémoire.
- Une seconde, la reconnaissance ou l'**authentification**, consiste à capturer une image **candidate**, qui est comparée à l'enregistrement réalisé lors de l'enrôlement. Un algorithme de comparaison retournera un score de similarité, et décidera si l'image candidate est suffisamment proche pour retourner un résultat positif, qui dépendra d'un seuil pré-réglé.

Dans le cas des petits capteurs, l'utilisateur doit présenter son doigt de manière répétitive pour enregistrer une surface de doigt suffisante afin d'augmenter les chances plus tard d'en reconnaitre une portion. Il est logiquement nécessaire de présenter la même portion de peau à l'enrôlement et à la reconnaissance pour reconnaitre quelqu'un.
On définit habituellement deux taux :

- **FAR = faux accepté** ; il s'agit pour celui-ci d'un problème de sécurité. Il doit être de l'ordre d'au moins 1 pour 10000 pour un système commercial, alors que la police requiert des taux de 1 pour 1 million pour exécuter des recherches dans des bases de données de millions de personnes.
- **FRR = vrai refusé** ; c'est généralement ce taux que les utilisateurs expérimentent *de facto*, il correspond à un problème de commodité. Il faut qu'il soit suffisamment faible pour ne pas rejeter le système, et est de l'ordre de 1% pour un bon système.

Ces deux taux sont liés : si on diminue le taux de vrai refusé pour que le système soit pratique, alors on augmente forcément la possibilité pour un imposteur d'entrer dans celui-ci. Ils sont couramment représentés sous forme de courbe ROC (*Receiver Operating Characteristic*) comme nous le verrons dans la partie résultat.

3 Calcul du score de similarité

Nous présentons ici les trois méthodes permettant de calculer un score de similarité entre deux images enrôlée et candidate, respectivement notées E et C par la suite. Ces méthodes sont extraites de l'état de l'art et servirons de référence en terme de performance de reconnaissance dans notre contexte de capteur biométrique de petite surface.
La première utilise globalement l'intensité des pixels des images pour calculer un score ; elle fait partie de la classe des méthodes appelées directes [7, 8]. Les deux suivantes utilisent des descripteurs pour mettre en correspondance les images, elles sont aujourd'hui largement plus répandues que les premières [9, 10].
Nous considerons que les images sont définies sur une grille régulière discrète de taille $(m\text{x}n)$ pixels de sorte que $E(x, y)$ corresponde à l'intensité à valeur dans $\mathbb{R}$ de l'image E au pixel de position $\mathbf{p} = [x, y]^T$.

3.1 Approche directe

La première méthode utilise la corrélation croisée centrée et normalisée pour calculer une surface de correlation $\gamma(u, v)$ entre les deux images par translation de l'une par rapport à l'autre :

$$\gamma(u,v) = \frac{\sum_{x,y\in\mathcal{S}}(E(x,y)-\mu^E)(C(x-u,y-v)-\mu^C)}{\sigma^E\sigma^C} \tag{1}$$

pour $u \in [-n+1, n-1]$ et $v \in [-m+1, m-1]$. Le support $\mathcal{S}$ contient l'ensemble des coordonnées des pixels $\{(x, y)\}$ tel que $x - u \in [1, n]$ et $y - v \in [1, m]$; il varie donc selon u et v. Les paramètres statistiques sont recalculés pour chaque valeur de u et v avec :

$$\mu^E = \frac{\sum_{x,y\in\mathcal{S}} E(x,y)}{\text{Card}(\mathcal{S})} \text{ et } \mu^C = \frac{\sum_{x,y\in\mathcal{S}} C(x-u,y-v)}{\text{Card}(\mathcal{S})},$$

ainsi que :

$$\sigma^E = \sqrt{\textstyle\sum_{x,y\in\mathcal{S}} \left(E(x,y)-\mu^E\right)^2},$$
$$\sigma^C = \sqrt{\textstyle\sum_{x,y\in\mathcal{S}} \left(C(x-u,y-v)-\mu^C\right)^2}.$$

Le score de similarité final retenu est le maximum de la surface de correlation :

$$s_{\text{cor}} = \max \gamma(u, v) \tag{2}$$

On peut relever au moins deux avantages à cette méthode : elle est invariante à un changement de luminosité global dans les images et il existe une implémentation rapide de celle-ci [11] ; ces deux critères l'ont souvent avantagée en pratique au dépend de métriques plus robustes [12]. En revanche, elle n'est pas invariante aux rotations et ses performances se dégradent suivant l'amplitude des déformations géométriques locales entre les deux images.
Pour compenser sa sensibilité aux rotations dans cette étude, nous calculons plusieurs scores de corrélation entre l'image E et l'image C tournée par pas de 5 degrés entre -30 et 30 degrés, et retenons le score maximal.

3.2 Approches par descripteurs

L'utilisation de descripteurs a supplanté depuis de nombreuses années les approches directes sur la question du recalage d'image [8], en plus d'étendre ses compétences dans des domaines comme la reconnaissance de forme. Le schéma d'utilisation de ceux-ci est assez commun selon les applications visées [7], nous résumons ici les étapes pertinentes pour notre sujet :

1. **identification de points saillants** : L'objectif est de rechercher dans les deux images E et C des coordonnées de pixels, $\mathbf{p}^E$ et $\mathbf{p}^C$ respectivement, tels que leur mise en correspondance soit facilitée par la suite. La littérature est abondante sur le sujet [13, 9, 10], mais il est admis à minima qu'un point saillant doit présenter des variations de contraste autour de lui ; autrement dit les dérivées spatiales de l'image dans l'entourage de ce point ne doivent pas être nulles. A l'issue de cette étape, nous obtenons les ensembles $\{\mathbf{p}^E\}$ et $\{\mathbf{p}^C\}$ pour les deux images, non nécessairement de même dimension.
2. **calcul de descripteurs** : Toujours dans l'optique de la mise en correspondance, le descripteur cherche à caractériser au mieux la région autour du point saillant. Il doit être discrimant et compact dans la quantité d'information qu'il contient pour un appariement efficace en terme de robustesse et vitesse de calcul. Une nouvelle fois les travaux sont nombreux sur ce sujet, exploitant la répartition spatiale des points pour capturer une forme [14], des histogrammes sur la distribution des gradients de l'image [9], ou des patches orientés [7] ; une vue plus exhaustive est donnée par [8]. Après cette étape, chaque point $\mathbf{p}_i$ est donc associé à un descripteur $\mathcal{D}_i$ qui peut prendre la forme d'un vecteur ou d'une matrice d'information variée (histogramme, patch, ...).
3. **mise en correspondance** : cette étape recherche les liaisons une à une entre les points $\{\mathbf{p}^E\}$ et $\{\mathbf{p}^C\}$ par comparaison de leurs descripteurs respectifs. La stratégie la plus simple consiste à rechercher pour chaque point d'indice i de E celui qui minimise la fonction distance h de son descripteur avec ceux de C :

$$\hat{k} = \underset{k}{\operatorname{argmin}}\, h(\mathcal{D}_i^E, \mathcal{D}_k^C), \tag{3}$$

où $\hat{k}$ est l'indice de l'élément de $\{\mathcal{D}^C\}$ choisi pour la mise en correspondance avec $\mathcal{D}_i^E$. La fonction h peut par exemple être une distance euclidienne ou de Hamming suivant la caractéristique du descripteur choisi. Pour obtenir une bijection entre les ensembles de points, la même procédure peut être réalisée en inversant les rôles des images, ainsi seuls les appariements qui satisfont le critère de coût minimum de façon symétrique sont retenus ; c'est l'approche qui est utilisée dans ce travail. Nous obtenons alors un nouvel ensemble de points liés entre les deux images $\{\bar{\mathbf{p}}^E, \bar{\mathbf{p}}^C\}$. Notons qu'il existe d'autres stratégies plus élaborées tel que l'algorithme hongrois pour trouver un couplage de poids minimum dans un graphe biparti [14], ou encore des méthodes de tri des descripteurs dans des arbres ou table de hash pour accélérer leur mise en correspondance [9, 7]. Il est également possible de considérer deux minimums plutôt qu'un pour améliorer la robustesse de cette étape par un test du ratio [9], la motivation étant de déterminer à quel point le choix d'une mise en correspondance serait meilleure que la suivante dans la liste.
Afin de satisfaire une contrainte géométrique liée à notre problème selon laquelle le doigt peut subir une rotation et une translation dans le plan du capteur entre deux acquisitions, une étape supplémentaire de minimisation de fonction permet d'exclure les appariements aberrants :

$$\hat{\theta}, \hat{\mathbf{t}} = \underset{\theta,\mathbf{t}}{\operatorname{argmin}} \sum_i \rho\big(\bar{\mathbf{p}}_i^E - R(\theta)(\bar{\mathbf{p}}_i^C + \mathbf{t})\big) \tag{4}$$

où $R(\theta)$ désigne la matrice de rotation dans le plan image et ρ est une fonction permettant un calcul robuste de $\hat{\theta}$ et de $\hat{\mathbf{t}}$. Ces derniers ne sont pas directement d'intérêt, ils ne servent ici que de contrainte dans la détermination d'un nouvel ensemble de points $\{\bar{\bar{\mathbf{p}}}^E, \bar{\bar{\mathbf{p}}}^C\}$ qui satisfont chacun :

$$\|\bar{\bar{\mathbf{p}}}_i^E - R(\theta)(\bar{\bar{\mathbf{p}}}_i^C + \mathbf{t})\| < \epsilon, \tag{5}$$

pour chaque indice i des éléments de l'ensemble de points, où $\|.\|$ désigne la norme euclidienne et ϵ est un scalaire de valeur suffisamment petite. Ceci nous permet de définir le score de similarité final entre E et C :

$$s_{\text{desc}} = \text{Card}(\{\bar{\bar{\mathbf{p}}}^E, \bar{\bar{\mathbf{p}}}^C\}). \tag{6}$$

Littéralement le score correspond au nombre de bonnes mises en correspondance entre les deux images à l'issue de toutes les étapes précédentes. La fonction ρ peut prendre des formes multiples, comme appartenir à la classe des M-estimateurs. En pratique nous utilisons l'algorithme de RANSAC [15] pour sa bonne adéquation avec le problème à résoudre.

Il faut préciser que la mise en correspondance proposée ici n'est pas optimale en terme de vitesse de calcul ; ce n'est pas l'objectif de cette étude qui se concentre sur les 2 premières étapes. Par expérience, nous avons constaté que RANSAC rendait l'algorithme général insensible au choix d'une méthode d'appariement de descripteurs plus élaborée que l'approche naïve et exhaustive que nous utilisons. Cependant il serait intéressant de chercher à s'affranchir de RANSAC par un meilleur appariement de descripteurs au départ. Nous pouvons maintenant détailler les différences entre les deux approches considérées dans ce papier.

SIFT. Cet algorithme qui a pour acronyme *Scale Invariant Feature Transform* a été très populaire depuis le début des années 2000 [9] et reste une référence incontournable en vision par ordinateur. Il optimise toutes les étapes précédemment décrites pour aboutir à une détection de points saillants et une mise en correspondance ayant pour caractéristiques principales :

- une grande robustesse aux transformations affines et aux changements d'échelle entre les images, ainsi qu'aux changements de luminosité (étape 1) ;
- une efficacité dans le temps de calcul par la construction de descripteurs compacts et discriminants (étape 2), et par leur indexation dans des arbres pour faciliter leur mise en correspondance (étape 3).

En pratique nous avons fait le choix de n'utiliser dans ce travail que la partie identification des points saillants et calcul du descripteur de la méthode SIFT. Ceci nous permet une certaine maîtrise de la mise en correspondance en conservant le schéma décrit précédemment, et facilite également la comparaison avec la seconde méthode. Les conséquences sont un temps de calcul supérieur dans notre cas, et un nombre de mise en correspondance erronée plus grand qui est corrigé par la suite à l'aide de RANSAC.
Les points saillants de SIFT, également définis comme points clés, sont détectés comme des extremas de la fonction DoG (*Difference of Gaussians*) appliquée à l'image pour différents facteurs d'échelle. Intuitivement, les contours de l'image sont estimés à différentes échelles par filtrage, et les extremas locaux le long de ces contours et entre échelles adjacentes sont sélectionnés. Après suppression des points de faible contraste et de faible courbure le long du contour, un raffinement subpixelique permet une localisation plus précise de ceux-ci. Enfin à chaque point est associé une orientation, déterminée par un calcul de gradient dans un environnement local qui dépend de l'échelle à laquelle celui-ci a été détecté ; ceci finalise l'étape 1.
Le descripteur SIFT cherche à caractériser l'orientation des gradients dans une fenêtre (16x16) pixels autour de chaque point clé par un classement ingénieux de celles-ci dans des histogrammes[9]. La robustesse du descripteur est assurée notamment par pondération des orientations estimées par la magnitude du gradient servant à leur calcul ainsi que par leur distance au point saillant, mais également par un calcul d'orientation qui est relatif à l'orientation même du point clé, et aussi par un seuillage des histogrammes pour améliorer l'indépendance aux conditions d'illumination. Le descripteur SIFT est résumé dans un vecteur contenant les valeurs des histogrammes, la fonction de distance h entre deux descripteurs est une distance euclidienne :

$$h(\mathcal{D}_i, \mathcal{D}_k) = \|\mathcal{D}_i - \mathcal{D}_k\|. \tag{7}$$

Ceci complète les étapes 2 et 3 dans le cas de SIFT.

Harris-SSD. Nous désignons par Harris-SSD l'utilisation conjointe de points saillants extraits par la méthode d'Harris [13] et de descripteurs dont la comparaison s'effectue par calcul d'une somme des différences entre éléments au carré (SSD : *Sum Square Differences*).
La détection de points saillants par l'algorithme d'Harris et Stephen s'appuie sur un développement de Taylor de la différence des intensités au carré entre l'image concernée et une version de celle-ci légèrement translatée. Cette expression peut être astucieusement réécrite sous une forme matricielle localement pour chaque pixel :

$$H_{x,y} = \nabla_{\sigma_H} E(x,y) \nabla_{\sigma_H} E(x,y)^T, \tag{8}$$

où $H_{x,y}$ désigne la matrice de Harris de taille (2x2) au pixel (x, y) de l'image E et $\nabla_{\sigma_H} E(x, y)$ le gradient selon x et y de E lissée par une fenêtre gaussienne de paramètre σ_H fixé à 1 dans notre étude. Une analyse en valeur propre de $H_{x,y}$ permet de déterminer si le pixel appartient à une région uniforme, à un contour, où encore à un coin. Ce dernier ayant pour avantage de lever toute ambiguïté spatiale, il devient privilégié dans l'extraction des points caractéristiques après seuillage de la fonction d'Harris sur l'image. Pour chaque point saillants, nous ajoutons un calcul d'orientation θ à partir de la valeur locale du gradient estimée pour l'image lissée :

$$\theta(x,y) = \arctan2\left(\nabla^y_{\sigma_\theta} E(x,y), \nabla^x_{\sigma_\theta} E(x,y)\right), \tag{9}$$

où le paramètre de lissage σ_θ est fixé à 4. Ainsi chaque point saillant est caractérisé par une position $\mathbf{p}$ et une orientation θ ; ceci complète l'étape 1.
Le descripteur utilisé ici correspond à un patch orienté extrait autour du point d'intérêt dont nous centrons et normalisons les intensités pour limiter sa sensibilité vis a vis des conditions de peau. Soit $I_{\mathbf{p},\theta}$ le patch de taille ($2w$+1x$2w$+1) extrait autour du point $\mathbf{p}$ d'orientation θ, nous avons :

$$\mathcal{D} = \frac{I_{\mathbf{p},\theta} - \mu^{I_{\mathbf{p},\theta}}}{\sigma^{I_{\mathbf{p},\theta}}}. \tag{10}$$

Par convention, le patch est extrait par interpolation relativement à θ de sorte à ce que les plus forts gradients soit détectés selon l'axe x de $I_{\mathbf{p},\theta}$, de même nous avons choisi $w = 7$ notre étude. Par extension, $\mathcal{D}$ est donc une matrice et la fonction de distance utilisée à un autre descripteur est :

$$h(\mathcal{D}_i, \mathcal{D}_k) = \sum_{x,y} \left(\mathcal{D}_i(x,y) - \mathcal{D}_k(x,y)\right)^2. \tag{11}$$

Ici h est également connue comme une fonction d'erreur du type SSD (*Sum of Square Differences*). Ceci finalise l'étape 2 et 3 pour la méthode Harris-SSD ; nous adoptons ici une version simplifiée de la méthode MOPS (*Multi-Scale Oriented Patches*) utilisée dans [7] pour la construction de mosaïque d'images.

4 Protocole d'évaluation

Nous détaillons dans cette partie le type des données utilisées pour comparer les méthodes présentées précédemment ainsi que les critères évalués. Les algorithmes ont été implémentés en python et utilisent en grande partie les fonctions de la bibliothèque scikit-image [16].

4.1 Données

Nous avons utilisé la base de données biométrique SDUMLA-HMT en libre accès sur internet [17]. Elle comprend des images acquises par 5 capteurs différents sur 106

personnes. Pour chaque personne, 6 doigts sont enregistrés : le pouce, l'index et le majeur de chacune des mains ; chacun des doigts est acquis 8 fois. Nos expérimentations sont faites sur les données du capteur capacitif FT-2BU [3]. De cette base, nous avons sélectionné 30 personnes aléatoirement, totalisant 1440 images de taille (152x200) pixels. Il s'agit de données labélisées puisque les acquisitions sont rangées par dossier selon les individus avec un sous dossier par doigt. En revanche, nous n'avons pas de connaissance sur la quantité de translation et de rotation d'un même doigt entre deux acquisitions.

4.2 Simulation d'un petit capteur

Pour simuler le processus d'acquisition d'un petit capteur, nous avons découpé une image de taille 70x70 pixels dans chaque acquisition. Ce format correspond à une surface de (5x5)mm pour un capteur de 350dpi, soit de l'ordre de grandeur de ceux du marché des smartphones.

Figure 3 – *Extraction sur une acquisition à gauche (contours rouges), et exemples d'images extraites à droite.*

Pour extraire une image centrée sur l'empreinte, nous utilisons un algorithme K-means qui sépare le fond de l'image de celle de l'empreinte. On détermine ensuite le centre comme le point situé au milieu du rectangle englobant l'empreinte segmentée. Enfin, on extrait la fenêtre de taille (70x70) centrée en ce point ; un exemple est en figure 3.

4.3 Séparation des données

Les images sont ensuite séparées en 2 catégories : images d'enrôlement qui correspondraient aux données enregistrées par l'utilisateur sur son smartphone lors de la première utilisation du système biométrique, et images candidates qui représenteraient des enregistrements authentiques ou frauduleux faits par le capteur lors de son utilisation. Pour chaque doigt, on sélectionne aléatoirement un nombre n d'acquisitions que l'on utilise comme données d'enrôlement, le reste devenant les données candidates.

4.4 Critères évalués

Une table de score est calculée en comparant chaque image candidate avec l'ensemble des images d'enrôlement de chaque doigt, nous obtenons un score par couple {image candidate, doigt enrôlé} qui est fonction des scores individuels de l'image candidate avec chacune des images d'enrôlement du doigt ; dans les résultats qui suivent nous utilisons le maximum de ceux-ci. La table de score permet de calculer les taux de FRR et FAR pour différents seuils. Les trois algorithmes sont comparés sur la base des courbes ROC obtenues sur les données extraites.

3. capteur de Miaxis Biometrics qui utilise une technologie de Fingerprint Cards, a priori un clone du capteur FPC10xx à 363dpi

5 Résultats

La figure 4 résume les performances en reconnaissance des trois algorithmes en fonction du nombre d'images retenues au cours de l'enrôlement ; plus la courbe ROC s'approche du point $(0,10^{-3})$, meilleur est l'algorithme.

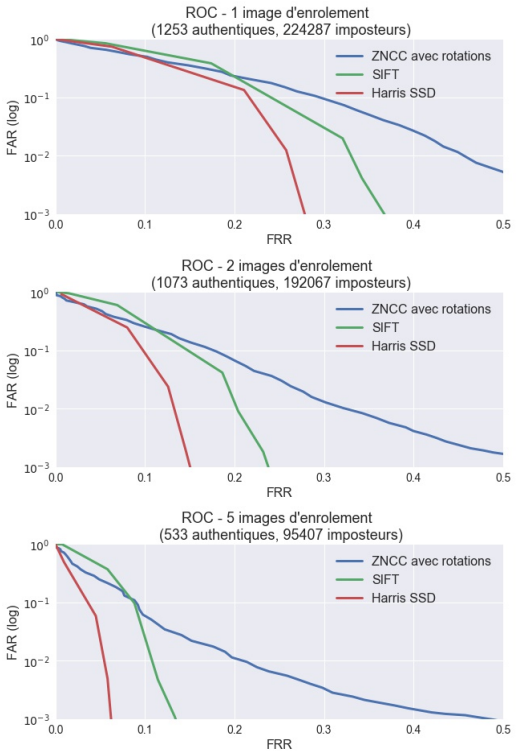


Figure 4 – *Courbles ROC pour les trois algorithmes selon le nombre d'images d'enrôlement.*

On constate assez logiquement qu'une augmentation du nombre d'images enrôlées permet de réduire nettement le FRR à FAR fixe, le FRR est presque divisé par trois pour les méthodes SIFT et Harris SSD en passant d'une à cinq images d'enrôlement. Il faut noter que ce nombre est bien supérieur à dix pendant la procédure d'enregistrement du doigt réalisée par un smartphone. La quantité des données de la base ne nous permet pas de reproduire un tel enrôlement, mais il est clairement déterminant pour augmenter les performances d'un algorithme d'authentification.
On remarque également que la méthode Harris SSD se démarque de SIFT, ceci contredit certains résultats de la littérature dont ceux de [5]. Une première explication est que SIFT est un algorithme performant dans des tâches génériques avec de nombreux degrées de liberté comme l'invariance au facteur d'échelle. Celle-ci n'est pas requise dans notre étude, et n'est d'ailleurs pas prise en compte dans Harris SSD. Ensuite, SIFT par conception cherche des ex-

trema le long des lignes appartenant à l'image, or ce sont des zones à forte ambiguité dans le cas des empreintes digitales. La figure 5 vient renforcer cette observation en comparant les points saillants détectés par les deux méthodes sur une même acquisition. On remarque que le détecteur d'Harris se concentre exclusivement sur les contours là où SIFT détecte des fonds de vallées ou des sommets de crêtes qui ont a priori moins de chances d'être discrimants.

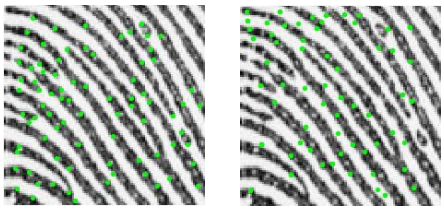


Figure 5 – *Points saillants de Harris et SIFT en verts.*

La seconde explication est liée à la nature des données. Les empreintes sont constituées d'une texture très répétitive avec une alternance de vallées et de crêtes. Un descripteur qui ne s'appuie que sur des histogrammes d'orientation des gradients peut perdre des informations locales sur la forme de l'empreinte autour du point saillant. La figure 6 montre des exemples de patchs extraits par Harris SSD qui semblent difficiles à résumer à l'aide du descripteur SIFT.

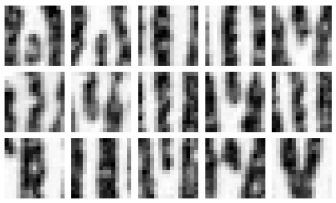


Figure 6 – *Exemples de patchs (15x15) pixels extraits autour des points saillants de Harris en figure 5.*

6 Perspectives

Les résultats obtenus sont préliminaires, la robustesse des algorithmes doit être testée sur plus de données avec des variations dans les conditions de peau plus importantes. Certains paramètres ont été fixés empiriquement par observation des résultats, comme la taille des fenêtres ou les coefficients de lissage pour Harris SSD, ils nécessiteraient une analyse de sensibilité plus poussée.
Il reste des marges de manoeuvre importantes pour accélérer l'appariement et probablement se passer de l'étape de RANSAC. Enfin la similarité entre les patchs retenus par la méthode d'Harris SSD sur l'ensemble des données nous suggère que des algorithmes d'apprentissage d'une base de représentation de ceux-ci pourraient ouvrir d'autres perspectives dans la façon de calculer le score de similarité.

Références

[1] J.F. Mainguet, W. Gong, et A. Wang. Reducing silicon fingerprint sensor area. Dans *International Conference on Biometric Authentication*, volume 3072, pages 301–308, Hong-Kong, 2004.

[2] R. Sanchez-Reillo. The Madrid Report, *Next Biometrics*, june 2017.

[3] C. Lee et S. Wang. Fingerprint feature extraction using gabor filters. *Electronics Letters*, 35(4) :288–290, Feb 1999.

[4] A. Jain, S. Prabhakar, L. Hong, et S. Pankanti. Filterbank-based fingerprint matching. *Transactions on Image Processing*, 9(5) :846–859, 2000.

[5] A. Karlsson et J. Hagel. Fingerprint matching-hard cases, *Faculty of engineering, Lund University*, 2016.

[6] A. Jain, A. Ross, et S. Prabhakar. An introduction to biometric recognition. *Transactions on Circuits and Systems for Video Technology*, 14(1) :4–20, Jan 2004.

[7] M. Brown, R. Szeliski, et S. Winder. Multi-image matching using multi-scale oriented patches. Dans *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 510–517, Jun 2005.

[8] R. Szeliski. *Computer vision : algorithms and applications*. Springer Science and Business Media, 2010.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[10] K. Mikolajczyk et C. Schmid. A performance evaluation of local descriptors. *Transactions on Pattern Analysis and Machine Intelligence*, 27(10) :1615–1630, 2005.

[11] J. P. Lewis. Fast normalized cross-correlation. *Vision interface*, 10(1) :120–123, 1995.

[12] S. Chambon et A. Crouzil. Mesures de corrélation robustes aux occultations. Dans *ORASIS-Journées jeunes chercheurs en vision par ordinateur*, Gerardmer, 2003.

[13] C. Harris et M. Stephens. A combined corner and edge detector. Dans *Fourth Alvey Vision Conference*, pages 147–151, 1988.

[14] S. Belongie, J. Malik, et J. Puzicha. Shape matching and object recognition using shape contexts. *Transactions on Pattern Analysis and Machine Intelligence*, 24(4) :509–522, 2002.

[15] M. Fischler et Ro. Bolles. Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6) :381–395, Juin 1981.

[16] S. van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, et the scikit-image contributors. scikit-image : image processing in Python. *PeerJ*, 2 :e453, 6 2014.

[17] Y. Yin, L. Liu, et X. Sun. Sdumla-hmt : A multimodal biometric database. Dans *Chinese Conference on Biometric Recognition*, pages 260–268. Springer-Verlag, 2011.

Validation de Métriques de Qualité de Données Biométriques

Z. Yao J.M Lebars C. Charrier C. Rosenberger

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
christophe.rosenberger@ensicaen.fr

Résumé

Le contrôle qualité de données biométriques fonctionne comme un péage pour décider si un échantillon biométrique peut être exploité pour définir le modèle d'un individu. Prendre en compte la qualité des données biométriques à l'enrôlement a un impact important sur la performance ultérieure du système biométrique. Des métriques sont proposées dans cet objectif par les chercheurs, leur pertinence doit être étudiée notamment pour évaluer leur utilité. Dans ce papier, nous proposons un cadre méthodologique permettant d'estimer la pertinence d'une métrique de qualité de données biométriques. La méthode proposée permet de comparer de façon objective deux métriques de qualité. Nous illustrons l'apport de cette méthodologie sur des empreintes digitales.

Mots clefs

Qualité d'une donnée biométrique, validation d'une métrique de qualité, empreinte digitale.

1 Introduction

La biométrie a pour objectif d'identifier ou de vérifier l'identité d'un individu à partir de caractéristiques morphologiques (visage, empreinte digitale, ...) ou comportementales (dynamique de frappe au clavier, voix, ...). A cette fin, le système biométrique exploite la référence biométrique de l'individu construite à partir d'une ou plusieurs captures de la donnée biométrique de l'individu. Afin de garantir les meilleures performances du système biométrique, l'on essaie en général de collecter pendant l'enrôlement la meilleure qualité possible de la donnée biométrique. Ainsi, la photo du visage fournie lors de la création d'un passeport doit suivre plusieurs contraintes [1] pour être utilisée. De même, lors de l'acquisition d'une empreinte digitale pour le passeport, une métrique est utilisée pour garantir un niveau de qualité suffisant [2]. Il s'agit de la métrique NFIQ (NIST Fingerprint Image Quality) proposée par le NIST en 2004.

La qualité d'une donnée biométrique dépend de plusieurs facteurs dont les conditions d'acquisition (environnement, bonne utilisation du capteur, ...), du capteur biométrique (problème de reglage, ...) ou de la donnée biométrique (doigt abimé, partie du visage occultée, ...). La figure 1 présente quelques exemples d'empreinte digitale de différentes qualités. Il est clair que la performance d'un algorithme de comparaison d'empreintes digitales sera différente dans tous ces cas.

Figure 1 – *Exemples d'empreintes digitales de qualités différentes (normal, doigt sec, doigt mouillé, plis, rides)*

Dans la dernière décennie, plusieurs travaux de recherche ont porté sur la définition de métriques de qualité de données biométriques pour le visage [3, 4], le réseau veineux [5] et surtout l'empreinte digitale [2, 6, 7]. Plusieurs problématiques demeurent quelque ce soit la modalité biométrique considérée : comment valider la pertinence d'une métrique de qualité d'une donnée biométrique ? comment démontrer objectivement la supériorité d'une métrique de qualité par rapport à une autre ? La principale contribution de ce papier est de proposer un cadre méthodologique visant à quantifier la pertinence d'une métrique de qualité d'une donnée biométrique. L'approche proposée est générique, elle peut être appliquée sur n'importe quelle modalité biométrique. L'autre avantage de la méthode est de pouvoir comparer de façon objective deux métriques de qualité. La méthode permet également d'estimer dans quelle mesure où une métrique est proche d'un jugement optimal, l'indicateur de pertinence est donc borné.

Le papier est organisé comme suit. La section 2 présente les méthodes de l'état de l'art pour analyser la performance d'une métrique de qualité d'une donnée biométrique. La méthode proposée est décrite dans la section 3. La section

4 présente l'application de cette méthodologie pour comparer deux métriques de qualité d'empreintes digitales. La section 5 conclue ce travail et propose différentes perspectives.

2 Travaux antérieurs

Phillips et al. [8] ont mis en évidence l'importance de l'évaluation des sysèmes biométriques. Cependant, certaines études antérieures sur l'évaluation de la qualité de données biométriques ne fournissent pas de protocoles totalement ouverts [9]. Cette méthode largement utilisée (notamment pour la normalisation de métriques) repose principalement l'étude de la corrélation statistique de scores légitimes et la valeur de la métrique des échantillons. Ce type d'approche nécessite une base de données biométriques très conséquente (plusieurs millions d'échantillons) difficilement accessible aux chercheurs.

D'autres études évaluent l'intérêt d'une métrique de qualité vis à vis d'éléments d'évaluation subjective [10]. Ratha et Bolle ont également proposé de mesurer la pertinence d'une métrique de qualité de données biométriques par comparaison statistique (avec le facteur de corrélation de Pearson) avec les valeurs d'autres métriques de qualité [11]. De même, Shen et al. [12] ont proposé d'affecter les empreintes digitales d'une base de données suivant leur qualité évaluée par une métrique et de comparer ces classes avec d'autres générées par d'autres métriques de qualité. Ces approches ne permettent pas de quantifier la relation entre la métrique de qualité et les performances correspondantes. En outre, ces tentatives sont plus ou moins liées aux observations subjectives lors de la validation de métriques de qualité. Cependant, ce type d'opération pourrait être facilement estimée en utilisant un générateur d'empreintes synthétiques tel que SFinGe [13].

Tabassi et al. [14] définissent la qualité de l'échantillon biométrique comme un prédicteur de la performance de reconnaissance en observant que des échantillons biométriques de bonne qualité produisent des scores légitimes relativement élevés et bien séparés des scores des imposteurs. Cependant, la prédiction dépend totalement de la performance de l'algorithme de comparaison utilisé. Grother et al. [9] ont proposé plusieurs approches d'évaluation associées aux niveaux de seuil de décision et de qualité correspondante, y compris la courbe DET et l'approche basée sur le test de Kolmogorov Smirnov (KS).

Ces dernières approches sont intéressantes mais sont difficiles à interpréter (car basé sur un test d'hypothèse), pas vraiement adaptées pour comparer quantitativement deux métriques de qualité. Nous proposons une méthode répondant à ces deux limitations et permettant également d'estimer la pertinence optimale à atteindre pour une base de données biométriques et un algorithme de comparaison.

3 Méthode proposée

Le principe de la méthode proposée de validation d'une métrique de qualité de données biométriques consiste à évaluer sa pertinence sur la tâche d'enrôlement d'un utilisateur (qui correspond à l'usage le plus important des métriques de qualité). Nous supposons avoir une base de d'échantillons biométriques de plusieurs utilisateurs (voir Figure 2) et un algorithme de comparaison de ces échantillons.

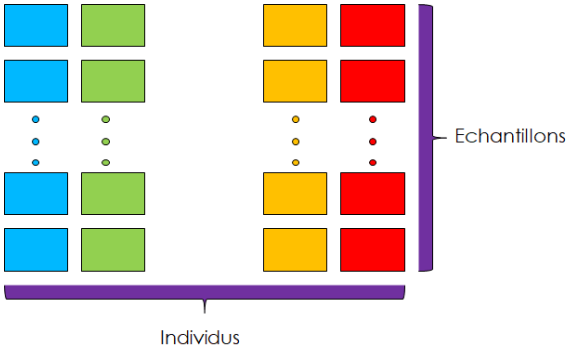


Figure 2 – *Description d'une base de données biométriques.*

Si l'on souhaite évaluer la performance d'un système biométrique sur cette base, il faut appliquer une heuristique pour choisir l'échantillon pour être la référence de l'utilisateur. Une fois ce choix réalisé, nous calculons l'ensemble des scores légitimes de la base en comparant les échantillons avec la référence de chaque utilisateur. On fait de même pour les scores d'imposture en comparant une référence avec les échantillons de tous les autres individus de la base (voir Figure 3). Ainsi, si on une base composée de N individus avec M échantillons par indididu, par ce procédé, on génère $N \times (M-1)$ scores légitimes et $N \times (M-1) \times (N-1)$ scores d'imposture. Ces scores permettent de calculer le taux de faux rejet et de fausse acceptation pour différentes valeurs du seuil de décision (à partir duquel on considère la preuve de vérification comme suffisante). Ceci permet de calculer différentes mesures classiques en biométrie : la courbe DET (évolution du faux rejet en fonction de la fausse acceptation), le taux d'égal erreur (EER) et l'aire en dessous de la courbe DET (AUC).

Revenons à présent sur le choix de l'échantillon comme référence de l'utilisateur, il existe plusieurs possibilités :

- Choix du premier échantillon comme réference de l'individu. C'est une pratique courante considérée comme un choix par défaut,
- Choix de la référence à partir d'une métrique de qualité. Dans ce cas, les échantillons sont évalués par cette métrique et le "meilleur" échantillon (au vu de cette métrique) est sélectionné comme référence de l'utilisateur,
- Choix du meilleur échantillon : En général, une

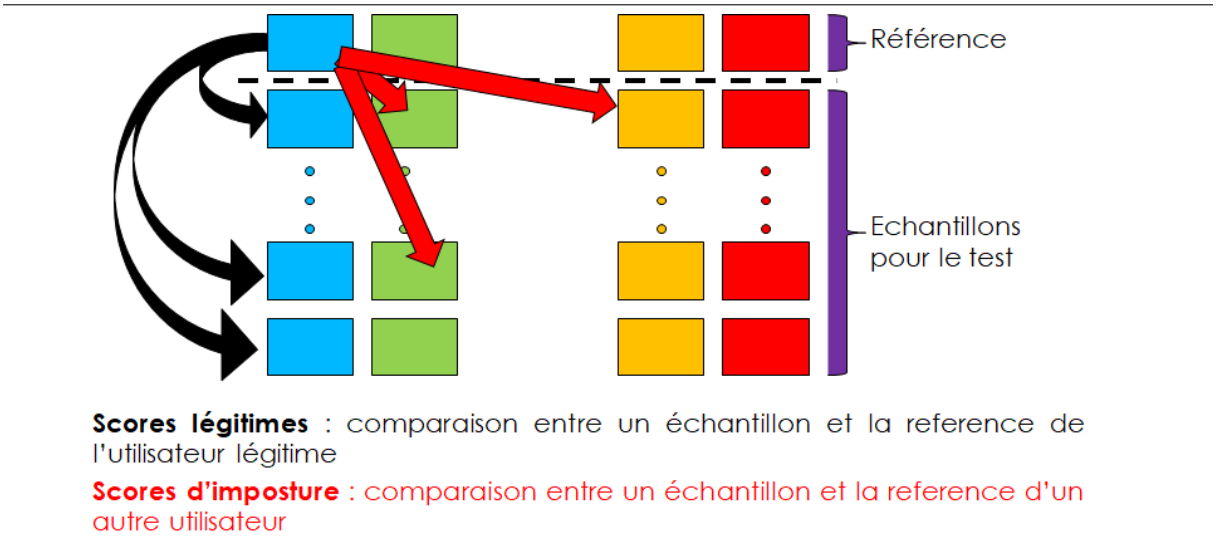


Figure 3 – *Explication du calcul des scores légitimes et d'imposture à partir d'une base de données biométriques.*

base de données biométriques contient un nombre fini et faible d'échantillons M de chaque individu. Il est possible de tester toutes les possibilités, c'est à dire estimer la valeur de l'AUC obtenue pour chaque possibilité d'échantillon comme référence. Il est donc possible de déterminer le choix optimal de l'échantillon de référence au vu de la performance du système biométrique (valeur minimale de l'AUC),

— Choix du plus mauvais échantillon : on reprend la procédure précédente en prenant l'échantillon conduisant à la valeur la plus élevée de l'AUC.

En applicant ces différentes heuristiques, on peut mesurer la performance du système biométrique pour une base de données biométriques et un algorithme de comparaison. Nous considérons dans cette étude la valeur de l'AUC qui est une mesure globale de performance à minimiser. Nous pouvons calculer $AUC_{meilleure}$ la valeur de l'AUC avec le choix optimal de la référence, $AUC_{mauvais}$ la valeur de l'AUC avec le choix de la plus mauvaise référence et $AUC_{metrique}$ la valeur de l'AUC avec le choix de la référence guidée par une métrique de qualité. Nous proposon de calculer la pertinence d'une métrique comme suit :

$$P = 1 - \frac{(AUC_{metrique} - AUC_{meilleure})}{(AUC_{mauvais} - AUC_{meilleure})} \quad (1)$$

La figure 5 présente la performance obtenue sur une base de données biométriques pour différentes heuristiques de choix de la référence. En noir est représentée la courbe DET en prenant l'échantillon le plus mauvais comme référence et en vert le meilleur choix. Cette figure montre que le choix de la référence va conduire à une performance du système entre un AUC de 0.0352 à 0.2338. En utilisant deux métriques, on obtient une performance de 0.0991 (bleu) et l'autre de 0.0788 (rouge). Cette figure donne deux informations, la première est que la métrique 1 (courbe bleu) est moins efficace que la métrique 2 (courbe rouge). En effet, la métrique 2 permet une meilleure performance du système biométrique. Deuxièmement, la métrique 2 a une pertinence de $P = 78\%$ contre $P = 67\%$ pour l'autre métrique. Cela signifie qu'il y a encore des possibilités d'amélioration pour effectuer le meilleur choix de la référence.

4 Application

Nous illustrons dans cette section l'intérêt de la méthode de validation d'une métrique de qualité. Nous présentons le résultat d'analyse de la pertinence d'une métrique de qualité récente d'empreintes digitales [15]. La métrique en question consiste à mesurer le nombre de pixels de bonne qualité dans une image d'empreinte digitale. Un pixel est considéré comme de mauvaise qualité s'il fait partie de l'arrière plan de l'empreinte ou s'il appartient à un bloc dont le gradient calculé sur l'image des crêtes n'est pas uniforme (voir Figure 6).

Nous avons utilisé 5 bases de données biométriques issues de la compétition FVC [16] avec l'algorithme de comparaison Bozorth3 proposé par le NIST [17]. Nous avons appliqué la méthodologie présentée dans la section précédente. Nous avons comparé cette métrique avec NFIQ [2].

Le tableau 5 présente la valeur de l'AUC obtenue sur les 5 bases de données biométriques en réalisant le choix de la référence de chaque individu à partir de la métrique NFIQ et celle du papier Yao et al. 2016. On peut constater que cette seconde métrique permet d'obtenir de meilleures performances sur les 3 dernières bases de données biomé-

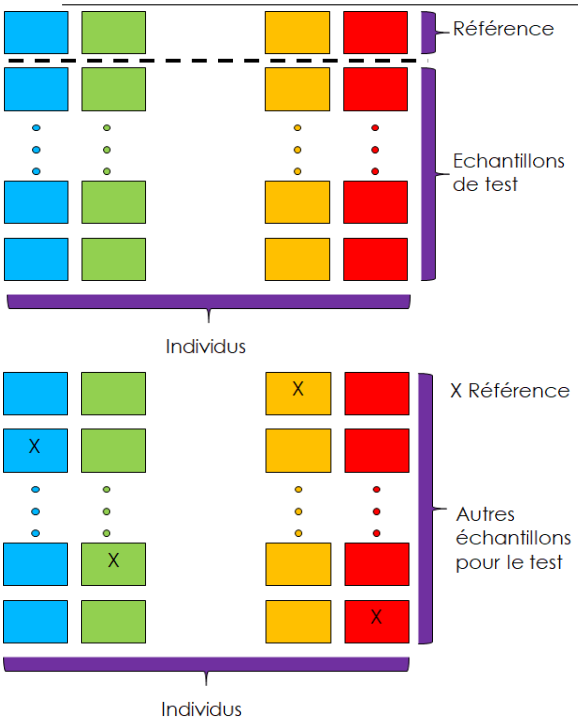


Figure 4 – *Choix de l'échantillon de référence dans une base de données biométriques*

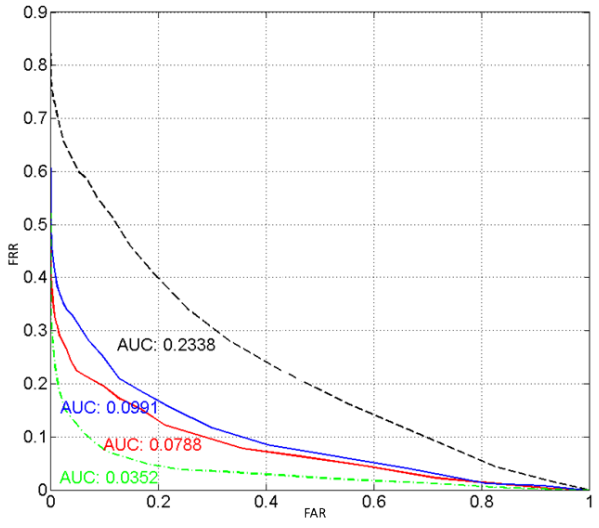


Figure 5 – *Représentation de la performance en fonction du choix de la référence : plus mauvais choix (noir), meilleur choix (vert), choix par la métrique 1 (bleu) et choix par la métrique 2 (rouge).*

triques. Sur les deux premières, NFIQ donne un résultat légèrement meilleur sur des bases plus simples (avec un AUC très proche de 0%). Dans la mesure où plusieurs bases de données biométriques ont été utilisées, on peut conclure que cette métrique propose une amélioration de NFIQ.

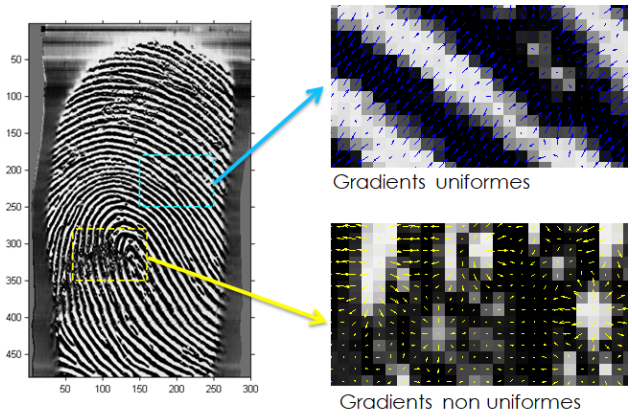


Figure 6 – *Principe de la métrique de qualité d'une empreinte digitale.*

5 Conclusion et perspectives

Nous avons proposé dans ce papier une méthodologie d'analyse de la pertinence d'une métrique de qualité de données biométriques. Le principe est de quantifier la performance du choix de l'échantillon de référence d'un individu. Nous avons pu mettre en évidence le caractère générique de l'approche et l'intérêt opérationnel de pouvoir quantifier la pertinence d'une métrique. L'intérêt de cette méthodologie a été illustrée pour comparer deux métriques de l'état de l'art sur cinq bases d'empreintes digitales.

Les perspectives de cette étude visent à proposer une métrique de qualité d'empreinte digitale par optimisation de la mesure de pertinence.

Références

[1] Matteo Ferrara, Annalisa Franco, et Davide Maltoni. Evaluating systems assessing face-image compliance with icao/iso standards. Dans *European Workshop on Biometrics and Identity Management*, pages 191–199. Springer, 2008.

[2] E Tabassi, CL Wilson, et C Watson. Fingerprint image quality (nfiq). nistir 7151. august 2004, 2011.

[3] Kamal Nasrollahi et Thomas B Moeslund. Face quality assessment system in video sequences. Dans *European Workshop on Biometrics and Identity Management*, pages 10–18. Springer, 2008.

[4] Pankaj Wasnik, Kiran B Raja, Raghavendra Ramachandra, et Christoph Busch. Assessing face image quality for smartphone based face recognition system. Dans *Biometrics and Forensics (IWBF), 2017 5th International Workshop on*, pages 1–6. IEEE, 2017.

[5] Huafeng Qin et Mounim A El Yacoubi. Deep representation for finger-vein image quality assessment.

Métrique	00DB2	02DB2	04DB1	04DB2	04DB3
NFIQ	0.22%	0.11%	2.66%	3.86%	1.89%
Yao et al. 2016	0.10%	0.20%	1.93%	3.24%	1.51%

Tableau 1 – *Valeur de l'AUC obtenue sur 5 bases de données biométriques en réalisant le choix de la référence de chaque individu à partir de la métrique NFIQ et celle du papier Yao et al. 2016.*

Figure 7 – *Exemples d'empreinte digitale issus des 5 bases de données biométriques.*

IEEE Transactions on Circuits and Systems for Video Technology, 2017.

[6] Zhigang Yao, Jean-Marie Le Bars, Christophe Charrier, et Christophe Rosenberger. Fingerprint quality assessment combining blind image quality, texture and minutiae features. Dans *ICISSP 2015*, 2015.

[7] Xinwei Liu, Marius Pedersen, Christophe Charrier, Patrick Bours, et Christoph Busch. The influence of fingerprint image degradations on the performance of biometric system and quality assessment. Dans *Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the*, pages 1–6. IEEE, 2016.

[8] P Jonathon Phillips, Alvin Martin, Charles L Wilson, et Mark Przybocki. An introduction evaluating biometric systems. *Computer*, 33(2) :56–63, 2000.

[9] Patrick Grother et Elham Tabassi. Performance of biometric quality measures. *IEEE transactions on pattern analysis and machine intelligence*, 29(4), 2007.

[10] Tai Pang Chen, Xudong Jiang, et Wei-Yun Yau. Fingerprint image quality analysis. Dans *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 2, pages 1253–1256. IEEE, 2004.

[11] Nalini K Ratha et Ruud Bolle. *Fingerprint image quality estimation*. IBM TJ Watson Research Center, 1999.

[12] LinLin Shen, Alex Kot, et Wai Mun Koo. Quality measures of fingerprint images. Dans *AVBPA*, pages 266–271. Springer, 2001.

[13] Raffaele Cappelli, D Maio, et D Maltoni. Sfinge : an approach to synthetic fingerprint generation. Dans *International Workshop on Biometric Technologies (BT2004)*, pages 147–154, 2004.

[14] Elham Tabassi et Patrick Grother. Fingerprint image quality. Dans *Encyclopedia of Biometrics*, pages 482–490. Springer, 2009.

[15] Z Yao, Christophe Charrier, et Christophe Rosenberger. Pixel pruning for fingerprint quality assessment. Dans *International Biometric Performance Testing Conference (IBPC)*, 2016.

[16] Fingerprint verification competition databases.

[17] Patricia A Flanagan. Nist biometric image software (nbis). 2010.

Authentification multi-biométrique sur mobile respectueuse de la vie privée

A. Ninassi S. Vernois C. Rosenberger

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
{alexandre.ninassi, sylvain.vernois, christophe.rosenberger}@ensicaen.fr

Résumé

Les smartphones sont de plus en plus utilisés pour différents services numériques tels que les réseaux sociaux ou le commerce électronique. L'authentification de l'utilisateur avec des mots de passe sur ces appareils n'est pas conviviale et n'offre pas un niveau de sécurité élevé de l'utilisateur. La biométrie devient une solution populaire pour atteindre cet objectif avec notamment l'intégration de capteurs d'empreintes digitales dans les smartphones. Dans cet article, nous proposons un nouveau protocole combinant l'empreinte digitale et une biométrie comportementale pour améliorer la sécurité de l'authentification des utilisateurs tout en préservant la facilité d'usage et le respect de la vie privée. Le comportement lors de la saisie d'une authentification d'un motif sur l'écran tactile du smartphone est considéré comme une solution rapide et simple d'usage. Nous pensons que la solution proposée offre de nombreux avantages en termes de sécurité, d'usage et de respect de la vie privée. Nous montrons au travers des résultats expérimentaux l'efficacité de la méthode proposée même en cas d'attaque.

Mots clefs

Authentification, biométrie révocable, multi-biométrie.

1 Introduction

Un sondage récent en 2016 [1] a montré que plus de 50% des utilisateurs de smartphones l'utilisent immédiatement après leur réveil. À mesure qu'un smartphone intègre de plus en plus d'informations personnelles (contacts, contenu du courrier, média ...) et est utilisé comme périphérique privilégié pour accéder à des services distants, une authentification forte de l'utilisateur devient nécessaire. L'authentification par code PIN est une solution simple, néanmoins, elle ne constitue pas une preuve d'identité solide car facile à contourner. Afin de résoudre ce problème, la biométrie est de plus en plus utilisée pour augmenter le niveau de confiance de l'authentification des utilisateurs. Néanmoins, les données biométriques sont sensibles et nécessitent une attention particulière en termes de sécurité et de respect de la vie privée. La protection des données biométriques doit être réalisée pendant le cycle de vie des données, du stockage à la manipulation. La cryptographie standard n'est pas en mesure d'assurer la protection des données lors de l'étape de comparaison (déchiffrement nécessaire). Plusieurs solutions sont proposées dans la littérature pour assurer la protection des données biométriques dont les algorithmes de crypto-biométrie [2, 3] ou les algorithmes de transformation [4, 5]. Pour plus de détails sur ces schémas, nous renvoyons le lecteur à ce papier [6].

En général, une authentification biométrique se réalise en deux étapes : l'enrôlement et vérification. La première consiste à générer la référence biométrique d'un utilisateur et à la stocker. Au cours de la vérification, une capture biométrique est comparée à la référence biométrique de l'individu présumé pour décision. Afin d'améliorer la sécurité de l'authentification des utilisateurs, il faut généralement combiner différents facteurs d'authentification. Cela peut être réalisé en utilisant différentes données biométriques pour définir un système multi-biométrique.

La principale contribution de cet article est de proposer un système multi-biométrique efficace et utilisable pour améliorer la sécurité de l'authentification des utilisateurs sur les smartphones. Nous combinons deux modalités biométriques, à savoir l'empreinte digitale et une biométrie comportementale. Nous supposons dans ce travail que le smartphone utilisé possède un capteur d'empreintes digitales. Cette hypothèse est réaliste car une enquête récente estime que 67% des smartphones en 2018 disposeront d'un capteur d'empreintes digitales [7]. L'empreinte digitale de référence de l'individu est stockée dans un élément sécurisé du smartphone pour en assurer sa protection. Nous utilisons également une modalité biométrique comportementale (façon de saisir un motif sur un écran tactile) avec un schéma de protection du modèle. Cette solution présente l'avantage d'être très simple à utiliser et très rapide. La référence biométrique est stockée dans le smartphone protégée sous la forme d'un BioCode (code binaire lié à la donnée biométrique) révocable en cas d'attaque.

Ce papier est organisé comme suit. La section 2 fournit un

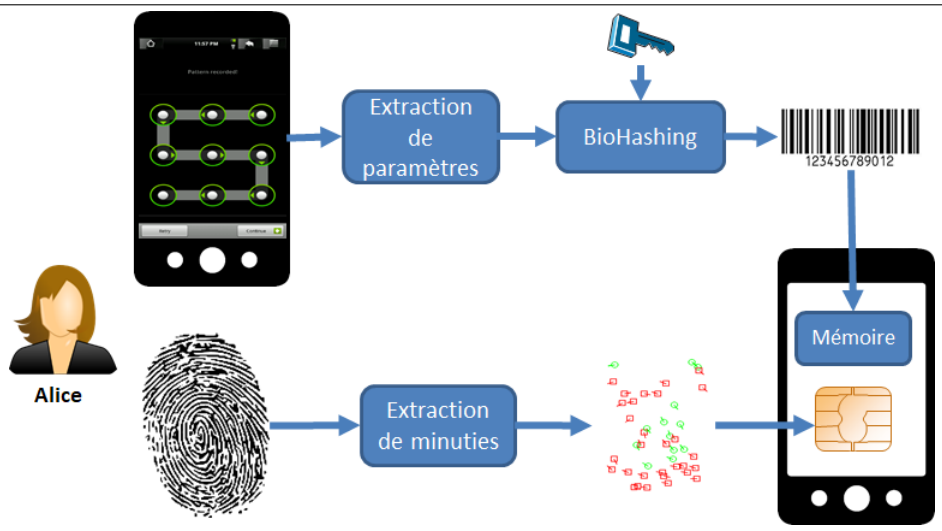


Figure 1 – *Enrôlement d'Alice*

bref état de l'art sur les solutions existantes pour l'authentification des utilisateurs sur un smartphone. La méthode proposée est décrite dans la section 3. La section 4 illustre la performance de la solution proposée à partir de résultats expérimentaux. Enfin, nous concluons et donnons des perspectives dans la section 5.

2 Travaux antérieurs

L'authentification biométrique sur mobile est un problème émergent, avec des références relativement croissantes. Un rapport du NIST [8] propose plusieurs recommandations concernant l'acquisition de données biométriques sur un smartphone et considère les modalités suivantes : l'empreinte digitale, le visage et l'iris. La plupart des papiers de la littérature sont consacrés à une modalité particulière. On peut citer par exemple les références [9, 10] sur la reconnaissance du locuteur. Plusieurs articles [11, 12] traitent de la reconnaissance basée sur la dynamique de frappe. De nombreux articles proposent d'utiliser l'écran tactile pour capturer des données biométriques [13]. La plupart de ces études utilisent des méthodes utilisées pour la dynamique de frappe au clavier ou de signature. Par exemple, la notion de TapPrint a été proposée par Miluzzo et al. [14] où la notion de dynamique de frappe est généralisée à l'écran tactile. La méthode proposée est basée sur le comportement de saisie d'un texte sur un écran tactile et exploite des informations de l'accéléromètre du smartphone. L'efficacité de la reconnaissance est comprise entre 80% et 90%. Le travail réalisé par Luca et al. [15] est très intéressant car il combine le mot de passe à base de motif et la biométrie. Ils ont proposé un système et l'ont testé avec 34 utilisateurs. Ils ont obtenu une performance de 19% pour la valeur FRR (faux taux de rejet) et 21 % pour le FAR (Fausse Acceptation). En 2013, une méthode a été proposée [16] combinant plusieurs informations avec le facteur de corrélation ou la DTW comme méthode de mesure de similarité. Le taux d'erreur égal (EER) est proche de 17 %.

Outre la littérature consacrée aux solutions biométriques pour l'authentification sur mobile liées à une modalité spécifique, plusieurs articles ou thèses récents proposent un aperçu complet du sujet [17, 18]. Même si l'utilisation de plusieurs modalités biométriques permet de limiter le taux d'erreur (notamment le taux de fausse acceptation), l'utilisation de données biométriques supplémentaires pose le problème de la protection de la vie privée. Peu de contributions considèrent ce problème sur un mobile. Nous proposons dans ce papier une nouvelle solution d'authentification combinant l'empreinte digitale et une modalité comportementale. La première modalité est présente dans la plupart des smartphones et le modèle de référence biométrique est stocké en toute sécurité dans un élément sécurisé. L'utilisation de l'approche comportementale permet d'améliorer le niveau de sécurité tout en fournissant une solution d'authentification pratique (interaction rapide) et facile à protéger.

3 Méthode proposée

Le principe général de la méthode proposée est donné par les figures 1 et 2. Pendant la phase d'enrôlement, Alice doit fournir son empreinte digitale au capteur du smartphone pour générer son template de référence. Le modèle calculé (ensemble des minuties détectées) est stocké sur un élément sécurisé dédié. Alice doit également entrer un chemin secret sur l'écran tactile. L'application calcule plusieurs paramètres en fonction de son comportement de saisie (vitesse, pression, façon de tenir le smartphone). Nous utilisons ensuite l'algorithme de BioHashing pour protéger ce modèle. Pour cet algorithme, nous avons besoin d'une clé secrète pour pouvoir révoquer le BioCode généré en cas d'attaque. Cette clé secrète peut

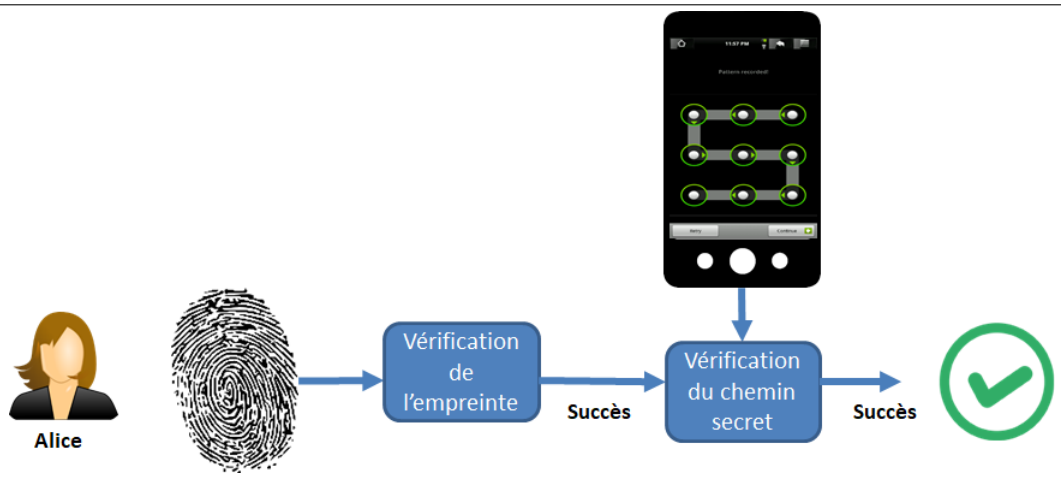


Figure 2 – *Vérification multi-biométrique*

être une représentation binaire du chemin dessiné et peut être concaténée avec d'autres informations telles que le numéro IMEI du smartphone (identifiant du mobile), le nom d'Alice, une valeur aléatoire ...

Au cours de la phase de vérification, Alice doit fournir son empreinte digitale pour s'authentifier. L'empreinte digitale capturée est comparée à la référence d'Alice stockée dans l'élément sécurisé du smartphone. Si son identité est vérifiée, elle doit entrer le motif sur l'écran tactile. Un BioCode est calculé et comparé au BioCode de référence d'Alice dans le smartphone. Si les deux systèmes biométriques acceptent la preuve d'Alice, elle est authentifiée. Nous sommes ici dans un cas de fusion de décision (n'ayant pas accès au score de comparaison de l'algorithme de comparaison des empreintes digitales).

3.1 Vérification de l'empreinte digitale

Nous proposons d'utiliser dans cet article l'empreinte digitale comme première modalité biométrique. La plupart des smartphones intègrent un capteur d'empreinte digitale. Alice doit s'enrôler en fournissant une ou plusieurs captures de son empreinte digitale. Le modèle de référence (un ensemble de minuties) est stocké dans l'élément sécurisé (SE) associé au smartphone ou au matériel du capteur d'empreintes digitales. La comparaison entre une nouvelle capture d'empreinte digitale et le modèle de référence d'Alice est également réalisée en SE et une valeur de décision est fournie (le score n'est pas disponible pour des raisons de sécurité).

Concernant le respect de la vie privée, cette solution est appropriée, la référence biométrique est stockée dans le SE. En terme de sécurité, la solution est intéressante même si le processus d'enrôlement est réalisé par l'utilisateur sans aucun contrôle. Nous pouvons nous attendre à ce qu'un smartphone soit un objet personnel, le processus d'enrôlement devrait être effectué par le propriétaire du smartphone (Alice dans ce cas). En ce qui concerne la performance, il est décrit par la valeur du taux de Fausse acceptation (FAR) correspondant au pourcentage d'attaques réussies par un imposteur. Pour les smartphones, le niveau ciblé du FAR est de 0.005% [19], correspondant au niveau de sécurité 3. Il est difficile de vérifier cette valeur sans le score fourni par l'algorithme correspondant. Le taux de faux rejet associé (FRR) correspondant aux problèmes de reconnaissance des utilisateurs légitimes est censé être inférieur à 2 $\%$. Aucune étude n'existe dans l'état de l'art sur l'évaluation des capteurs commerciaux d'empreintes digitales sur smartphones. La raison est qu'il faudrait avoir un nombre d'utilisateurs très important pour fournir des résultats significatifs.

3.2 Chemin secret biométrique

Le système biométrique que nous proposons d'utiliser dans cette étude a pour but d'accroître la sécurité pour un contrôle d'accès logique rapide sur un smartphone. Nous proposons de reconnaître l'utilisateur par la connaissance d'un mot de passe représenté par un chemin secret. Cette approche pour entrer un mot de passe est plus rapide et plus conviviale pour un appareil mobile. Deuxièmement, le comportement de l'utilisateur lors de la saisie du chemin est analysé. De nombreuses informations peuvent être collectées lors du processus de capture :

- Position X : la position horizontale du doigt sur l'écran tactile est enregistrée pendant la capture,
- Position Y : la position verticale du doigt sur l'écran tactile est également enregistrée,
- Pression : la pression du doigt sur l'écran tactile est capturée (fournie par le système d'exploitation),
- Taille du doigt : nombre de pixels où le doigt est en contact avec l'écran tactile,
- Accéléromètres : trois angles correspondant à l'orientation du smartphone.

Dans cette étude, nous n'avons utilisé que les informations

de position X et Y en calculant également la dérivée première et seconde de chaque signal. Comme le temps nécessaire pour dessiner le même chemin peut être différent pour chaque capture, les signaux sont sous-échantillonnés à une longueur fixe. Une description de taille constante est nécessaire pour utiliser ce modèle comme entrée pour l'algorithme de BioHashing que nous détaillons dans la section suivante.

3.3 Protection par BioHashing

L'algorithme Biohashing est appliqué aux données biométriques représentées par un vecteur à valeur réelle de longueur fixe et génère un modèle binaire appelé BioCode de longueur inférieure ou égale à la taille d'origine. Cet algorithme a été initialement proposé pour le visage et les empreintes digitales par Teoh *et al.* dans [4]. L'algorithme de biohashing peut être appliqué sur toutes les modalités biométriques, qui peuvent être représentées par un vecteur de valeurs réelles de longueur fixe. Cette transformation nécessite un secret lié à l'utilisateur. La comparaison des BioCodes est réalisée par le calcul de la distance de Hamming. L'algorithme de Biohashing transforme le modèle biométrique $T = (T_1, \dots T_n)$ dans un modèle binaire appelé BioCode $B = (B_1, \dots B_m)$, avec $m \leq n$, comme suit :

1. m vecteurs aléatoires orthonormés $V_1, \dots, V_m$ de la longueur n sont générés à partir d'un secret servant de germe du tirage aléatoire (typiquement avec l'algorithme de Gram Schmidt).
2. Pour $i = 1, \dots, m$, calcul du produit scalaire $x_i =< T, V_i >$.
3. Calcul du BioCode $B = (B_1, \dots, B_m)$ avec le processus de quantification :

$$B_i = \begin{cases} 0 & \text{si} \quad x_i < \tau \\ 1 & \text{si} \quad x_i \geq \tau, \end{cases}$$

Où τ est un seuil donné, généralement égal à 0.

La performance de cet algorithme est assurée par le produit scalaire avec les vecteurs orthonormés, comme c'est détaillé dans [20]. Le processus de quantification garantit la non-inversibilité des données (même si $n = m$), car chaque coordonnée de l'entrée T est une valeur réelle, alors que le BioCode B est binaire. Enfin, le germe aléatoire garantit la diversité et les propriétés de révocabilité.

4 Validation du système

Dans cette section, nous présentons des résultats expérimentaux pour la validation du système proposé.

4.1 Protocole

Nous détaillons le protocole que nous avons suivi dans cette étude. Dans ce travail, nous avons d'abord utilisé un ensemble de données biométriques capturées lorsque les utilisateurs dessinent un seul chemin. Les données ont été collectées sur un téléphone portable Nexus 4 avec un écran tactile d'une résolution de 800 x 1280 pixels. Le motif était le même pour tous les utilisateurs et est défini par le code de modèle suivant "1235987". Cette configuration expérimentale peut être considérée comme le pire cas où un attaquant connaît le modèle à dessiner. 34 utilisateurs ont participé à cette expérience et chaque utilisateur a fourni 15 échantillons décrits par 8 signaux sous-échantillonnés à 200 valeurs (normalisation du temps), nous avons utilisé la première et la deuxième dérivée des signaux X et Y. Nous avons également ajouté le temps total pour dessiner le chemin secret. Ainsi, la taille du modèle biométrique comportemental est 1601 (en concaténant tous les signaux sous-échantillonnés et le temps de saisie). Au total, nous avons un sous-ensemble de $34\ fois 15 = 510$ captures biométriques de taille 1601 à valeurs réelles pour le modèle biométrique. Compte tenu de la configuration du BioHashing, nous définissons les valeurs des paramètres comme suit :

- Taille des paramètres d'entrée : n = 1601,
- Taille du BioCode : m = 750 (choix arbitraire, il faut $m < n$ pour garantir la non inversibilité),
- Comme le chemin est le même pour tous les utilisateurs, dans le calcul du BioCode de référence, le code de motif est défini sur "1235987" pour tous les utilisateurs,
- Algorithme de comparaison : distance de Hamming.

En ce qui concerne l'empreinte digitale, nous avons utilisé les bases d'empreintes digitales FVC2002 DB2, FVC2004 DB1 et FVC2004DB3 [21]. La figure 3 présente une image de chaque base de données. On peut voir que les empreintes digitales sont très différentes et représentatives des différents types d'empreintes digitales (acquises avec des capteurs utilisant différentes technologies). Ces bases de données ont des empreintes digitales de 100 individus avec 8 échantillons par personne. Afin de constituer une base chimérique de données multi-biométriques, nous avons pris en compte les empreintes digitales des 34 premiers individus. Pour chaque jeu de données FVC, nous disposons de $34 \times 8 = 272$ échantillons d'empreintes digitales.

(a) (b) (c)

Figure 3 – *Un exemple de chaque base : (a) FVC2002 DB2, (b) FVC2004 DB1, (c) FVC2004 DB3*

Afin d'évaluer la performance de la méthode proposée,

nous utilisons la méthodologie suivante. Nous utilisons le premier échantillon de chaque utilisateur comme modèle de référence, pour le chemin secret, nous utilisons ces données pour calculer le *BioCode de Référence*. Comme nous n'avons pas accès au matériel du capteur d'empreinte digitale (c'est-à-dire la valeur du score correspondant), nous simulons le résultat du score en considérant l'algorithme Bozorth3 fourni par le NIST [22]. Cet algorithme ne fournit pas une performance équivalente aux algorithmes de comparaison commerciale (OCC), la performance est donc sous-estimée. Nous calculons les résultats légitimes comme suit. Nous considérons toutes les empreintes de référence et nous les comparons avec chaque échantillon disponible appartenant au même individu. Nous considérons deux fois ces scores parce que le modèle biométrique comporte 14 échantillons. Pour le chemin secret biométrique, nous comparons le Biocode de référence avec tous les autres BioCodes du même individu. Nous obtenons $14 \times 34 = 476$ scores légitimes pour chaque base de données FVC. Nous avons un processus similaire pour simuler une attaque d'imposteur en considérant tous les échantillons biométriques appartenant à un autre utilisateur. Nous obtenons $14 \times 34 \times 33 = 15708$ scores d'imposture pour chaque base de données FVC. Compte tenu de ces deux ensembles de scores, nous pouvons calculer leur distribution afin d'estimer dans quelle mesure les scores des imposteurs sont différents des légitimes. Deuxièmement, nous calculons la valeur du taux d'erreur égal (EER) qui est une mesure bien connue en biométrie qui mesure le comportement du système biométrique lorsque le seuil de décision est configuré pour avoir le même nombre du taux de faux rejetés et les faux acceptés.

4.2 Résultats

Tout d'abord, nous essayons d'estimer l'efficacité de chaque système biométrique que nous combinons. La figure 4 fournit les courbes DET (de l'anglais Detection Error Trade Off)) du système d'empreintes digitales sur les trois bases de données. La valeur EER est entre 5,2 % et 8 %. Nous pourrions nous attendre en utilisant un système commercial une performance nettement meilleure, cette valeur estime une borne supérieure de l'erreur. En ce qui concerne la performance de reconnaissance par le chemin secret biométrique, avec une simple distance euclidienne, nous obtenons un EER de 27.4 %. En appliquant le BioHashing dans le meilleur des cas (secret seulement connu par l'utilisateur légitime), nous obtenons une reconnaissance parfaite avec une valeur EER de 0 % (voir Figure 5). Dans le pire des cas (secret connu par l'imposteur), la performance est similaire à celle obtenue sans protection.

La figure 6 fournit la distribution du score en combinant l'empreinte digitale et le chemin de secret biométrique dans le meilleur scénario. Nous obtenons pour chacun une reconnaissance parfaite pour toutes les bases de données d'empreintes digitales. C'est évidemment un excellent ré-

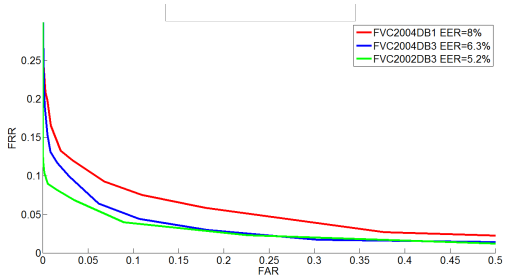


Figure 4 – *Courbe DET de la performance de la reconnaissance d'empreintes digitales sur les 3 bases FVC avec Bozorth3 comme algorithme de comparaison.*

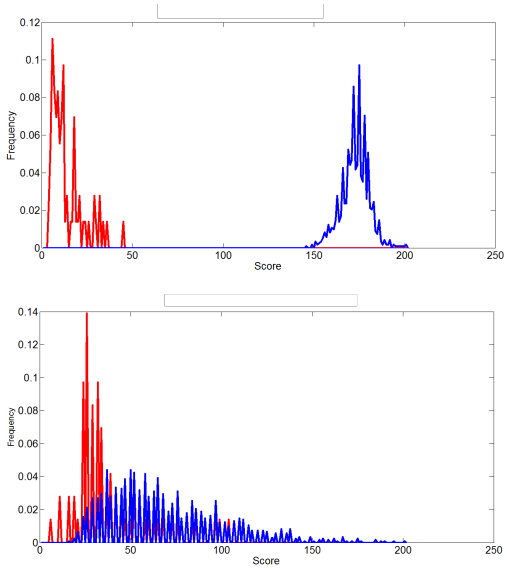


Figure 5 – *Distribution des scores après protection (haut) sans attaque, (bas) secret connu.*

sultat et améliore les résultats en appliquant uniquement le système d'empreinte digitale (voir Figure 4).
Maintenant, nous devons considérer le scénario du pire des cas lorsque l'imposteur a obtenu le secret associé à l'algorithme de BioHashing. Nous supposons que le seuil fixé pour le système d'empreinte digitale est celui associé à la valeur EER (il pourrait être plus strict). Par exemple, pour la FVC2004DB1, nous obtenons un FAR égal à 8 %. Nous fixons la valeur de seuil pour le chemin secret biométrique avec la même approche. Nous avons calculé le taux de fausse acceptation dans le pire des cas et il vaut 28.7%. Cela signifie que si l'imposteur connaît le secret, il a 28.7% de chance de casser le système. En considérant le système multi-biométrique, il a 8% chances de casser le système d'empreinte digitale (sur FVC2004DB1) et 28.7 % pour le chemin secret biométrique. Comme ces événements sont indépendants, nous pouvons estimer le taux de fausse acceptation (FAR) du système multi-biométrique sur FVC2004DB1 à $8\% \times 28.7\% = 2.3\%$. Pour tous les ensembles de données d'empreinte digitale, le FAR est entre

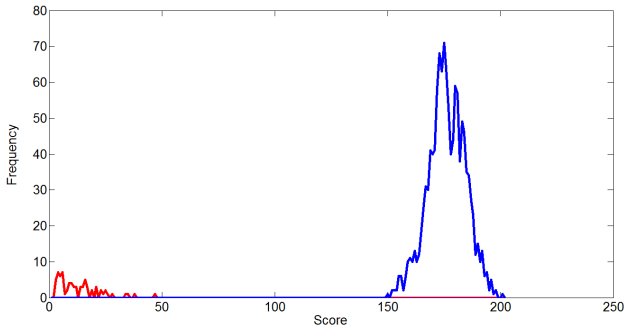


Figure 6 – *Distribution des scores du système multi-biométrique sur la base 2002DB3.*

1.5 % à 2.3 % pour le système multi-biométrique si l'imposteur connaît le chemin secret et le secret. Nous pouvons considérer ce résultat comme très intéressant compte tenu de toutes les informations nécessaires à l'imposteur pour cette attaque (possession temporaire du téléphone, fourniture d'une empreinte digitale proche de celle de l'utilisateur légitime, connaissance du chemin secret, fourniture d'une saisie du chemin proche de celle de l'utilisateur légitime).

5 Conclusion et perspectives

Dans cet article, nous proposons un système multi-biométrique d'authentification pour smartphones en combinant la reconnaissance de l'empreinte digitale à l'aide de son capteur intégré et d'un système biométrique comportemental. Le système proposé est très rapide et simple d'usage pour les utilisateurs car tous ces systèmes de vérification sont couramment utilisés. L'utilisation de la reconnaissance de l'empreinte digitale permet de limiter l'attaque possible du chemin secret lorsque le secret associé à l'algorithme BioHashing est obtenu par l'imposteur. L'utilisation du second système biométrique permet d'augmenter la sécurité de l'authentification des utilisateurs. Dans le meilleur des cas, nous obtenons une reconnaissance parfaite sur les bases de données testées et un taux de fausse acceptation inférieur à 2.3% dans le pire des cas (l'imposteur doit avoir accès au smartphone, connaît le chemin secret et la clé secrète associée à l'algorithme BioHashing).

Nous avons l'intention, à l'avenir, d'intégrer d'autres systèmes biométriques tels que les systèmes de reconnaissance vocale et faciale.

Remerciements

Les auteurs souhaitent remercier la société United Biometrics pour son soutien financier.

Références

[1] Ramona Sukhraj. 31 mobile marketing statistics to help you plan for 2017, 2016.

[2] H. Chabanne, J. Bringer, G. Cohen, B. Kindarji, et G. Zemor. Optimal iris fuzzy sketches. Dans *IEEE first conference on biometrics BTAS*, 2007.

[3] M. Barni, T. Bianchi, D. Catalano, M. Di Raimondo, R. Donida Labati, P. Failla, D. Fiore, R. Lazzeretti, V. Piuri, A. Piva, et Fabio Scotti. A privacy-compliant fingerprint recognition system based on homomorphic encryption and fingercode templates. Dans *BTAS 2010*, 2010.

[4] A.B.J. Teoh, D. Ngo, et A. Goh. Biohashing : two factor authentication featuring fingerprint data and tokenised random number. *Pattern recognition*, 40, 2004.

[5] A. Nagar, K. Nandakumar, et A. K. Jain. Biometric template transformation : A security analysis. *Proceedings of SPIE, Electronic Imaging, Media Forensics and Security XII*, 2010.

[6] C. Rathgeb et A. Uhl. A survey on biometric cryptosystems and cancelable biometrics. *EURASIP J. on Information Security*, 3, 2011.

[7] Statista. Penetration of smartphones with fingerprint sensors worldwide from 2014 to 2018, 2016.

[8] S. Orandi et R. M. McCabe. Mobile id device. best practice recommendation. NIST Special Publication 500-280, 2009. Available from : http ://www.nist.gov/itl/iad/ig/upload/MobileID-BPRS-20090825-V100.pdf.

[9] A. Kounoudes, A. Antonakoudi, V. Kekatos, et P. Peleties. Combined speech recognition and speaker verification over the fixed and mobile telephone networks. Dans *Proceedings of the 24th IASTED International Conference on Signal processing, Pattern Recognition, and Applications*, pages 228–233, 2006.

[10] A. Roy, M. Magimai.-Doss, et S. Marcel. A fast parts-based approach to speaker verification using boosted slice classifiers. *IEEE Trans. on Information Forensics and Security*, 7 :241–254, 2012.

[11] S. Hwang, S. Cho, et S. Park. Keystroke dynamics-based authentication for mobile devices. *Computer & Security*, 28 :85–93, 2009.

[12] T.-Y. Changa, C.-J. Tsaib, et J.-H. Lina. A graphical-based password keystroke dynamic authentication system for touch screen handheld mobile devices. *The Journal of Systems and Software*, 85 :1157 ?1165, 2012.

[13] N. Sae-Bae, N. Memon, et K. Isbister. Investigating multi-touch gestures as a novel biometric modality. Dans *IEEE Fifth International Conference on Biometrics : Theory, Applications and Systems (BTAS)*, 2012.

[14] E. Miluzzo, A. Varshavsky, S. Balakrishnan, et R. Choudhury. Tapprints : your finger taps have fingerprints. Dans *Proceedings of the 10th international conference on Mobile systems, applications, and services*, 2012.

[15] A. De Luca, A. Hang, F. Brudy, C. Lindner, et H. Hussmann. Touch me once and i know it's you ! : implicit authentication based on touch screen patterns. Dans *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 2012.

[16] Michael Beton, Vincent Marie, et Christophe Rosenberger. Biometric secret path for mobile user authentication : A preliminary study. Dans *Computer and Information Technology (WCCIT), 2013 World Congress on*, pages 1–6. IEEE, 2013.

[17] Abdulaziz Alzubaidi et Jugal Kalita. Authentication of smartphone users using behavioral biometrics. *IEEE Communications Surveys & Tutorials*, 18(3) :1998–2026, 2016.

[18] Attaullah Buriro. *Behavioral Biometrics for Smartphone User Authentication*. Thèse de doctorat, University of Trento, 2017.

[19] William E. Burr, Donna F. Dodson, Elaine M. Newton, Ray A. Perlner, W. Timothy Polk, Sarbari Gupta, et Emad A. Nabbus. Nist special publication 800-63-2 : Electronic authentication guideline. Rapport technique, NIST, 2013.

[20] A. B.J. Teoh, Y. W. Kuan, et S. Lee. Cancellable biometrics and annotations on biohash. *Pattern Recognition*, 41 :2034–2044, 2008.

[21] Davide Maltoni, Dario Maio, Anil Jain, et Salil Prabhakar. *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.

[22] Kenneth Ko. Users guide to export controlled distribution of nist biometric image software (nbisec). *NIST Interagency/Internal Report (NISTIR)-7391*, 2007.

Authentification basée sur des habitudes d'appel et garante de la vie privée

Julien Hatin[1,2], Estelle Cherrier[1], Jean-Jacques Schwartzmann[1,2] et Christophe Rosenberger[1]
[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
[2] Orange Labs, 14000 Caen, France

Résumé

L'authentification transparente sur les téléphones mobiles souffre de problèmes liés à la protection de vie privée et en particulier lorsque des données biométriques sont en jeu. Dans cet article, nous proposons une solution pour concilier ces deux points en utilisant l'algorithme de Biohashing sur des données comportementales issues d'un téléphone mobile. Le scénario d'authentification est testé sur une base de données composée de 100 utilisateurs et offre des résultats prometteurs avec un EER de 1% dans le meilleur des cas.

Mots clefs

Authentification, Biométrie comportementale, Protection de la vie privée

Résumé

1 Introduction

L'authentification auprès de différents services passe aujourd'hui de plus en plus souvent par des terminaux mobiles, principalement les smartphones. Malgré l'existence de solutions matérielles permettant de stocker des secrets et de capturer des attributs biométriques, l'authentification auprès des services et applications repose toujours au final sur un mot de passe.
Ceci nuit à l'usabilité et s'avère être un problème de sécurité puisque les mots de passe sont généralement trop faibles et/ou enregistrés de manière permanente sur le smartphone, car il est impossible pour un utilisateur de retenir un nombre grandissant de mots de passes complexes et tous distincts. Une solution élégante pour pallier ce problème est d'utiliser des techniques d'authentification transparente. Afin d'expliquer ce qu'est l'authentification transparente, il convient de citer Nathan Clarke : [1].

Definition 1 (Authentification transparente)**.** *Une authentification transparente peut être réalisée par n'importe quelle approche d'authentification qui est en mesure d'obtenir l'échantillon requis pour la vérification de manière non intrusive.* [1]

La biométrie comportementale est particulièrement adaptée pour obtenir des échantillons dédiés à une authentification transparente. En effet, la biométrie comportementale authentifie quelqu'un grâce à la manière dont il/elle exécute une action (manière de signer, façon de taper au clavier, démarche...). Les modalités de biométrie comportementale sont donc bien moins intrusive que les modalités physiologiques (empreintes digitales, iris, visage...).
Par ailleurs, avec un smartphone, il est maintenant possible d'extraire l'ensemble des informations de contexte : le contexte permet d'avoir accès aux informations liées au comportement de l'utilisateur, comme ses habitudes d'appel ou encore sa position. De plus, le téléphone mobile est déjà largement utilisé dans des buts d'authentification aux services en ligne et est vu comme un compagnon naturel pour l'authentification [2].
Cependant, même si l'authentification transparente facilite le processus, les informations comportementales peuvent largement mettre à mal la protection de la vie privée des utilisateurs. Un autre problème lié au stockage de ces informations dans une base centralisée est la non révocabilité de ces données. Ces deux problématiques ne permettent donc pas de construire des bases de données centralisées contenant des informations comportementales.
Dans cet article, nous proposons une solution d'authentification transparente sur smartphone, garante de la vie privée, qui permette également de révoquer les données comportementales. L'article est organisé de la façon suivante : dans la section 2, nous présentons un rapide état de l'art. La section 3 présente notre solution, et détaille l'algorithme de BioHashing, ainsi que l'architecture proposée. Enfin les résultats obtenus sont présentés dans la section 4, et la section 5 présente des perspectives à ces travaux.

2 Etat de l'art

En Mai 2016, Google a annoncé la mise en service d'un mécanisme d'authentification continue et transparente pour remplacer le couple {login, password} [3].
D'autre part, les méthodes d'authentification comportementale proposant une authentification transparente sont en plein essor. Ceci est principalement dû au projet Américain : "Active Authentication" [4]. Le département de la défense américaine propose de remplacer le mot de passe par une authentification transparente. Plus précisément, cela signifie que les utilisateurs devront désormais s'authentifier en utilisant des techniques biométriques.
Les auteurs de [5] prouvent que combiner la localisation à un facteur d'authentification standard comme le code PIN

1. Traduit de l'anglais : Transparent authentication can be achieved by any authentication approach that is able to obtain the sample required for verification non-intrusively.

augmente la confiance globale dans cette authentification. De plus, l'article montre que les deux localisations principales qui ressortent pour un utilisateur sont : (i) L'habitation et (ii) le lieu de travail. Ceci implique de continuellement savoir où l'utilisateur se trouve.

Les auteurs de [6] utilisent la localisation associée aux informations d'appels pour authentifier un utilisateur. Ils obtiennent un EER (Equivalent Error Rate) de 5.4% en utilisant les 6 derniers appels téléphoniques. Cependant la protection de la vie privée n'est pas prise en compte : du point de vue des utilisateurs, ces données (localisation et informations d'appel) peuvent paraître extrêmement intrusives du point de vue de leur vie privée.

A notre connaissance, il existe peu de solutions qui prennent en compte la protection de la vie privée des utilisateurs. Les auteurs de [7] utilisent un schéma de chiffrement homomorphe. Dans [8], les auteurs résolvent ce problème en stockant directement les données dans le téléphone mobile. Un serveur d'autorisation délègue alors le rôle de l'authentification au téléphone mobile. Cela permet de réduire les risques de fuite de données mais cela ne résout par le problème de révocabilité.

La contribution principale de ce papier est de proposer une solution garante de la vie privée tout en permettant de révoquer les données comportementales.

3 Proposition

La solution proposée combine différentes informations issues des capteurs du téléphone mobile et utilise l'algorithme de Biohashing afin d'assurer à la fois la révocabilité des données sensibles et la protection du caractère personnel de ces données.

3.1 Biohashing

Les données biométriques sont des données personnelles et par conséquent elles ne sont pas révocables. L'algorithme de Biohashing est un cas particulier de biométrie révocable [9, 10].

Comme tout système biométrique, les systèmes révocables ont deux étapes :

- L'enrôlement : Un template (modèle de l'utilisateur) est transformé puis stocké comme référence.
- La vérification : Un échantillon est comparé au template de référence afin d'obtenir une distance ou un score de similitude.

Le principe du BioHashing, illustré à la figure 1, est de combiner (i) une clé secrète K spécifique à l'utilisateur avec (ii) un échantillon de données biométriques exprimé comme un vecteur de taille fixe $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Pour assurer sa protection, la clé K est stockée dans l'élément sécurisé du téléphone mobile (voir la section 3.2).

Le processus de Biohashing est divisé en deux étapes :

- Une projection aléatoire : la clé K est utilisée comme graine pour générer m vecteurs aléatoires $r_j \in \mathbb{R}^n$, $j = 1, \ldots, m$ et $m \leq n$. Après orthonormalisation avec la méthode de Gram-Schmidt [11], ces vecteurs sont regroupés en vecteurs colonnes d'une matrice $O = (O_{i,j})_{i,j \in [1,n] \times [1,m]}$. Une projection sur cette matrice de l'échantillon biométrique $(f_1, \ldots, f_n)$ est alors calculée.
- Une quantification : Cette étape est dédiée à la transformation en un vecteur binaire des données réelles précédemment obtenues après projection. Cette transformation est réalisée avec un simple seuil. Plus précisément, un vecteur binaire $B = (B_1, \ldots, B_m)$ appelé Biocode est obtenu. L'objectif de cette étape est de garantir l'irréversibilité du processus.

La partie suivante détaille l'architecture proposée, qui inclut 'algorithme du BioHashing.

3.2 Architecture

L'architecture globale est composée d'un client (le téléphone mobile) et d'un serveur d'authentification. Dans cet article, le téléphone mobile collecte les données comportementales de l'utilisateur quand celui-ci passe un appel ou envoie un sms. L'approche utilisée peut être étendue à d'autres échantillons issus de capteurs différents de ceux utilisés dans ces travaux. Ceci s'adapte particulièrement pour combiner des informations de géolocalisation avec n'importe quel autre groupe de capteurs.

Architecture cliente. Afin d'évaluer notre proposition, nous avons accès aux données suivantes :

- La position géographique de l'antenne depuis laquelle l'appel a été réalisé
- Le numéro de téléphone de l'appelé

Le processus de vérification est réalisé en ligne. Avant d'être envoyées au serveur, les données doivent d'abord être protégées avec l'algorithme de BioHashing. La figure 2 détaille cette architecture.

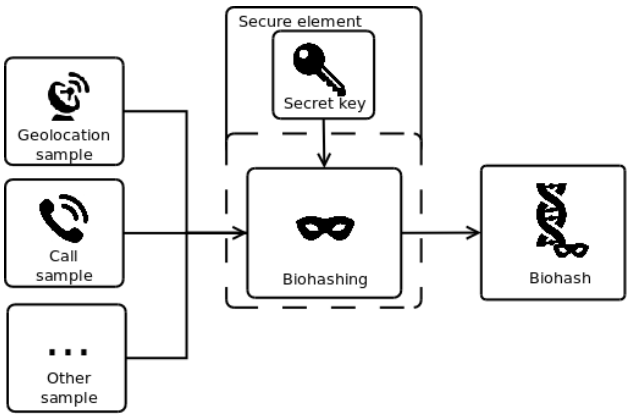


Figure 2 – *Architecture côté client*

Une fois sous forme de Biocode, les données peuvent alors être envoyées en ligne. Afin d'éviter des attaques par rejeu, il est nécessaire d'utiliser un canal sécurisé. Ceci peut être fait en utilisant une connexion TLS [12].

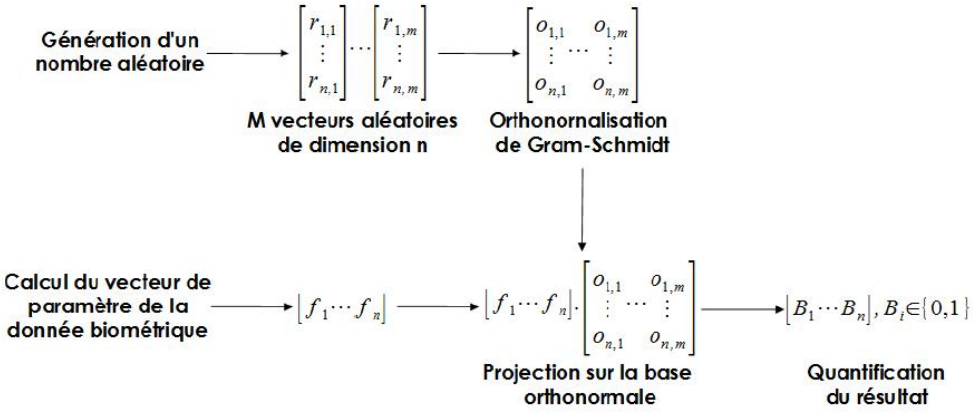


Figure 1 – *Génération du Biocode*

Architecture serveur. Le serveur reçoit des BioCodes en continu, à chaque fois qu'un utilisateur passe un appel. La première étape est de stocker le BioCode en base. Les données stockées permettent de créer un template. Lorsque suffisamment de données sont enrôlées pour un utilisateur, le serveur passe alors en mode vérification. La figure 3 illustre cette architecture.

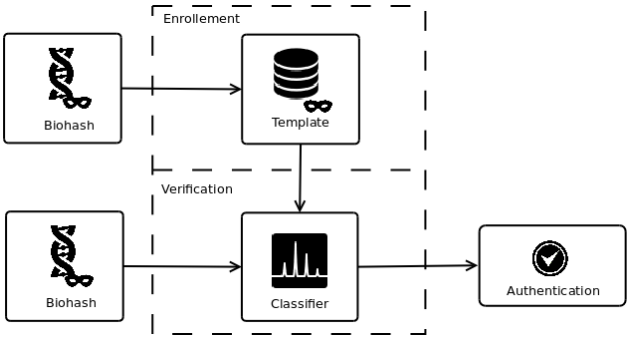


Figure 3 – *Architecture côté serveur*

4 Résultats expérimentaux

4.1 Base de données

La base données utilisée pour évaluer ces travaux contient l'historique des communications de 100 personnes pendant un mois. Ces données sont extraites des informations du réseau d'un opérateur téléphonique.

Les données présentes dans cette base sont :

- La latitude et la longitude de l'antenne
- Le numéro de l'appelant
- Le numéro de l'appelé
- Le type (Appel ou SMS)

Dans la section suivante, nous présentons nos résultats expérimentaux.

Tableau 1 – *Taille de la base d'expérimentation*

Données	Enrôlement	Vérification
Max	666	666
Min	16	14
Moy	157.5	109

4.2 Protocole expérimental

Le correspondance entre les donnée enrôlés et les données présentées est dans la plupart des recherche biométrique réalisé à l'aide d'algorithmes standard de classification.

Dans ces travaux, nous utilisons des classifier ne nécessitant qu'une seule classe. De cette manière il est possible de travailler uniquement sur les données d'un utilisateur.

Ceci représente des conditions d'utilisation dans un cadre industriel où la proportion d'individu malicieux et théoriquement infini.

Dans ces travaux nous utilisons deux classifieurs différents :

- One Class SVM
- Distance au plus proche voisin (KNN)

One Class SVM. Ce classifier est communément utilisé avec des problèmes à deux classes. Le modèle représente les points de l'espaces qui sont séparés par un hyperplan. Chacune des classes est alors situé d'un côté de l'hyperplan.

Dans un One class SVM, les données utilise une fonction de base radiale. Nous recherchons alors l'hyperplan qui entoure au mieux les données enrôlés.

KNN. Cette algorithme place l'ensemble de nos features dans un graphique multidimensionel. Lors de la présentation d'une nouvelle feature, la distance aux voisins est calculé à l'aide d'une distance (par exemple une distance euclidienne). Les points les plus proches sont alors sélectionnés et permettent de définir la classe dans laquelle classer

la feature.
Afin d'utiliser cette algorithme avec une seule classe il est possible de retourner les distances aux voisins les plus proches.

4.3 Résultats

Sans protection. Dans un premier temps nous évaluons notre solution sans utiliser de technique de préservation de la vie privée sur les données.

Tableau 2 – *One Class SVM sans protection de la vie privée*

FRR (%)	FAR (%)
29.54	1.23

Tableau 3 – *KNN sans protection de la vie privée*

Nombre de voisins	EER (%)	Seuil correspondant
1	8.39	0.15
2	8.39	0.30
3	9.15	0.49
4	9.28	0.67
5	9.77	0.86

Meilleur cas. Dans le meilleur des cas, l'attaquant ne dispose pas du téléphone mobile de la victime. Il ne peut donc qu'envoyer ses propre données.

Tableau 4 – *One Class SVM dans le meilleur cas*

FRR (%)	FAR (%)
35.19	0

Tableau 5 – *KNN dans le meilleur cas*

Nombre de voisins	EER (%)	Seuil correspondant
1	1.04	0.30
2	1.10	0.62
3	1.09	0.95
4	1.16	1.29
5	1.19	1.63

Pire cas. Dans le pire des cas, l'attaquant dispose du téléphone mobile de la victime. Il peut donc utiliser ses propres données qu'ils va Biohashé avec la clé de la victime.

Tableau 6 – *One Class SVM dans le pire des cas*

[h !] FRR (%)	FAR (%)
34.60	2.68

Tableau 7 – *KNN dans le pire des cas*

Nombre de voisins	EER (%)	Seuil correspondant
1	10.45	0.23
2	10.16	0.47
3	10.65	0.72
4	10.69	0.98
5	10.76	1.24

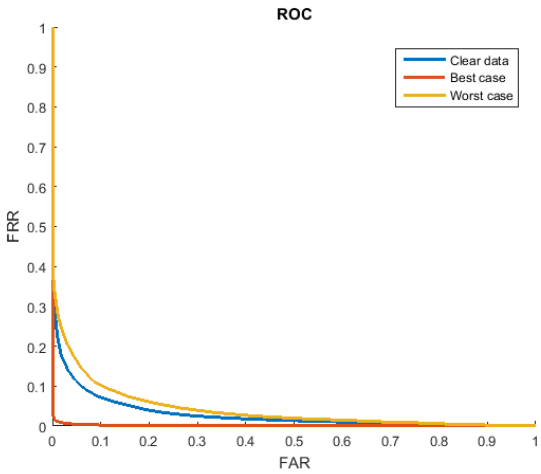


Figure 4 – *Courbe ROC des différents scénarios*

5 Conclusion

La protection des données personnelles est un des problèmes majeur de l'authentification transparente [15].
Notre proposition, basée sur l'algorithme de Biohashing, permet d'authentifier un utilisateur avec des résultats similaires à l'état de l'art tout en permettant de protéger la vie privée de l'utilisateur.
La table 8 compare notre solution à l'état de l'art. Cela permet de mettre en avant les deux contributions majeures de cette article qui permettent d'envisagé des solution d'authentification transparente comme facteur d'authentification pour des services en ligne :

— La révocabilité de la solution
— La protection des données à caractère personnelles

Enfin, ce framework offre des performances biométriques et en temps de calcul suffisant pour envisager une intégration dans les systèmes d'authentification actuels. A la suite

Tableau 8 – *Comparaison avec les autres solutions*

Ref.	No. d'utilisateurs	Features	Privacy	Performance	Révocable
Li et al.[6]	71	Location & call	Aucun	EER :8.8% avec 1 échantillon et EER :5.3% avec 6 échantillons	✗
Savaenee et al.[13]	30	analyse linguistique, dynamique de frappe et analyse comportementale	Aucun	EER :3.3%.	✗
Tanviruzzaman et al.[2]	13	gps et démarche	None	EER :10%	✗
Fridman et al.[14]	200	GPS	Aucun	FAR :11% et FRR :6%	✗
Fridman et al.[14]	200	SMS, gps, applications, web browsing	Aucun	ERR :5% après 1 minute et EER :1% après 30 minutes	✗
Safa et al.[7]	Non disponible	appels, Localisation, Réseaux Wi-Fi, sites webs visités	3-round protocol between the device and carrier	Non disopnible	✓
Notre framework	100	appels, localisation	Biohashing, privacy leakage : 4 bits	EER :1.04% dans le meilleurs cas, EER :10.45% dans le pire des cas	✓

de ces travaux, des études supplémentaires sont mené pour intégrer ces contributions dans des solutions industrielles.

Références

[1] Nathan Clarke. *Transparent User Authentication Biometrics, RFID and Behavioural Profiling*. Springer, 2011.

[2] Mohammad Tanviruzzaman et Sheikh Iqbal Ahamed. Your phone knows you : Almost transparent authentication for smartphones. Dans *Computer Software and Applications Conference (COMPSAC), 2014 IEEE 38th Annual*, pages 374–383. IEEE, 2014.

[3] Google. Google Abacus project. `http://www.androidcentral.com/project-abacus-atap-project-aimed-killing-password`. [Online ; accessed 10-July-2016].

[4] Richard P Guidorizzi. Security : Active authentication. *IT Professional*, 15(4) :4–7, 2013.

[5] Eiji Hayashi, Sauvik Das, Shahriyar Amini, Jason Hong, et Ian Oakley. Casa : Context-aware scalable authentication. Dans *SOUPS '13 Proceedings of the Ninth Symposium on Usable Privacy and Security*, 2013.

[6] Fudong Li, Nathan Clarke, Maria Papadaki, et Paul Dowland. Active authentication for mobile devices utilising behaviour profiling. *International Journal of Information Security*, 2013.

[7] Nashad Ahmed Safa, Reihaneh Safavi-Naini, et Siamak F Shahandashti. Privacy-preserving implicit authentication. Dans *IFIP International Information Security Conference*, pages 471–484. Springer, 2014.

[8] Richard Chow, Markus Jakobsson, Ryusuke Masuoka, Jesus Molina, Yuan Niu, et Zhexuan Song. Authentication in the clouds : A framework and its application to mobile users. Dans *Proceedings of the 2010 ACM workshop on Cloud computing security workshop*, 2010.

[9] N.K. Ratha, J.H. Connell, et R. Bolle. Enhancing security and privacy in biometrics-based authentication system. *IBM Systems J.*, 37(11) :2245–2255, 2001.

[10] R.M. Bolle, J.H. Connell, et N.K. Ratha. Biometric perils and patches. *Pattern Recognition*, 35(12) :2727–2738, 2002.

[11] Åke Björck. Solving linear least squares problems by gram-schmidt orthogonalization. *BIT Numerical Mathematics*, 7(1) :1–21, 1967.

[12] Tim Dierks. The Transport Layer Security (TLS) Protocol Version 1.2. RFC 5246, Octobre 2015.

[13] Hataichanok Saevanee, Nathan Clarke, Steven Furnell, et Valerio Biscione. Text-based active authentication for mobile devices. Dans *ICT Systems Security and Privacy Protection*, pages 99–112. Springer, 2014.

[14] Lex Fridman, Steven Weber, Rachel Greenstadt, et Moshe Kam. Active authentication on mobile devices via stylometry, application usage, web browsing, and gps location. *arXiv preprint arXiv :1503.08479*, 2015.

[15] V. M. Patel, R. Chellappa, D. Chandra, et B. Barbello. Continuous user authentication on mobile devices : Recent progress and remaining challenges. *IEEE Signal Processing Magazine*, 33(4) :49–61, July 2016.

Etude de la dynamique de frappe au clavier pour l'authentification sur les terminaux mobiles avec code PIN

Wael Elloumi, Olivier Maas
Worldline
19, rue de la Vallée Maillard 41000 Blois
{wael.elloumi, olivier.maas}@wordline.com

Résumé

Dans ce papier, nous proposons d'étudier la dynamique de frappe au clavier (DDF), qui est une modalité biométrique comportementale, afin d'authentifier l'utilisateur selon la façon dont il saisit son code confidentiel sur un terminal mobile. Notre solution combine 3 facteurs d'authentification dont un facteur de connaissance (un code confidentiel à 6 chiffres), un facteur biométrique comportemental (la DDF) et un facteur de possession (le téléphone) et ce dans une démarche en parfaite conformité avec les réglementations CNIL et règlement européen sur la protection des données personnelles. Nous présentons également une étude expérimentale sur l'entropie de la clé biométrique générée par la DDF pour quantifier la sécurité de notre système contre les attaques des imposteurs.

Mots clefs

Authentification mobile, Biométrie comportementale, Dynamique de frappe au clavier, Entropie de la dynamique de frappe au clavier.

1 Introduction

Fortement intégrés dans le quotidien des consommateurs, les services bancaires sur mobile (Mobile Banking) sont en pleine expansion. Ce progrès amène des risques tels que l'augmentation de la fraude liée à l'usurpation d'identité ou à la réutilisation du même mot de passe pour plusieurs services, etc.

Bien que les facteurs de connaissance (un mot de passe, un code PIN, etc.) constituent la forme d'authentification la plus couramment utilisée, ils restent vulnérables car ils peuvent être volés, partagés ou même obtenus par une simple observation (shoulder surfing). Quant aux facteurs de possession (un Token, une carte à puce, un smartphone, etc.), ils doivent être avec l'utilisateur tout le temps ce qui implique un risque de perte et de vol. Les facteurs d'inhérence (la biométrie) représentent la seule forme d'authentification qui nécessite la présence physique de l'utilisateur. Cependant, ils peuvent être compromis par les attaques de présentation (spoofing). Pour une authentification forte, plusieurs facteurs peuvent être combinés. Néanmoins, de telles solutions sont souvent coûteuses, compliquées à mettre en place ou peu appréciées en termes d'expérience utilisateur.

Compte tenu des menaces des services bancaires sur mobile, ce travail présente une étude de l'authentification des utilisateurs sur les terminaux mobiles en combinant deux facteurs d'authentification: un code PIN à six chiffres et la DDF. Notre solution est peu coûteuse, non intrusive, complètement transparente pour l'utilisateur et conforme aux réglementations CNIL et aux nouveaux règlements européens sur la protection des données personnelles (PSD2[1], GDPR[2] et eIDAS[3]). De plus, elle permet d'améliorer l'expérience utilisateur et d'augmenter la sécurité de l'ensemble du mécanisme d'authentification tout en conservant sa facilité d'utilisation.

Les principales contributions de ce travail sont les suivantes:

• Etude des différentes données (les temps et les données spatiales) et classificateurs afin de quantifier leur impact sur la précision du système.
• Évaluation du système proposé dans des conditions réalistes.
• Réalisation d'une étude expérimentale sur l'estimation de l'entropie des clés biométriques générées par notre système afin de mesurer sa sécurité contre les attaques des imposteurs.

Le reste de ce papier est organisé comme suit. La section 2 dresse un état de l'art sur la DDF. La section 3 présente un aperçu des différentes étapes de notre système. La section 4 détaille les résultats obtenus. La section 5 décrit notre étude expérimentale sur l'estimation de l'entropie des clés biométriques générées par la DDF. Enfin, la section 6 conclut.

2 Etat de l'art

La DDF est une modalité biométrique comportementale qui permet d'authentifier un utilisateur selon sa façon de saisir un texte (rythme de frappe, temps, etc.) sur un clavier standard ou virtuel pour les terminaux avec écrans

[1] La nouvelle directive européenne sur les services de paiement (en anglais : Payment Services Directive 2, PSD2)
[2] Le Règlement général sur la protection des données (en anglais : General Data Protection Regulation, GDPR)
[3] Le règlement sur l'identification électronique et les services de confiance pour les transactions électroniques (en anglais : Electronic ID and Trust Services, eIDAS)

tactiles. Bien qu'initialement, comparée à d'autres modalités biométriques, elle n'a pas été suffisamment étudiée, la DDF est devenue progressivement une thématique de recherche active grâce sa contribution au renforcement de la cybersécurité ainsi que de sa commodité pour l'utilisateur.
Cette section présente un aperçu sur les modes de fonctionnement étudiés, les données biométriques extraites et les approches de comparaison ou de classification des systèmes d'authentification par la DDF.

2.1. Mode de fonctionnement

L'utilisation de la DDF à des fins d'identification et de vérification a été explorée avec du texte statique ou dynamique (avec du texte libre). Les systèmes de DDF par texte statique utilisent un texte préenregistré lors de l'étape de l'enrôlement (une passphrase, un mot de passe ou un code PIN) pour permettre l'authentification de l'utilisateur à un moment bien précis (pour se connecter à une session par exemple) [1, 2]. Cependant, les systèmes de DFF par texte dynamique reposent sur du texte libre pour permettre l'authentification continue de l'utilisateur pendant toute la session [3, 4]. Par conséquent, pour réduire les risques de sécurité, les systèmes de DDF par texte statique permettent une authentification par deux facteurs en combinant un facteur de connaissance (mot de passe, passephrase ou code PIN) et un facteur d'inhérence (la DDF). Cependant, les systèmes de la DDF par texte dynamique reposent uniquement sur seul facteur d'authentification (la DDF) mais peuvent vérifier l'identité de l'utilisateur en permanence [5].
Dans une étude récente, Teh et al. [6] ont rapporté que la plupart des recherches sur la DDF concerne le mode vérification (74%) comparées à celles du mode d'identification (26%). Ils ont également mentionné que la majorité des études de la DDF portaient sur l'authentification par texte statique (77%) comparées à celles de l'authentification par texte dynamique (23%). Les statistiques fournies par [7] reflètent également les mêmes tendances: 89% pour le mode vérification contre 5% pour le mode d'identification et 83% pour l'authentification par texte statique contre 10% pour l'authentification par texte dynamique.

2.2. Données biométriques

Les données de la DDF sont les caractéristiques ou les primitives qui sont extraites au moment où l'utilisateur saisit du texte (statique ou dynamique) afin de créer une référence ou une signature biométrique (gabarit biométrique). Les caractéristiques couramment utilisées ont été initialement basées sur des informations de temps qui sont principalement le temps de pression des touches (durée d'appui), les temps de latence entre l'appui des touche (ou temps de vol), le temps entre le relâchement d'une touche et la pression d'une autre, le temps entre deux relâchements de touches. Suivant les études, d'autres informations peuvent être extraites à partir des mêmes données brutes comme le digraphe qui représente le temps nécessaire à la saisie de deux lettres (le délai entre la pression de la première touche et le relâchement de la suivante), le temps de séquence (trigraphes, n-graphes) et la vitesse de frappe globale. Il faut noter que certaines études considèrent le digraphe comme étant le temps de latence. Depuis l'arrivée des appareils mobiles tactiles, souvent équipés de capteurs sophistiqués, d'autres types de données ont également été explorés. Elles peuvent être regroupées en deux catégories: des caractéristiques spatiales et des caractéristiques de mouvement. Les données spatiales incluent la pression du doigt appliquée sur les touches, la surface de pression et la position de pression, tandis que les données de mouvements incluent les mesures brutes extraites de l'accéléromètre et du gyroscope. Afin d'améliorer les performances de leurs systèmes d'authentification, de nombreux travaux sur la DDF ont combiné les données conventionnelles (basées sur le temps) avec celles des terminaux mobiles tactiles [3, 8].

2.3. Méthodes de classification

Les travaux sur la DDF ont exploré plusieurs techniques de classification existantes qui peuvent être regroupées en trois grandes catégories: les approches basées sur les distances, les méthodes statistiques simples et les méthodes d'apprentissage automatique (machine learning). Les méthodes de classification basées sur les distances sont les plus utilisées grâce à leur simplicité et facilité de mise en œuvre. Les distances le plus étudiées sont la distance Euclidienne [2], la distance de Manhattan [2], la distance « scaled Manhattan » [9], la distance de Mahalanobis [10] et la distance de Bhattacharyya [11]. Les méthodes statistiques ont été bien explorées aussi car elles représentent un bon compromis entre temps de calcul et précision. De plus, ce type d'approches nécessite beaucoup moins de données d'apprentissage comparées à d'autres méthodes plus complexes. Ces méthodes utilisent des mesures statistiques telles que la moyenne, la médiane et l'écart-type. Les techniques statistiques les plus étudiées sont la densité de probabilité Gaussienne [1], le classificateur Bayésien [12], le modèle de Markov caché [13], les modèles de mélange Gaussien [14] et la probabilité pondérée.
En ce qui concerne les approches d'apprentissage automatique, certes elles nécessitent plus de données comparées à celles des deux premières catégories et elles sont plus complexes à mettre en œuvre mais elles peuvent donner de bons résultats lorsque le système est bien entraîné. Parmi les techniques explorées dans la littérature on peut citer les séparateurs à vaste marge (SVM) [8], les réseaux de neurones [4], les arbres de décision [15], les forêts aléatoires [3], les K plus proches voisins [12], les K-moyennes [16], le Boosting [17], et la logique floue [18].

Afin d'obtenir de meilleures performances, de nombreuses études ont combiné deux ou plusieurs techniques de classification [1, 19, 20]. D'autres travaux ont mené une étude comparative de plusieurs techniques de classification pour évaluer et comparer leurs performances sur un pied d'égalité [15, 21].
A l'issu de cette étudie bibliographique, nous avons pu identifier les choix techniques et méthodologiques appropriés pour notre cas d'utilisation et qui seront détaillés dans la section suivante.

3 Présentation du système

Le but de notre système est d'améliorer la sécurité de l'authentification avec code PIN sur les terminaux mobiles en rajoutant une couche supplémentaire de contrôle de sécurité basée sur la DDF. Notre solution concerne la vérification par texte statique (avec code PIN à six chiffres). Elle comprend deux étapes: une phase d'enrôlement qui permet de générer une référence ou un modèle biométrique à partir des caractéristiques extraites des captures. Ensuite, dans la phase d'authentification, l'utilisateur qui veut accéder au système saisit son code PIN et les données extraites de cette capture (données de test) sont comparées avec la référence biométrique de l'utilisateur cible, enregistrée lors de la phase d'enrôlement. Un score de concordance entre les deux modèles est alors calculé pour la prise de décision (utilisateur authentique ou imposteur).

1. Enrôlement

La phase d'enrôlement se compose essentiellement de l'extraction des caractéristiques et de la génération de la référence biométrique.

Extraction des données : Dans cette étude, nous proposons de combiner les données conventionnelles (basées sur le temps) et spatiales. Au total, notre système utilise neuf types de données. Les caractéristiques basées sur le temps incluent le temps de pression sur une touche (noté PR), trois temps de latence différents et le temps de saisie global (noté VT). Les temps de latence utilisés sont le temps entre l'appui des touches (noté PP), le temps entre le relâchement d'une touche et la pression d'une autre (noté RP) et le temps entre deux relâchements de touches (noté RR). Les caractéristiques spatiales incluent la pression sur les touches (noté TP), la surface de pression (noté TS) et la position de pression (Xpos et Ypos). Au total, pour un code PIN à six chiffres, 46 caractéristiques sont extraites. Le tableau 2 résume la description et le nombre des caractéristiques utilisées par notre système.

Génération de la référence biométrique : La génération de la référence biométrique consiste à transformer les caractéristiques extraites (données biométriques brutes) à partir des captures en une forme compacte qui représente

TABLEAU 1 : DESCRIPTION DES DONNEES EXTRAITES POUR UN CODE PIN A SIX CHIFFRE.

Symbole	Description	nombre
PR	Durée d'une pression relâchement	6
PP	Durée entre deux pressions	5
RP	Durée d'un relâchement pression	5
RR	Durée entre deux relâchements	5
VT	Temps de saisie global	1
TP	Pression sur les touches	6
TS	Surface de contact avec les touches	6
Xpos	Abscisse de la position d'appui sur les touches	6
Ypos	Ordonnée de la position d'appui sur les touches	6

le modèle ou le gabarit de la DDF de l'utilisateur. Le fichier du modèle biométrique est généré à l'issue de l'étape de l'enrôlement afin d'être utilisé dans l'authentification. Ce fichier est mis à jour, à chaque authentification réussie, afin de suivre l'évolution de la DDF de l'utilisateur (adaptation du modèle).
L'en-tête du fichier du modèle biométrique contient les données suivantes: le nom de l'utilisateur, le nombre d'échantillons d'enrôlement n, le nombre d'authentification réussie (facultatif) et le score moyen d'authentification (facultatif). Les deux derniers champs pourraient être utilisés pour le réajustement du seuil d'appariement (seuil adaptatif). Le reste du fichier contient les données suivantes pour chacune des 46 caractéristiques extraites: la somme des carrés des caractéristiques SUMSQ, la moyenne (μ), l'écart type σ et l'écart absolu moyen (MAD).
Soit F un ensemble de n échantillons dont chacun est composé de k caractéristiques $f_{i1}, f_{i2}, \dots, f_{ik}$, la somme des carrés des caractéristiques, la moyenne, l'écart type et l'écart absolu moyen de chaque caractéristique f_j de l'ensemble F sont respectivement définis comme suit :

$$\mathrm{SUMSQ}_{f_j} = \sum_{i=1}^{n} f_{ij}^2 \qquad (1)$$

$$\mu_{f_j} = \frac{1}{n}\sum_{i=1}^{n} f_{ij} \qquad (2)$$

$$\sigma_{f_j} = \sqrt{\frac{\sum_{i=1}^{n}\left(f_{ij} - \mu_{f_j}\right)^2}{n}} \qquad (3)$$

$$MAD_{f_j} = \frac{1}{n}\sum_{i=1}^{n}\left|f_{ij} - \mu_{f_j}\right| \qquad (4)$$

2. Authentification

La phase d'authentification consiste à vérifier l'identité de l'utilisateur en comparant le modèle de la DDF soumis par un sujet (modèle de test) au modèle de référence. Elle se compose principalement de l'appariement des modèles et de l'adaptation de la référence biométrique.

Appariement : Dans ce papier, nous étudions trois techniques de classification différentes : une approche statistique basée sur la fonction de la densité de probabilité Gaussienne (GPD), une méthode basée sur la

distance « scaled Manhattan » alors que la troisième est basée sur la forme de la courbe qui est la mesure de similarité des directions (DSM).
L'approche de la densité de probabilité Gaussienne (GPD) a été détaillée dans [1]. Dans la phase d'enrôlement, la moyenne μ_{f_j} et l'écart type σ_{f_j}de chaque caractéristique sont calculés. Ensuite, dans la phase d'authentification, pour chaque type de caractéristique, un score GPD S_{GPD} entre un modèle de référence et un modèle de test est calculé comme suit:

$$S_{GPD} = \frac{\sum_{i=1}^{c} e^{-\left(\frac{(f_i - \mu_i)^2}{2\sigma_i^2}\right)}}{k} \quad (5)$$

avec f_i les données de test d'une caractéristique donnée, μ_i et σ_i sont, respectivement, la moyenne et l'écart type de la même caractéristique dans le modèle de référence et c représente le nombre de caractéristiques présents dans le code PIN pour un type de caractéristique donné. Soit ft l'ensemble de types de caractéristiques. Le score global GS_{GPD} est calculé comme étant la moyenne des scores de tous les types de caractéristiques.

$$GS_{GPD} = \frac{\sum_{i=1}^{card\,(ft)} S_{GPD_i}}{card(ft)} \quad (6)$$

Dans notre système, $card(ft) = 9$, comme le montre le tableau 1.
La méthode de classification basée sur la distance « scaled Manhattan » (SMD) a été décrite dans [9]. Dans la phase d'enrôlement, la moyenne μ_{f_j} et l'écart absolu moyen MAD_{f_j}de chaque caractéristique est calculé. Dans la phase de test, un score de similarité D_{SMD} est alors estimé, pour chaque type de caractéristique, en calculant la distance « scaled Manhattan » entre les modèles de référence et de test.

$$S_{SMD} = \frac{\sum_{i=1}^{c} \left| f_i - \mu_{f_j} \right|}{MAD_{f_j}} \quad (7)$$

Le score global GD_{SMD} est calculé comme étant la moyenne des scores de tous les types de caractéristiques.
La méthode basée sur la mesure de similarité des directions a été décrite dans [1]. A la différence des deux approches de classification précédentes, cette méthode est basée sur la comparaison des directions de frappe de deux touches consécutives entre les données de test et le modèle de référence. Pour chaque deux touches consécutives, elle consiste à comparer le sens de la courbe pour les données de test avec le sens de la courbe du modèle de référence. Si les deux ont la même direction (même signe), il s'agit d'un appariement et un compteur m est alors incrémenté. Dans la phase de test, pour chaque type de caractéristique, un score S_{DSM} entre le modèle de test et celui de référence est calculé comme suit :

$$S_{DSM} = \frac{m}{d-1} \quad (8)$$

Avec d le nombre de caractères du code PIN (6 caractères dans notre cas). Le score global GS_{DSM} est calculé comme étant la moyenne des scores de tous les types de caractéristiques.
Adaptation de la référence biométrique : La biométrie comportementale est sensible au problème de vieillissement du modèle de référence biométrique vu que le comportement de l'utilisateur change naturellement au cours du temps. Pour remédier à ce problème, nous avons intégré un mécanisme d'adaptation ou de mise à jour du modèle de référence après chaque authentification réussie. Afin d'éviter les problèmes liés au stockage de données, au lieu d'enregistrer les données de toutes les captures de l'utilisateur, notre méthode consiste à mettre à jour uniquement quelques informations du modèle de référence initial qui sont les suivantes : Le nombre d'échantillons de l'enrôlement n dont sa valeur mise à jour est calculée comme suit:

$$n_u = n + 1 \quad (9)$$

La somme des carrés des caractéristiques (SUMSQ) dont sa valeur mise à jour est calculée comme suit :

$$SUMSQ_{f_{ju}} = \sum_{i=1}^{n_u} f_{ij}^2 \quad (10)$$

La moyenne μ dont sa valeur mise à jour est calculée comme suit :

$$\mu_{f_{ju}} = \frac{\sum_{i=1}^{n} f_{ij} + f_{nuj}}{n_u} \quad (11)$$

L'écart type σ en utilisant le théorème de König-Huygens qui relie la variance et la moyenne. Sa valeur mise à jour est calculée comme suit :

$$\sigma_{f_{ju}} = \sqrt{\frac{\sum_{i=1}^{n_u} {f_{ij}}^2}{n_u} - {\mu_{f_{ju}}}^2} \quad (12)$$

4 Résultats expérimentaux

4.1. Protocole expérimental

Notre prototype expérimental comporte un smartphone Nexus 5X doté d'un écran tactile sensible à la pression et d'une version Android 6.0. Les données de temps sont obtenues à l'aide de la classe système d'Android alors que les données spatiales sont obtenues en utilisant l'API « MotionEvent » d'Android. Toutes les caractéristiques de temps sont mesurées en ms (millisecondes).
Soixante-dix collaborateurs de notre société, dont l'âge est compris entre 23 à 60 ans, ont participé à l'expérience. Notre protocole de test est le suivant: dans la phase d'enrôlement, nous avons demandé à chaque participant de saisir le code PIN à six chiffres "024680" 15 fois à l'aide

d'un clavier virtuel statique. Ensuite, dans l'étape d'authentification, chaque participant enrôlé a réalisé 5 authentifications, ce qui nous a permis d'obtenir 350 enregistrements authentiques. Nous avons demandé aussi à 27 participants d'agir comme imposteur en effectuant 5 tentatives d'imposture pour 3 sujets enrôlés choisis au hasard, ce qui nous a permis d'obtenir 410 enregistrements d'imposteurs. Pour évaluer la performance de notre solution avec les différentes méthodes de classification testées, nous avons utilisé les 3 indicateurs de performance les plus utilisés: le taux de faux rejets (en anglais False Rejection Rate : FRR) qui est la proportion des d'utilisateurs légitimes rejeté par erreur. Le taux de fausse acceptation (en anglais False Acceptance Rate : FAR) qui est la proportion des imposteurs acceptés par erreur. Le taux d'égales erreurs (en anglais Equal Error Rate : EER) qui est le taux d'erreur lorsque le FAR est égal au FRR. Il est utilisé pour déterminer la précision globale du système et aussi pour se comparer à d'autres systèmes.

1.1. Résultats

Nous avons testé six méthodes de comparaison différentes dont trois combinent la fonction GPD et la distance « scaled Manhattan » (SMD). Afin de combiner les scores de GPD et SMD, trois règles de fusion différentes ont été comparées: le score moyen (13), le vote AND (14) et le théorème bayésien (15). Le FAR, le FRR et le EER obtenus avec les six approches de classification sont présentés dans les tableaux 2 et 3.

$$S_f = \frac{S_{GPD}+S_{SMD}}{2} \quad (13)$$

$$authentique = \begin{cases} accepter, & S_{GPD} > th,\ S_{SMD} > th \\ rejeter, & si\ non \end{cases} \quad (14)$$

$$S_f = P(S_{GPD}|S_{SMD}) = \frac{S_{GPD}*S_{SMD}}{(S_{GPD}*S_{SMD})+((1-S_{GPD})*(1-S_{SMD}))} \quad (15)$$

avec th est un seuil d'appariement et S_f représente le score final après la fusion des scores S_{GPD} et S_{SMD}.
La figure 1 représente les courbes ROC ainsi que l'EER moyen de chaque méthode de classification. L'EER moyen de chacune de ces méthodes correspond à l'intersection de sa courbes ROC avec la droite d : FAR=FRR. Les résultats obtenus montrent que la méthode SMD et celle du score moyen du GPD et SMD donnent le meilleur EER moyen suivies de l'approche de combinaison bayésienne. En revanche, la méthode de fusion par le vote AND n'est pas pertinente vu que ses résultats sont similaires à ceux de la méthode GPD. Les résultats montrent également que la fonction GPD donne le meilleur FAR suivie par l'approche de fusion du score moyen. Compte tenu des trois indicateurs de performance (FAR, FRR et EER), la méthode du score moyen donne les meilleures performances. Par conséquent, nous avons sélectionné cette approche de fusion pour notre système d'authentification. Il est important de noter que la référence biométrique est estimée à partir de 15 enrôlements seulement. Par conséquent, la performance de notre système pourrait être considérablement améliorée quand le modèle sera mieux entraîné à l'aide du mécanisme d'adaptation.

5 Analyse de sécurité

Dans cette section, nous proposons de quantifier la sécurité de notre système contre les attaques des imposteurs. Pour ce faire, nous générons d'abord une clé biométrique à partir des données de la DDF extraites. Ensuite, nous mesurons empiriquement la force de cette clé en estimant son entropie « Guessing Entropy»

5.1. Génération d'une bio-clé

Dans la littérature [22, 23, 24, 25, 26], les méthodes connues pour la génération de clés cryptographiques à partir des mesures biométriques se composent essentiellement de deux étapes : l'estimation d'un descripteur et la génération de la clé cryptographique.

Les descripteurs des caractéristiques doivent séparer les utilisateurs dans la mesure où les descripteurs produits par le même utilisateur sont suffisamment similaires (une faible variation intra-utilisateur), tandis que ceux produits par des utilisateurs différents sont suffisamment différents (une grande variation inter-utilisateur). Afin de générer ces descripteurs, nous avons utilisé la méthode de Monrose et al. [22]. Chaque caractéristique extraite est représentée sur un bit qui peut avoir 3 valeurs différentes : 0 ou 1 si elle est distinctive et " ⊥ " (indéfini) si non. La sélection des caractéristiques distinctives se fait de la manière suivante. Soient A l'ensemble des comptes utilisateurs et m le nombre des caractéristiques $\varphi_1, \varphi_2, \dots, \varphi_m$ extraites lors de l'authentification au compte a. $\varphi(a,l)$ est la mesure de la caractéristique φ lors de la l^{éme} tentative d'authentification au compte a. Soient B l'ensemble de tous les descripteurs, μ_{ai} et σ_{ai} la moyenne et l'écart type des mesures de la caractéristique φ_i lors des h dernières authentifications réussites au compte a. Le descripteur b_a pour le compte a est défini comme suit :

$$b_a(i) = \begin{cases} 0, & si\ \mu_{ai} + k\sigma_{ai} < t_i \\ 1, & si\ \mu_{ai} - k\sigma_{ai} > t_i \\ \perp, & si\ non \end{cases} \quad (16)$$

Avec k et t_i des paramètres du système dont les valeurs sont déterminées d'une façon empirique. Dans nos tests, la valeur de k a été déterminée empiriquement comme présenté dans la section suivante et le seuil t_i a été choisi comme étant la moyenne globale de la caractéristique φ_i

TABLEAU 2 : LE FAR ET LE FRR DE SIX METHODES DE COMPARAISON OBTENUS AVEC DIFFERENT SEUILS

	FAR				**FRR**			
	Seuil				***Seuil***			
Méthode	***50%***	***60%***	***70%***	***80%***	***50%***	***60%***	***70%***	***80%***
GPD	0.293	0.122	0.017	0.012	0.105	0.281	0.611	0.918
IC (95%)	0.044	0.032	0.013	0.011	0.042	0.052	0.048	0.027
SMD	0.902	0.807	0.6	0.237	0.011	0.017	0.028	0.088
IC (95%)	0.029	0.038	0.047	0.041	0.011	0.014	0.017	0.03
DSM	0.773	0.549	0.322	0.105	0.026	0.063	0.173	0.386
IC (95%)	0.041	0.048	0.045	0.03	0.016	0.025	0.04	0.051
score moyen	0.646	0.402	0.144	0.015	0.006	0.031	0.177	0.649
IC (95%)	0.046	0,047	0.034	0.012	0.008	0.018	0.04	0.05
Vote AND	0.293	0.122	0.017	0.012	0.1	0.277	0.609	0.917
IC (95%)	0.044	0.032	0.013	0.011	0.031	0.047	0.051	0.029
Bayésienne	0.646	0.559	0.427	0.268	0.006	0.014	0.029	0.08
IC (95%)	0.046	0.048	0.048	0.043	0.008	0.012	0.017	0.028

TABLEAU 3 : LES EER OBTENUS POUR LES SIX METHODES DE COMPARAISON.

Méthode	**GPD**	**SMD**	**DSM**	**Score moyen**	**Vote And**	**Bayésienne**
EER	0.2	0.16	0.25	0.165	0.2	0.175

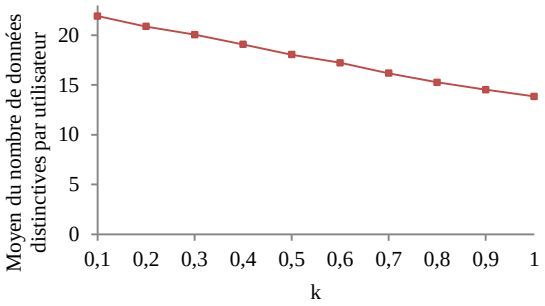


Figure 2 - *Variation du nombre moyen des caractéristiques distinctives par utilisateur en fonction du paramètre k.*

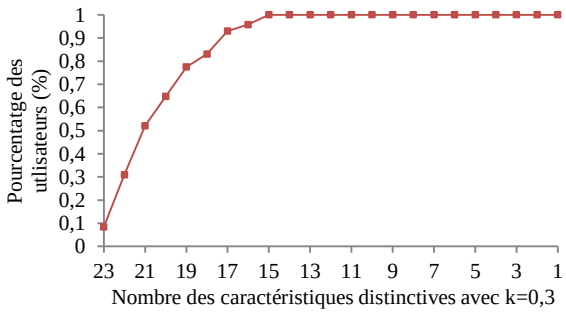


Figure 3 - *Variation du pourcentage des utilisateurs en fonction du nombre de caractéristiques distinctives avec k=0.3*

pour tous les utilisateurs. A la fin de cette étape, on obtient un descripteur b_a pour chaque compte utilisateur a. Ce

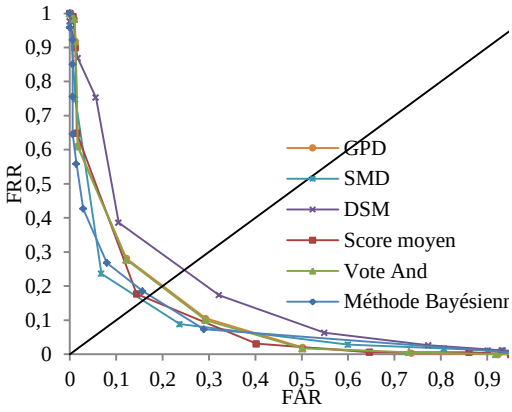


Figure 1 - *Les courbes ROC et l'EER moyen des six méthodes de classification*

descripteur est ensuite utilisé pour la génération d'une clé biométrique dont la longueur maximale est m (dans le où tous les caractéristiques sont distinctives).

5.2. Estimation de l'entropie

Afin d'estimer l'entropie d'une modalité biométrique, plusieurs travaux utilisent le « Guessing Entropy » [22, 23, 26] qui peut être calculé à partir des données empiriques. Le « Guessing Entropy » permet de déterminer le nombre moyen de tentatives à faire par un imposteur pour deviner la clé authentique. Il est défini comme suit :

Soient $A = \{a_1, \dots, a_l\}$ l'ensemble des l comptes utilisateurs présents dans la base et $B = \{b_1, \dots, b_l\}$ leurs descripteurs correspondants. Comme il est possible que différentes personnes possèdent le même descripteur (leurs caractéristiques biométriques se ressemblent), on note $B' = \{b_1, \dots, b_P\}$ l'ensemble des descripteurs parmi B qui sont uniques et ordonnés par leur fréquence f tel que $f(b_1) \geq f(b_2) \geq \cdots \geq f(b_n)$. Le nombre de tentatives qu'un imposteur doit faire est alors :

$$E_G = \sum_{i=1}^{P} (i.f(b_i)) \qquad (17)$$

Plus précisément, le « Guessing Entropy » (E_G) représente le nombre de descripteurs de B' qu'un imposteur aurait besoin d'examiner pour trouver la clé authentique d'un compte utilisateur a choisi au hasard. De plus, E_G suppose que l'attaquant connaît la fréquence $f(b)$ de chaque élément dans B' et donc examine les éléments de B' dans un ordre optimal pour minimiser sa valeur. Par conséquent, la meilleure valeur possible de E_G est $min\left\{2^m, \frac{|A|+1}{2}\right\}$ mais peut être beaucoup plus inférieure si plusieurs utilisateurs génèrent le même descripteur.

Pour estimer le « Guessing Entropie » de notre système, nous avons sélectionné les caractéristiques qui sont indépendantes les unes des autres et qui sont: PR, RP, TP et TS. En se référant au tableau 1, le nombre de ces

caractéristiques indépendantes est $m = 23$. Ensuite, nous avons estimé un descripteur de caractéristiques pour chaque compte utilisateur de notre base de données expérimentale, composée de 71 utilisateurs et 1065 échantillons d'enrôlement. Enfin, nous avons calculé la fréquence de chacun des descripteurs obtenus. La figure 2 montre la variation du nombre moyen des caractéristiques distinctives par utilisateur en fonction de k. Comme l'on peut le voir, une augmentation de la valeur de k engendre une diminution du nombre des caractéristiques distinctives d. Cependant, une diminution de la valeur de k peut produire des doublons (le même descripteur pour deux utilisateurs différents). Dans notre étude, la plus petite valeur de k pour laquelle tous les descripteurs sont uniques est $k = 0{,}3$. Cette valeur de k nous permet d'obtenir le «Guessing Entropy » maximal: $E_G = \sum_{i=1}^{71}(i.(\frac{1}{71})) = \frac{71+1}{2} = 36$ tentatives. Ce choix nous permet d'avoir aussi 21 caractéristiques distinctives pour le compte moyen comme le montre la figure 2. De plus, avec $k = 0{,}3$, environ 78% des utilisateurs ont plus de 19 caractéristiques distinctives et tous les utilisateurs ont plus de 15 caractéristiques distinctives, comme illustré dans la figure 3.

6 Conclusion

Le but principal de cette étude est d'améliorer la sécurité de l'authentification avec code PIN sur les terminaux mobiles en rajoutant une couche supplémentaire de contrôle de sécurité transparente et non intrusive basée sur la DDF. Six méthodes de classification ont été évaluées et comparées dans des conditions réalistes et les résultats obtenus révèlent que la méthode de fusion basée sur le score moyen du GPD et SMD donne les meilleures performances.

Nous avons aussi mené une analyse expérimentale de la sécurité de notre système contre les attaques externes des imposteurs et les résultats obtenus montrent que notre solution nous permet d'obtenir la valeur maximale du « Guessing Entropy ».

Plusieurs améliorations et perspectives peuvent s'envisager dans la continuité des travaux de cette étude comme l'exploration d'autres types de caractéristiques de la DDF telles que les caractéristiques de mouvement (données inertielles) et la validation de notre méthode sur un plus grand nombre d'utilisateurs. Il serait aussi intéressant d'étudier la combinaison de la DDF avec d'autres modalités biométriques, telle que la reconnaissance de visage, afin de créer un système d'authentification biométrique multimodal.

Références

[1] P. S. Teh, A.B. J. Teoh, C. Tee and T. S. Ong, "Keystroke dynamics in password authentication enhancement," Journal of Expert Systems with Applications, 37(12):8618–8627, 2010.

[2] M. J. Coakley, J. V. Monaco, and C. C. Tappert, "Keystroke Biometric Studies with Short Numeric Input on Smartphones," Biometrics: Theory, Applications and Systems (BTAS), 2016..

[3] T. Feng, X. Zhao, B. Carbunar, and W. Shi, "Continuous mobile authentication using virtual key typing biometric," IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 1547–52, 2013. doi:10.1109/TrustCom.2013.272.

[4] B. Draffin, J. Zhu and J. Zhang, "KeySens: passive user authentication through micro-behavior modeling of soft keyboard interaction," Mobile computing, applications, and services, lecture notes of the institute for computer sciences, social informatics and telecommunications engineering. Springer International Publishing, 184–201, 2014.

[5] I. Deutschmann, P. Nordstrom, and L. Nilsson, Continuous authentication using behavioral biometrics," IT Professional, 15(4):12–15, 2013.

[6] P. S. Teh, N. Zhang, A.B. J. Teoh and K. Chen, "A survey on touch dynamics authentication in mobile devices," Journal of computers & security, 59:210–235, June 2016.

[7] P. S. Teh, A. B. J. Teoh, and S. Yue, "A Survey of Keystroke Dynamics Biometrics," The Scientific World Journal, vol. 2013, 2013.

[8] L. Jain, J.V. Monaco, M.J. Coakley, and C.C. Tappert, "Passcode keystroke biometric performance on smartphone touchscreens is superior to that on hardware keyboards," International Journal of Research in Computer Applications & Information Technology, 2(4):29–33, 2014.

[9] L.C.F. Araújo, L. H. R. Sucupira, M. G. Liz´arraga, L. L. Ling, and J. B. T. Yabu-uti, "User authentication through typing biometrics features," 1st International Conference on Biometric Authentication (ICBA), 3071:694–700. Springer-Verlag, Berlin, 2004.

[10] K. Killourhy and R. xion, "The effect of clock resolution on keystroke dynamics," 11th International Symposiumon Recent Advances in Intrusion Detection, 331–350, Cambridge, Mass, USA, 2008.

[11] R. Janakiraman and T. Sim, "Keystroke dynamics in a general setting," Advances in Biometrics, Proceedings, 4642:584–593, Springer, Berlin, Germany, 2007.

[12] D. Buschek, A. De Luca, and F. Alt, "Improving accuracy, applicability and usability of keystroke biometrics on mobile touchscreen devices," 33rd

Annual ACM Conference on Human Factors in Computing Systems, CHI'15, 2015.

[13] J.V. Monaco and C.C. Tappert, "The Partially Observable Hidden Markov Model with Application to Keystroke Biometrics," arXiv preprint, 2016.

[14] D. Hosseinzadeh and S. Kri shnan, "Gaussian mixture modeling of keystroke patterns for biometric applications," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(6):816–826, 2008.

[15] M. Antal, LZ. Szabó, "Keystroke dynamics on Android platform," 8th International Conference Interdisciplinarity in Engineering, INTER-ENG 2014, 820-826, Romania, 2014.

[16] E. Al Solami, C. Boyd, A. Clark, and I. Ahmed, "User-representative feature selection for keystroke dynamics," 5th International Conference on Network and System Security (NSS '11), 229–233, September 2011.

[17] N. Bartlow and B. Cukic, "Evaluating the reliability of credential hardening through keystroke dynamics," 17th International Symposium on Software Reliability Engineering (ISSRE), 117–126, 2006.

[18] S. Zahid, M. Shahzad, S.A. Khayam, and M. Farooq, "Keystroke-based user identification on smart phones," 12th International Symposium on Recent Advances in Intrusion Detection, 224–243, Saint-Malo, France, September 2009.

[19] S. Hocquet, J. Y. Ramel, and H. Cardot, "User classiflcation for keystroke dynamics authentication," Advances in Biometrics, 4642:531–539, 2007.

[20] S. Dhage, P. Kundra, A. Kanchan , and P. Kap, "Mobile authentication using keystroke dynamics," International Conference on Communication, Information Computing Technology (ICCICT), 1–5, 2015, doi:10.1109/ICCICT.2015.7045746.

[21] K. S. Killourhy, and R. A. Maxion, "Comparing Anomaly-Detection Algorithms for Keystroke Dynamics," 39th Annual International Conference on Dependable Systems and Networks (DSN-2009), 125-134, Estoril, Lisbon, Portugal, June 29-July 2, 2009.

[22] F. Monrose, M.K. Reiter, and S.G. Wetzel, "Password hardening based on keystroke dynamics," International Journal on Information Security, 1(2):69–83, February 2002.

[23] F. Monrose, M.K. Reiter, Q. Li, and S. Wetzel, "Cryptographic key generation from voice," IEEE Symposium on Security and Privacy, 202-213, 2001.

[24] W. Zhang, Y. Chang and T. Chen, "Optimal thresolding for key generation based on biometrics," IEEE Conference on Image Processing, 2004.

[25] L. Ballard, S. Kamara and M.K. Reiter, "The Practical Subtleties of Biometric Key Generation" 17th conference on Security symposium, 61-74, 2008.

[26] W. Zhang, C. Zhang and T. Chen, "Security analysis for key generation systems using face images," International Conference on Image Processing (ICIP '04), 5:3455 – 3458, 2004.

Analyse de la taille minimale d'un bloc de pixels afin d'obtenir une valeur significative de l'entropie : application à la correction d'images chiffrées

P. Puteaux W. Puech

LIRMM, UMR 5506
CNRS, Université de Montpellier
Montpellier, France
{pauline.puteaux, william.puech}@lirmm.fr

Résumé

De nombreuses techniques de cryptographie ont été présentées pour protéger la confidentialité et l'intégrité des images. Le propriétaire de l'image la chiffre à l'aide d'une clef de chiffrement et transfert l'image chiffrée résultante sur le réseau. Si le destinataire de cette image est autorisé à accéder à son contenu en clair, il doit pouvoir la reconstruire sans perte. Cependant, cela n'est pas toujours possible : durant sa transmission, l'image chiffrée peut être bruitée. Dans ce cas, certaines parties de l'image ne peuvent pas être correctement déchiffrées. Pour pallier ce problème, nous proposons d'utiliser l'entropie de Shannon. Nous commençons par analyser le sens de cette métrique en fonction de la taille des blocs. Ensuite, nous décrivons un exemple complet d'utilisation de l'entropie pour la suppression des erreurs dans les images chiffrées bruitées. Les résultats expérimentaux montrent qu'une telle approche, basée sur l'entropie, peut être utilisée en pratique pour corriger parfaitement les images chiffrées bruitées.

Mots clefs

Sécurité multimédia, chiffrement d'images, correction d'images bruitées, analyse statistique.

1 Introduction

Ces dernières années, la sécurité des données visuelles est devenue un sérieux problème. Pour cette raison, de plus en plus d'images sont transférées ou stockées dans le domaine chiffré. Bien que les algorithmes cryptographiques soient efficaces pour la protection des données, ils sont aussi très sensibles au bruit et le remplacement d'un seul bit dans les données chiffrées suffit à rendre difficile la reconstruction de l'image originale. Introduite par Shannon en 1948, l'entropie permet de mesure la quantité moyenne d'information contenue dans un message [1]. Aucune des méthodes précédentes de correction d'images chiffrées bruitées n'est basée sur l'entropie de Shannon. En effet, à cause du caractère éparse de l'échantillon lorsqu'une petite taille de bloc est considérée, utiliser directement l'entropie, sans adapter son calcul, n'est pas possible.

Pour cette raison, dans ce papier, nous nous sommes intéressés à l'analyse de la signification de cette mesure statistique en fonction de la taille des blocs considérée. Nous proposons alors d'adapter le calcul de l'entropie de manière à pouvoir l'utiliser en pratique.

Dans la partie 2, nous introduisons l'état de l'art actuel sur le traitement des images dans le domaine chiffré, et en particulier, les méthodes de correction d'images chiffrées bruitées. Dans la partie 3, nous présentons notre analyse sur le sens de l'entropie en fonction de la taille des blocs considérée. Les résultats expérimentaux sont exposés et discutés dans la partie 4. Enfin, la conclusion et des pistes d'extension de ce travail sont proposées dans la partie 5.

2 Etat de l'art

Les méthodes de chiffrement sont utilisées pour garantir la confidentialité d'un contenu multimédia en rendant le plus alétoire possible une partie ou la totalité de celui-ci. Ces systèmes cryptographiques peuvent être symétriques, quand la même clef secrète est utilisée lors du chiffrement et du déchiffrement (AES, DES) ou asymétriques, quand un couple de clefs – publique et privée – intervient (RSA, cryptosystème de Paillier). De plus, dans les cryptosystèmes symétriques, les données peuvent être chiffrées de manière indépendante ou en faisant intervenir le contenu chiffré lors des étapes précédentes [2]. Par ailleurs, ces données chiffrées peuvent être endommagées durant leur transmission sur un canal bruité ou par insertion de données cachées. Même si la bonne clef secrète est connue pendant la phase de déchiffrement, il devient alors difficile de reconstruire l'image originale sans erreur. Pour lutter contre ce problème, des méthodes de correction des erreurs pour les images chiffrées bruités ont été proposées. Les codes correcteurs d'erreur classiques introduisent de la redondance dans le contenu digital. Après la détection d'une erreur, la correction peut être effectuée de deux façons différentes : directement (*Forward Error Correction, FEC*) ou à l'aide d'une demande automatique de répétition (*Automatic Repeat Request, ARQ*). Des approches de correction des erreurs préservant la confidentialité du contenu ont aussi été proposées. Hu *et al.* ont décrit une technique où un double

chiffrement est utilisé pour réaliser un débruitage basé sur le calcul de moyennes non locales [3]. Certains auteurs suggèrent de recourir au partage de données secrètes, comme SaghaianNejadEsfahani *et al.* dans [4] ou Lathey et Atrey dans [5]. Récemment, Pedrouzo-Ulloa *et al.* ont présenté une méthode de correction des erreurs où ils combinent des équations polynomiales homomorphiques et des opérations de seuillage [6]. D'autres méthodes permettent de supprimer le bruit en effectuant une analyse statistique de chaque bloc de l'image chiffrée pendant la phase de déchiffrement pour déterminer s'il est bien déchiffré ou encore chiffré [7, 8].
Bien que certains papiers se soient intéressés au calcul de l'entropie dans un bloc de pixels [9], aucune des méthodes de correction d'images chiffrées bruitées n'est parvenu à utiliser l'entropie en tant que mesure locale.

3 Analyse de l'entropie en fonction de la taille des blocs

Dans cette partie, nous étudions d'abord la mesure de l'entropie d'ordre zéro en fonction de la taille des blocs. Dans un second temps, nous exploitons la redondance entre les pixels, propriété très utile du domaine clair. Pour cela, nous construisons la carte des distances entre pixels voisins et effectuons une analyse de l'entropie de cette carte.

3.1 Entropie d'ordre zéro

Soit X une image de taille $m \times n$ pixels avec l niveaux de gris α_i $(0 \leq i < l)$, de probabilité associée $p(\alpha_i)$. L'entropie d'ordre zéro d'une image X, exprimée en *bit par pixel* (*bpp*) et positive, est :

$$H(X) = -\sum_{i=0}^{l-1} p(\alpha_i) \log_2(p(\alpha_i)). \tag{1}$$

Dans le cas particulier où les l niveaux de gris α_i ont la même probabilité, la valeur de l'entropie d'ordre zéro est maximale et est égale à :

$$H(X) = -\sum_{i=0}^{l-1} \frac{1}{l} \log_2\left(\frac{1}{l}\right) = \log_2(l)\ bpp. \tag{2}$$

Si l'algorithme de chiffrement est efficace, les valeurs des pixels de l'image chiffrée sont générés pseudo-aléatoirement. De ce fait, la distribution des niveaux de gris de l'image tend vers la distribution uniforme.
La valeur de l'entropie d'une image chiffrée codée sur l niveaux de gris (H_{chiffre}) est alors très proche de l'entropie maximale :

$$H_{\text{chiffre}} \approx \log_2(l)\ bpp. \tag{3}$$

Dans le domaine clair, la distribution des pixels peut être approchée par une distribution normale. Dans le domaine discret, la loi normale de paramètres (μ, σ^2) est semblable à la loi binomiale de paramètres (n, p) :

$$\mathcal{B}(n,p) \sim \mathcal{N}(\mu, \sigma^2) = \mathcal{N}(np, np(1-p)). \tag{4}$$

De plus, si une variable aléatoire X suit la loi binomiale de paramètres n et p, sa probabilité d'être égale à α est :

$$P(X = \alpha) = \binom{n}{\alpha} p^{\alpha}(1-p)^{n-\alpha}. \tag{5}$$

Par conséquent, la valeur de l'entropie correspondante est :

$$H(X) = \log_2\left(\sqrt{2\pi e \cdot np(1-p)}\right) + O\left(\frac{1}{n}\right)\ bpp. \tag{6}$$

Ainsi, dans le domaine clair, la valeur de l'entropie d'une image avec l niveaux de gris (H_{clair}) est approchée par celle de la loi binomiale :

$$H_{\text{clair}} \approx \log_2\left(\sqrt{2\pi e(l-1)p(1-p)}\right) + O\left(\frac{1}{l-1}\right)\ bpp. \tag{7}$$

Si nous comparons la valeur de l'entropie d'ordre zéro d'une image en clair et celle d'une image chiffrée, nous aimerions avoir :

$$\begin{aligned} \log_2\left(\sqrt{2\pi e(l-1)p(1-p)}\right) + O\left(\frac{1}{l-1}\right) &< \log_2(l), \\ \sqrt{2\pi e(l-1)p(1-p)} + O\left(\frac{1}{l-1}\right) &\leq l, \\ 2\pi e(l-1)p(1-p) + O\left(\frac{1}{l-1}\right) &\leq l^2. \end{aligned} \tag{8}$$

Si l est grand, cette inégalité est toujours vraie car $\lambda l \ll l^2$, avec $\lambda < l$. Donc si l est suffisamment grand, l'entropie d'une image en clair est inférieure à celle d'une image chiffrée :

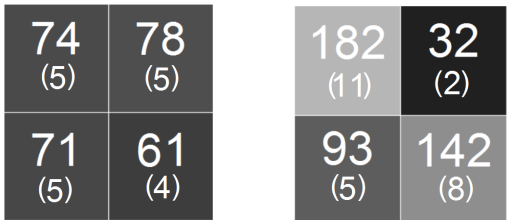
$$H_{\text{clair}} < H_{\text{chiffre}}. \tag{9}$$

Nous proposons maintenant de considérer des blocs de k pixels dans une image codée sur l niveaux de gris à la place de l'image dans sa globalité pour pouvoir définir le concept d'entropie locale. Soit B, un bloc de k pixels dans une image avec l niveaux de gris. L'entropie locale (*i.e.* à l'intérieur d'un bloc B) est majorée par la valeur minimale entre la taille du bloc k et le nombre de niveaux de gris l de l'image :

$$H_{(k,l)}(B) \leq \log_2(\min(k,l))\ bpp. \tag{10}$$

En effet, si la taille du bloc est plus grande que le nombre de niveaux de gris, l'entropie maximale correspond à l'équiprobabilité entre tous les niveaux de gris. Inversement, s'il y a plus de niveaux de gris que de pixels dans le bloc, la valeur de l'entropie maximale est atteinte lorsque tous les pixels ont des valeurs différentes. Dans ce cas, l'échantillon de pixels est éparse car certaines valeurs de niveaux de gris ne sont pas présentes dans le bloc B. Pour cette raison, la mesure de l'entropie peut être éronnée et un bloc clair peut être considéré comme chiffré.
Le problème est illustré à l'aide d'un exemple présenté fig. 1 où nous considérons un bloc de taille $k = 2 \times 2$

pixels avec $l = 256$ niveaux de gris et sa version chiffrée (caractères de grande taille). Dans le bloc de l'image en clair, même si les valeurs des pixels sont relativement proches, puisqu'elles sont toutes différentes, l'entropie est maximale, $H_{(2\times2,256)_{\text{clair}}} = \log_2(\min(4, 256)) = 2$ *bpp*.
Comme nous avons dans les deux cas une valeur maximale de l'entropie, nous ne pouvons pas distinguer un bloc de l'autre en utilisant classiquement l'entropie d'ordre zéro car le nombre de niveaux de gris est bien plus élevé que la taille du bloc. Pour résoudre ce problème, nous proposons de quantifier le nombre de niveaux de gris pour le calcul de l'entropie de façon à diminuer la valeur de l. L'idée est de trouver le meilleur compromis entre la taille des blocs k et le nombre de niveaux de gris l dans l'image.
Si nous reconsidérons l'exemple précédent fig. 1, nous montrons que si nous appliquons une quantification uniforme (valeurs entre parenthèses) à l'image pour se ramener à 16 niveaux de gris, nous levons l'ambiguïté à différencier un bloc en clair de sa version chiffrée. En effet, dans le bloc en clair, trois des quatre valeurs des pixels sont dans le même intervalle $[\![64, 79]\!]$ et sont donc codées avec le même niveau de gris 5. L'entropie correspondante est alors : $H_{(2\times2,16)_{\text{clair}}} = -\frac{3}{4} \cdot \log_2\left(\frac{3}{4}\right) - \frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) = 0.81$ *bpp*.



(a) Bloc de 2×2 pixels dans l'image en clair $H_{(2\times2,16)_{\text{clair}}} = 0.81$ *bpp*.

(b) Bloc de 2×2 pixels dans l'image chiffrée $H_{(2\times2,16)_{\text{chiffre}}} = 2$ *bpp*.

Figure 1 – *Mesure locale de l'entropie dans un bloc de* 2×2 *pixels et sa version chiffrée avec AES (*$l = 256$ *niveaux de gris pour la mesure initiale de l'entropie d'ordre zéro (en caractères de grande taille) et après requantification de l'image* $l = 16$ *(entre parenthèses)).*

3.2 Entropie de la carte des distances

Lors du calcul de l'entropie d'ordre zéro, nous ne tenons pas compte de la corrélation entre les pixels voisins dans le domaine clair. En effet, les valeurs des pixels du même voisinage sont très proches. Cela n'est pas le cas dans le domaine chiffré : la corrélation entre les pixels est très faible puisque les pixels sont générés pseudo-aléatoirement.
Pour exploiter cette corrélation, nous construisons la carte des distances D pour l'image originale X. Les valeurs des distances correspondent aux valeurs absolues entre deux pixels voisins :

$$\forall d \in D,\ d = d(x, x') = |x - x'|\,, \tag{11}$$

avec x et x' deux pixels voisins dans une image X.
Comme l'image originale, la carte des distances est aussi codée sur l niveaux de gris. A l'aide de l'eq. (1), comme chaque valeur de la distance d_i $(0 \leq i < l)$ a la probabilité $p(d_i)$, l'entropie de la carte des distances est :

$$H(D) = -\sum_{i=0}^{l-1} p(d_i) \log_2(p(d_i)). \tag{12}$$

Dans le domaine chiffré, la probabilité théorique associée à la valeur de la distance d est :

$$P(D = d) = \begin{cases} \frac{2(l-d)}{l^2} & \text{if } 1 \leq d \leq l-1, \\ \frac{1}{l} & \text{if } d = 0. \end{cases} \tag{13}$$

En effet, la distribution des distances n'est pas uniforme, comme illustré en fig. 2.a : elle dépend de la valeur originale des pixels dans la paire de voisins. Par exemple, si un pixel x dans la paire est égal à 128, la valeur de la distance est entre 0 et 128, quelque soit la valeur de x' :

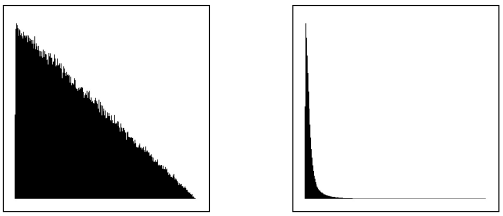
$$\forall x',\ d(x, x') \leq 128,\ P(D > 128 \mid X = 128) = 0. \tag{14}$$

En considérant cette valeur de la probabilité, l'entropie théorique de la carte des distances dans le domaine chiffré est :

$$\begin{aligned} H^D_{\text{chiffre}} &= \left[-\sum_{i=1}^{l-1} \frac{2i}{l^2} \log_2\left(\frac{2i}{l^2}\right)\right] - \frac{1}{l} \log_2\left(\frac{1}{l}\right), \\ &\geq \log_2\left(\frac{l}{2}\right) \textit{ bpp}. \end{aligned} \tag{15}$$

Dans une image en clair, la distribution des valeurs des distances est semblable à une distribution géométrique, comme illustré fig. 2.b. Si une variable aléatoire D suit une loi géométrique de paramètre p, sa probabilité d'être égale à d est :

$$P(D = d) = (1 - p)^{d-1} p. \tag{16}$$



(a) (b)

Figure 2 – *Histogrammes de la carte des distances : a) Dans une image chiffrée, b) Dans une image en clair.*

En conséquence, la valeur théorique de l'entropie pour la distribution des distances dans le domaine clair est :

$$\begin{aligned} H^D_{\text{clair}} &= -\sum_{i=1}^{l-1} \left((1-p)^{i-1} p\right) \log_2\left((1-p)^{i-1} p\right), \\ &\leq \log_2\left(\frac{l}{2}\right) \textit{ bpp}. \end{aligned} \tag{17}$$

D'après l'eq. (15) et l'eq. (17), on a :

$$H^D_{\text{clair}} < H^D_{\text{chiffre}}. \tag{18}$$

La valeur de l'entropie de la carte des distances dans le domaine clair est donc inférieure à celle mesurée dans le domaine chiffré.

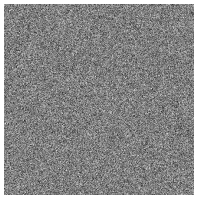
4 Résultats expérimentaux

Dans cette partie, nous décrivons les résultats expérimentaux obtenus pour illustrer notre analyse. Comme expliqué dans la partie précédente, la valeur de l'entropie dans le domaine clair est plus petite que dans le domaine chiffré. Nous pouvons alors utiliser ce résultat pour savoir si un bloc de pixels dans une image est bien déchiffré (*i.e.* en clair) ou encore chiffré (c'est-à-dire mal déchiffré). Cependant, il peut y avoir des cas d'erreur quand la valeur de l'entropie dans le domaine clair est supérieure à la valeur mesurée dans le domaine chiffré. En utilisant différentes tailles de bloc k et différents nombre de niveaux de gris l, nous évaluons le nombre d'erreurs avec l'entropie d'ordre zéro et celle de la carte des distances. Un exemple de l'utilisation de l'entropie pour corriger une image chiffrée bruitée est alors présenté.

4.1 Entropie d'ordre zéro

Nous commençons par mesurer l'entropie d'ordre zéro dans l'image de test Lena, de taille 512×512 pixels et codée sur 256 niveaux de gris, illustrée fig. 3.a. La fig. 3.b illustre l'image chiffrée correspondante en utilisant AES en mode ECB. Nous pouvons voir qu'il n'y pas d'ambiguïté à différencier l'image en clair de l'image chiffrée car la valeur globale de l'entropie est inférieure à celle de l'image chiffrée. Généralement, la valeur de l'entropie dans une image en clair avec 256 niveaux de gris est entre 6 *bpp* et 7.5 *bpp*. Dans le domaine chiffré, elle est très proche de 8 *bpp* (eq. (3)).

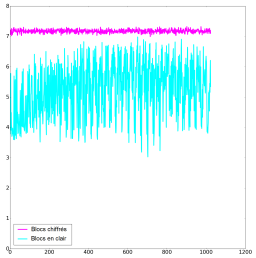
(a) Image originale $H_{\text{clair}} = 7.5940$ *bpp*.



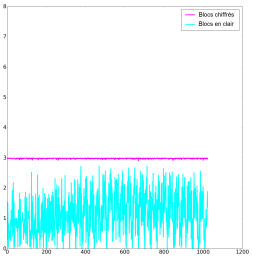
(b) Image chiffrée $H_{\text{chiffre}} = 7.9991$ *bpp*.

Figure 3 – *Mesure de l'entropie dans l'image de Lena et sa version chiffrée avec AES (images de taille* 512×512 *pixels et codées sur* 256 *niveaux de gris).*

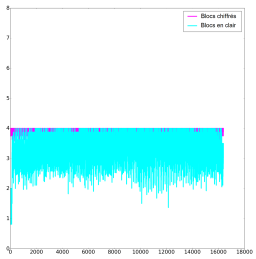
Nous proposons maintenant de mesurer l'entropie locale dans ces deux images, pour différentes tailles de bloc k. Les fig. 4.a–b illustrent la valeur de l'entropie dans chaque bloc de taille $k = 2^{2\times4} = 16 \times 16$ pixels dans l'image de Lena (en cyan) et sa version chiffrée (en magenta), pour $l = 256$ et $l = 8$ niveaux de gris. Si nous comparons les valeurs mesurées dans le domaine clair avec celles du domaine chiffré, nous pouvons voir qu'il n'y a pas d'erreur, quelque soit le nombre de niveaux de gris l considéré car la taille de bloc est suffisamment grande. Dans la fig. 4.c–d, la même analyse est conduite sur des blocs de plus petite taille, $k = 2^{2\times2} = 4 \times 4$ pixels. Dans ce cas, il y a de nombreuses erreurs si nous considérons $l = 256$ niveaux de gris car le nombre de niveaux de gris est trop élevé par rapport à la taille des blocs : la distribution des niveaux de gris est éparse. Cependant, nous pouvons voir que si le nombre de niveaux de gris est réduit, il est plus facile de faire la distinction entre les blocs en clair et les blocs chiffrés, comme illustré fig. 4.d pour $l = 8$, mais toutes les erreurs ne sont tout de même pas corrigées.



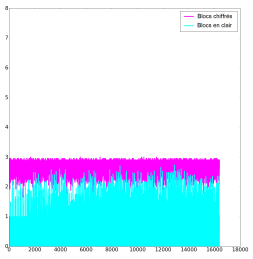
(a) $k = 16\times16, l = 2^8 = 256$.



(b) $k = 16 \times 16, l = 2^2 = 8$.



(c) $k = 4 \times 4, l = 2^8 = 256$.

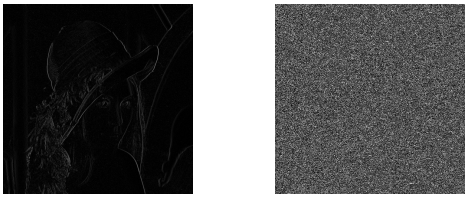


(d) $k = 4 \times 4, l = 2^2 = 8$.

Figure 4 – *Comparaison des valeurs de l'entropie locale* $H_{(k,l)}$ *(blocs de taille* $k = 16 \times 16$ *et* $k = 4 \times 4$ *pixels) suivant le nombre* l *de niveaux de gris considéres dans l'image originale de Lena (cyan) et sa version chiffrée (magenta).*

4.2 Entropie de la carte des distances

Dans l'idée d'exploiter la corrélation importante entre les pixels du même voisinage dans le domaine clair, sachant que cette corrélation est beaucoup plus faible dans le domaine chiffré, nous construisons la carte des distances. La fig. 5.a correspond à la carte des distances (dans le sens horizontal) de l'image de Lena en clair (fig. 3.a) et la fig. 5.b est celle de l'image chiffrée correspondante (fig. 3.b). Dans la fig. 3.a, nous pouvons voir qu'il y a de nombreux pixels sombres correspondant à une valeur de la distance entre deux pixels voisins proche de zéro. En d'autres termes, les valeurs des pixels voisins sont fortement corrélées. A l'inverse, la fig. 5.b semble avoir été pseudo-aléatoirement générée.



(a) (b)

Figure 5 – *Carte des distances pour : a) L'image originale de Lena et b) Sa version chiffrée.*

Dans la fig. 6, nous considérons des blocs de 4×4 pixels et comparons le pourcentage d'erreurs (*i.e.* quand les blocs de l'image en clair ont une valeur de l'entropie plus élevée que dans le domaine chiffré) obtenu en utilisant l'entropie d'ordre zéro (en bleu) avec celui associé à la mesure de l'entropie de la carte des distances (en rouge). Nous pouvons voir que quelque soit le nombre de niveaux de gris considéré lors du calcul, le nombre d'erreurs est toujours plus faible avec l'entropie de la carte des distances qu'avec l'entropie d'ordre zéro. En particulier, avec l'entropie de la carte des distances, le nombre d'erreurs est proche de zéro entre 2^2 et 2^6 niveaux de gris. Cela met en évidence l'importance de prendre en compte la corrélation entre les pixels dans le domaine clair.

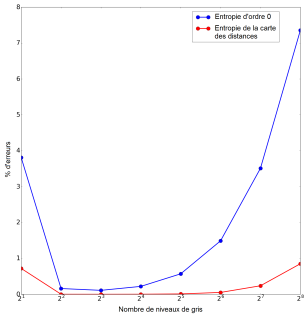


Figure 6 – *Comparaison entre le pourcentage d'erreurs en utilisant l'entropie d'ordre zéro et l'entropie de la carte des distances pour des blocs de taille $k = 4 \times 4$ en faisant varier le nombre l de niveaux de gris considérés lors du calcul (moyenne sur les blocs de 1000 images choisies aléatoirement dans la base BOWS-2 [10]).*

Le tableau 1 présente le nombre maximal de niveaux de gris l à considérer lors du calcul de l'entropie en fonction de la taille de bloc k considérée. Quand la taille du bloc est plus grande que le nombre de niveaux de gris de l'image (*i.e.* 256), il n'est pas nécessaire de quantifier l'image. Dans le cas contraire, quand il y a plus de niveaux de gris que de pixels dans le bloc, la distribution des niveaux de gris est éparse. Le meilleur compromis entre le nombre de niveaux de gris et la taille de bloc consiste à choisir un nombre de niveaux de gris inférieur à la taille du bloc. En effet, la meilleure quantification consiste à avoir des blocs dans le domaine clair relativement homogènes et des niveaux de gris uniformément distribués dans le domaine chiffré.

Taille des blocs k (pixels)	2×2	4×4	8×8	16×16	$\geq 32 \times 32$
Nombre max. de niveaux de gris l	8	8	16	64	256
% des erreurs (ordre zéro)	4.7856	0.1066	0.0017	0	0
% des erreurs (distances)	4.0589	0.0012	0	0	0

Tableau 1 – *Nombre maximal de niveaux de gris l à considérer pour minimiser le nombre d'erreurs en fonction de la taille des blocs k et pourcentage d'erreurs associé en utilisant l'entropie d'ordre zéro et l'entropie de la carte des distances (moyenne sur 1000 images de taille 512×512 choisies aléatoirement dans la base BOWS-2 [10]).*

4.3 Application : correction d'images chiffrées bruitées

Dans le but de présenter une application à notre analyse, dans l'expérience suivante nous utilisons l'entropie pour supprimer les erreurs contenues dans une image chiffrée bruitée.
L'algorithme 1 présente les étapes à suivre pour corriger une image chiffrée bruitée :

Algorithme 1 : Algorithme de correction des erreurs.

Données : Image chiffrée bruitée I_{ne} de taille $m \times n$ pixels
Taille de bloc k et nombre de niveaux de gris l
Résultat : Image reconstruite I de taille $m \times n$ pixels
pour chaque $B \in I_{ne}$, B de taille k **faire**
  $H_{\text{min}} \leftarrow H_{(k,l)}(B)$;
  /* Initialisation de la valeur minimale de l'entropie */
  $B_{\text{clair}} \leftarrow D_{\text{AES}}(B)$;
  /* Initialisation de la valeur en clair de B */
  pour $i \leftarrow 0$ **à** $k-1$ **faire**
    pour $j \leftarrow 0$ **à** 7 **faire**
      pour $\alpha \leftarrow 0$ **à** 1 **faire**
        /* Recherche de la valeur qui minimise l'entropie */
        si $H_{(k,l)}(D_{AES}(B_{p(i)_{b(j)=\alpha}})) < H_{min}$ **alors**
          $H_{\text{min}} \leftarrow H_{(k,l)}(D_{\text{AES}}(B_{p(i)_{b(j)=\alpha}}))$;
          $B_{\text{clair}} \leftarrow D_{\text{AES}}(B_{p(i)_{b(j)=\alpha}})$;
/* Tous les blocs B_{clair} de I ont été reconstruits */

Nous avons appliqué notre méthode à l'image de Lena (fig. 3.a) de taille 512×512 pixels et codée sur 256 niveaux de gris. Cette image a été chiffrée avec AES, en mode ECB, sur des blocs de taille 4×4 pixels (fig. 3.b). Pendant sa transmission, l'image chiffrée a été bruitée aléatoirement, avec un BER de 2.6×10^{-3}, ce qui affecte aléatoirement environ un bit d'un pixel tous les trois blocs (fig. 7.a). Comme illustré dans la fig. 7.b, si nous déchiffrons directement cette image sans effectuer de correction, certains blocs de l'image originale sont mal reconstruits. Dans la fig. 7.c, l'entropie d'ordre zéro est utilisée pour corriger les erreurs durant la phase de déchiffrement. Nous appliquons l'algorithme 1,

(a) Image chiffrée bruitée.

(b) Déchiffrement immédiat de l'image chiffrée bruitée, sans correction.

(c) Correction et déchiffrement en utilisant l'entropie d'ordre zéro.

(d) Correction et déchiffrement en utilisant l'entropie de la carte des distances.

Figure 7 – *Exemple d'application de l'utilisation de l'entropie pour corriger une image chiffrée bruitée (image originale présentée fig. 3.a et sa version chiffrée fig. 3.b , BER =* 2.6×10^{-3}*).*

avec des blocs de taille $k = 4 \times 4$ pixels et $l = 8$ niveaux de gris (valeur optimale d'après le tableau 1). Nous pouvons constater que la plupart des blocs faux ont été corrigés mais certains restent mal déchiffrés au niveau des contours. Cependant, si nous utilisons l'entropie de la carte des distances, comme présenté fig. 7.d, l'image originale est parfaitement reconstruite : notre approche de correction des erreurs permet de supprimer toutes les erreurs de transmission. Par ailleurs, si nous considérons des blocs de plus grande taille lors du calcul de l'entropie, le risque d'avoir plus d'un bit erroné par bloc devient non négligeable : dans ce cas, la méthode de l'algorithme 1 n'est pas efficace.

5 Conclusion

Dans ce travail de recherche, nous avons réalisé une analyse de l'utilisation de l'entropie de Shannon pour corriger des images chiffrées bruitées. Comme la valeur de l'entropie d'ordre zéro dans un bloc de pixels d'une image en clair est inférieure à celle mesurée dans le domaine chiffré, il est possible de savoir si un bloc est correctement déchiffré durant la phase de reconstruction de l'image. Cependant, il peut y avoir certains cas d'erreur lorsque l'entropie dans le domaine clair est plus grande que dans le domaine chiffré, en particulier lorsque nous considérons des blocs de très petite taille. Une première idée pour réduire le nombre d'erreurs consiste à adapter le nombre de niveaux de gris de l'image par quantification. De plus, nous avons pu observer une amélioration des résultats obtenus en calculant l'entropie de la carte des distances puisque, de cette façon, la corrélation des pixels voisins dans le domaine en clair est exploitée. Dans de futurs travaux, nous serons intéressés par réaliser la même analyse sur la mesure de l'entropie jointe pour considérer à la fois la valeur des pixels et celle des distances dans le but de réduire le nombre de cas d'erreur. De plus, nous pourrons étendre notre approche à la correction d'images couleur et de vidéos chiffrées bruitées.

Références

[1] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 :379–423, 1948.

[2] Wade Trappe et Lawrence C Washington. *Introduction to cryptography with coding theory*. Pearson Education India.

[3] Xianjun Hu, Weiming Zhang, Honggang Hu, et Nenghai Yu. Non-local denoising in encrypted images. Dans *International Conference on Internet of Vehicles*, pages 386–395. Springer, 2014.

[4] Sayed M SaghaianNejadEsfahani, Ying Luo, et S-c S Cheung. Privacy protected image denoising with secret shares. Dans *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 253–256. IEEE, 2012.

[5] Ankita Lathey et Pradeep K Atrey. Image enhancement in encrypted domain over cloud. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(3) :38, 2015.

[6] Alberto Pedrouzo-Ulloa, Juan Ramón Troncoso-Pastoriza, et Fernando Pérez-González. Image denoising in the encrypted domain. Dans *Information Forensics and Security (WIFS), 2016 IEEE International Workshop on*, pages 1–6. IEEE, 2016.

[7] William Puech, Marc Chaumont, et Olivier Strauss. A reversible data hiding method for encrypted images. Dans *Electronic Imaging 2008*, pages 68191E–68191E. International Society for Optics and Photonics, 2008.

[8] Naveed Islam, Zafar Shahid, et William Puech. Denoising and error correction in noisy AES-encrypted images using statistical measures. *Image Commun.*, 41(C) :15–27, Février 2016.

[9] Yue Wu, Yicong Zhou, George Saveriades, Sos Agaian, Joseph P Noonan, et Premkumar Natarajan. Local shannon entropy measure with statistical tests for image randomness. *Information Sciences*, 222 :323–342, 2013.

[10] Patrick Bas et Teddy Furon. Image database of BOWS-2. http ://bows2.ec-lille.fr/.

Stéganographie et Stéganalyse des images JPEG Couleur

Papa Mamadou Ndiaye[3,4] Marc Chaumont[1,2] Mehdi Yedroudj[1] Ahmad Zakaria[1]

[1] UNIVERSITE MONTPELLIER, UMR5506-LIRMM, F-34095 Montpellier Cedex 5, France

[2] UNIVERSITE DE NIMES, F-30021 Nîmes Cedex 1, France

[3] ECOLE SUPERIEURE POLYTECHNIQUE DE DAKAR, 5005 Dakar - Fann, Sénégal

[4] CNRS, UMR5506-LIRMM, F-34392 Montpellier Cedex 5, France

{ndiaye, chaumont, yedroudj, zakaria}@lirmm.fr

Résumé

JPEG est aujourd'hui le format d'image le plus couramment utilisé pour l'échange d'images. Bien que cela en fasse un standard naturel pour la stéganographie moderne, il n'en demeure pas moins qu'il n'y a pas de contributions pour l'insertion dans des images JPEG en couleur. Les approches d'insertion considèrent en effet uniquement l'insertion dans une image JPEG en niveaux de gris, principalement parce que l'insertion dans des images en niveau de gris est déjà un problème difficile. Dans cet article, nous étudions, de manière pratique, la question de l'insertion dans une image JPEG couleur. La question principale consiste à déterminer comment doit être effectuée la répartition du message, c'est-à-dire des bits à insérer, entre les composantes de couleurs (Y, Cr, Cb) qui ont été quantifiées. Après avoir rappelé l'état de l'art, nous donnons de premiers résultats expérimentaux indiquant que l'insertion doit principalement être effectuée dans la composante de luminance.

Mots clefs

JPEG, Couleur, Stéganographie, Stéganalyse.

1 Introduction

La stéganographie est l'art de dissimuler des informations dans un support anodin et cela sans éveiller la suspicion d'une tierce personne. La stéganalyse est la discipline duale de la stéganographie et consiste à déceler une dissimulation de données dans le support. La stéganographie a été largement appliquée aux images numériques dont les images JPEG. Cependant, si d'importantes contributions ont été apportées en stéganographie et stéganalyse des images JPEG en niveaux de gris [1, 2, 3, 4, 5], rien n'existe pour autant en terme de stéganographie moderne d'image JPEG en couleur. L'insertion dans des images en couleur a été proposée récemment dans [6, 7, 8, 9]. Les auteurs proposent dans cet articles d'insérer les bits indépendamment dans chacun des canaux RGB, et ceci en coupant le message en trois parties de même taille. Ils notent que l'insertion est sous-optimale, c'est-à-dire que le résultat pourrait être plus « sûr » si l'insertion était faite en prenant en compte les trois canaux simultanément. Les auteurs remarquent également que l'insertion optimale ne donne pas les résultats escomptés en pratique.

Dans ce papier nous souhaitons également insérer dans les trois composantes (Y, Cb et Cr) d'une image JPEG mais il est évident que la proportion de bits à insérer dans chacun des canaux ne doit pas être égale. Dans ce document nous étudions donc l'impact (en termes de détectabilité par le stéganalyste) de l'insertion d'un message en fonction de la proportion insérée dans chacun des canaux. Cette étude préliminaire est basée sur une insertion via l'algorithme de l'état de l'art J-UNIWARD [1]. Cet algorithme établit une carte de coûts de détectabilité pour chacun des coefficients DCT quantifiés et utilise ensuite cette carte pour réaliser une insertion adaptative à travers l'utilisation d'un codage par STC [10]. La sécurité de nos schémas d'insertion dans des images JPEG couleur sera évaluée en extrayant des caractéristiques inspirées du Spatial-Color Rich Model [8] et en les fournissant à un classificateur d'ensemble [11]. Les caractèristiques inspirées du Spatial-Color Rich Model sont composées du SRMQ1 ainsi que de co-occurrences de résiduels obtenus à partir des composantes couleurs. Nous nous plaçons dans l'espace de couleurs YCbCr pour calculer ces caractéristiques. Nous présentons notre proposition dans la section 2 puis le protocole dans la section 3. En section 4 nous analysons et interprétons nos résultats.

2 Propositions

En confrontant plusieurs variantes du même schéma stéganographique qui différent uniquement de par la proportion du message inséré dans la luminance, nous tentons de déterminer la bonne proportion à répartir dans la luminance et les chrominances lors de l'insertion. Nous présentons dans cette section les aspects liés à l'insertion dont la stratégie de répartition du budget.

2.1 Insertion indépendante

Les approches de stéganographie couleur exploitent chaque composante de l'image en y insérant une certaine portion du message. Cette insertion peut être opérée au sein de chaque canal indépendamment des autres composantes en considérant qu'une image en couleur est une combinaison de trois images en niveaux de gris. Dans [7], Abdulrahman et al. adoptent cette stratégie de stéganographie couleur en la préférant à l'approche qui considére une unique composante (concaténation des trois canaux RGB) et en effectuant l'insertion. Cette dernière approche est en pratique plus détectable. Il pourrait également être légitime de penser qu'effectuer une synchronisation entre les canaux de couleurs permettrait d'accroître la sécurité des schémas stéganographiques. En s'inspirant des travaux dans des images en niveau de gris de Denemark et Fridrich [12], on pourrait en effet contraindre l'algorithme d'insertion à favoriser les changements de même direction sur les trois canaux YCbCr [1] . De la même façon, on pourrait s'inspirer de la synchronisation dans des images couleur RGB proposée dans l'algorithme CMD-C [13]. Malheureusement, des erreurs dans le protocole du papier CMD-C invalident les résultats obtenus ainsi que les conclusions. On peut penser que préserver la corrélation entre canaux afin d'augmenter la sécurité empirique des images couleurs pourrait être intéressant ; toutefois, la corrélation des composantes Y,Cr,Cb est très faible. La question de la synchronisation reste donc une question ouverte et elle ne sera pas traitée dans cet article.

Dans cette étude préliminaire, nous choisissons donc d'insérer indépendamment dans chacune des composantes et cela sans tenir compte des possibilités de synchronisation.

2.2 Répartition du budget

Nous souhaitons comparer la détectabilité d'une technique distribuant les bits du message sur les trois canaux couleurs Y, Cb et Cr, et celle d'une approche consistant à insérer intégralement le message dans la luminance. Nous choisissons l'algorithme J-UNIWARD 'niveau de gris' pour implémenter chacune des deux approches. On désignera par 'payload relatif' à une composante, le nombre de bits du message inséré dans une composante. Il est important de noter que les schémas stéganographiques JPEG insèrent les messages dans les coefficients DCT quantifiés. Les paramètres d'entrée de JUNIWARD sont le fichier image JPEG et un message binaire dont la taille est exprimée en bpnzac (bits par coefficient AC non nul). Pour une comparaison objective à budget constant, on n'utilisera pas l'unité bpnzac. En effet, le nombre de coefficients AC non nuls varie en fonction des images, ce qui fait que pour un même nombre de bpnzac inséré dans deux images, celles-ci contiendraient à priori des messages de tailles différentes.

Nous préférons donc travailler à budget total constant et donc utiliser l'unité bits par pixels (bpp). Ainsi, deux images stéganographiées avec le même nombre de bpp, contiendront exactement le même nombre de bits. Notons que lorsque l'insertion est habituellement effectuée en bpnzac avec des budgets entre 0.1 et 0.5 bpnzac. Pour rester dans la même gamme et donc étudier la sécurité dans une gamme de taille de message similaire, le budget doit être compris entre 0.005 bpp et 0.03 bpp. Par soucis de commodité, nous testerons 6 valeurs P de taille de message exprimés en bpp :

$$P \in \{0.005; 0.010; 0.015; 0.020; 0.025; 0.030\} \quad (1)$$

Pour chaque valeur de P, nous insérerons une proportion α dans la luminance et la proportion $\beta = \frac{1-\alpha}{2}$ dans chacune des chrominances. Nous testerons 6 valeurs différentes pour α :

$$\alpha \in \{85\%; 90\%; 93\%; 95\%97\%; 100\%\} \quad (2)$$

Pratiquement, pour un P fixé, le budget en nombre de bits est obtenu en multipliant P par le nombre de pixels de l'image. Ce budget est réparti entre les canaux de couleur en considérant que la proportion α est celle allouée à la luminance. La portion du budget restante est équitablement répartie entre les deux chrominances. On déduit ensuite le nombre de bpnzac à insérer dans chaque composante en faisant un rapport entre la portion de budget affectée à chaque composante et son nombre de coefficients AC non nuls. Cette manipulation permet d'utiliser J-UNIWARD et sa version simulée en lui passant une composante et la valeur de bpnzac.

3 Protocole Expérimental

3.1 Base d'images

Notre base d'images de couverture est construite à partir des 10000 images RAW de la BOSSBase 1.0 [6] en suivant les étapes suivantes :

— Conversion des images de la BOSSBase en 10000 images couleur 512x512 au format PPM en utilisant successivement les primitives ufraw et convert de ImageMagick. Le code source bash utilisé pour ces opérations est disponible sur `http://www.lirmm.fr/~chaumont/BOSSJPEG/macroProductPPM.sh` .

— Compression des images PPM en des images JPEG de facteur de qualité QF=75 en utilisant les primitives imwrite de Matlab.

Nous n'appliquons aucun sous-échantillonnage aux chrominances des images obtenues dans la mesure où nos recherches nous ont indiqué que le format de sous-échantillonnage le plus représentatif est le 4 :4 :4. En effet, le tableau 1 établit le format de sous échantillonnage des chrominances opéré par trois réseaux sociaux importants en mettant en relief le nombre d'images téléchargées chaque jour. Les indications sur les formats de sortie ont été obtenues en publiant une image non compressée sur chaque réseau et en la récupérant par la suite. On dénote deux fois

1. direction positive : insertion +1 ; direction négative : insertion -1

Tableau 1 – *Compression des images sur les réseaux sociaux*

Réseaux sociaux	images publiées par jour	Format
Facebook	205 millions	4 :2 :0
Twitter	342 millions	4 :2 :0
Whatsapp	700 millions	4 :4 :4

plus de trafic d'images sur Whatsapp que sur Facebook et Twitter réunis. Pour cette raison, nous choisissons de calquer notre modèle de compression JPEG sur celui de Whatsapp. Nous disposons au terme de cette phase d'une base d'images de couverture que nous nommons "Cover75444".

3.2 Elements de stéganalyse

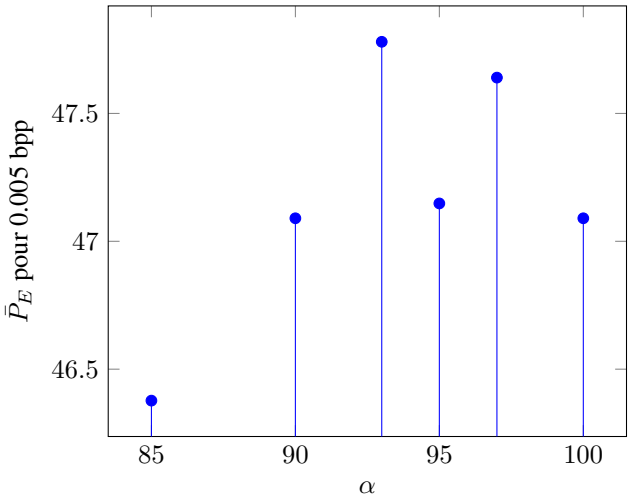
Pour évaluer les performances des schémas proposés, il est nécessaire de prendre en compte aussi bien l'aspect spatial que colorimétrique dans la phase de stéganalyse. Le Spatio-Color Rich Model (SCRMQ1) [3] a deux composantes principales qui prennent en compte cela.
La première composante est le Spatial Rich Model (SCRMQ1) [14] qui consiste en un vecteur de 12753 valeurs obtenues en calculant séparément sur chaque composante couleur des bruits résiduels via l'utilisation de nombreux filtres passe-haut, puis à quantifier les images obtenus avec un pas de quantification de 1. Ensuite les valeurs trop élevées sont tronquées (l'indice de troncature T est égal à 2). De là, des matrices de co-occurrences (horizontales et verticales) sont calculées sur chacune des images de bruits résiduels. Les trois matrices de co-occurrences sont fusionnées pour garder une même dimension que le vecteur d'origine, soit 12753 valeurs.
La seconde composante calculée dans le Color Rich Model est obtenue en nous basant sur les résiduels de chaque canal et en calculant des co-occurrences transversales. Il en résulte 5404 features (CRMQ1).
Ainsi, le Spatial Color Rich Model comprend 18157 features au total. Etant donné que la stéganographie JPEG s'opère dans un espace de couleur YCbCr, nous proposons une alternative au SCRM et que nous calculons directement à partir des composantes YCbCr. Le filtrage des résiduels ainsi que le calcul des cooccurrences s'effectuent de la même manière que pour le SCRMQ1 et les 18157 features résultantes seront appelées YCbCr-SCRM. Elles sont ensuite utilisées par le classificateur d'ensemble [11] pour la phase d'apprentissage et de test. Ce classificateur est formé par un ensemble de détecteurs binaires implémentés via le calcul des déterminants linéaires de Fisher et agissant chacun sur des portions réduites des caractéristiques de l'image. Nous utilisons le classificateur dans la version qui minimise la probabilité d'erreur de classification :

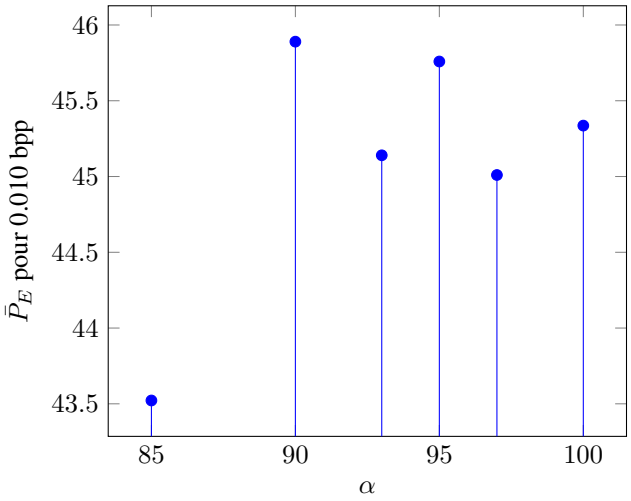
$$P_E = \min_{P_{FA}} \frac{1}{2}(P_{FA} + P_{MD})$$

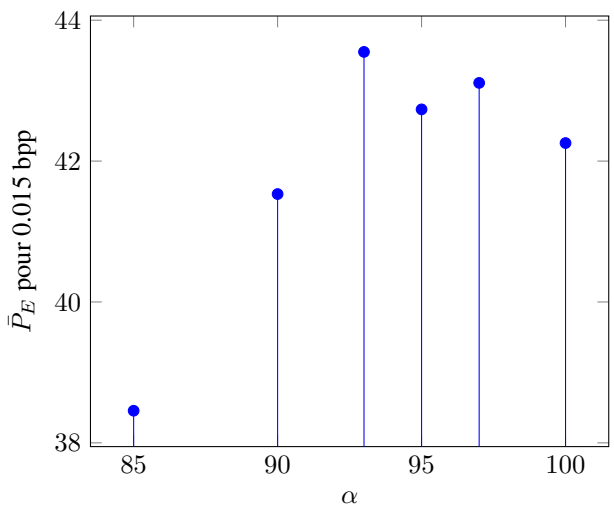
avec P_{FA} la probabilité de fausse alarme et P_{MD} la probabilité de détection ratée. On calcule la moyenne des probabilités d'erreurs P_E sur 10 différents scénarii d'apprentissage et de test au cours desquels les 10000 images de couverture et les 10000 images stéganographiées sont réparties en deux parties égales de telle sorte que chaque image stego soit associée à son image de couverture correspondante.

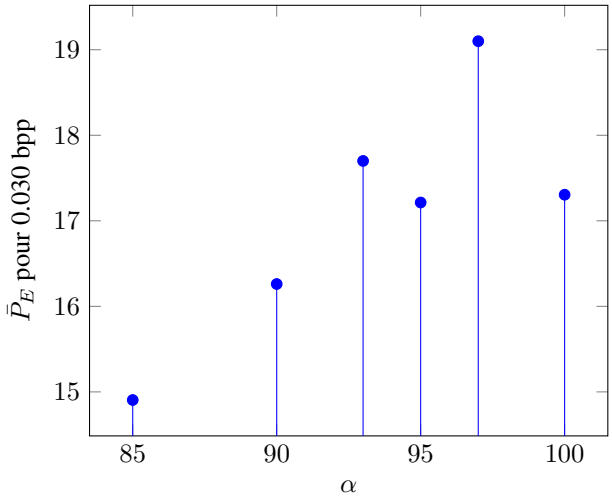
4 Résultats et discussion

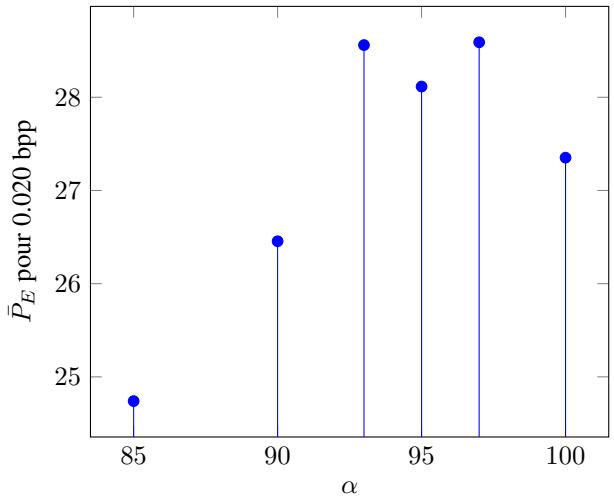
On présente dans cette partie les résultats des travaux menés en nous basant sur le protocole expérimental défini en section 2. Au total, six courbes correspondant chacune à un budget en bpp fixé (Eq. 1) et à une proportion α définie (Eq. 2). En abscisse, l'on retrouve les proportions de répartition du budget et en ordonnées la moyenne des probabilités d'erreur pour un payload particulier. Nous conviendrons de noter α* la proportion optimale de message à affecter à la luminance.











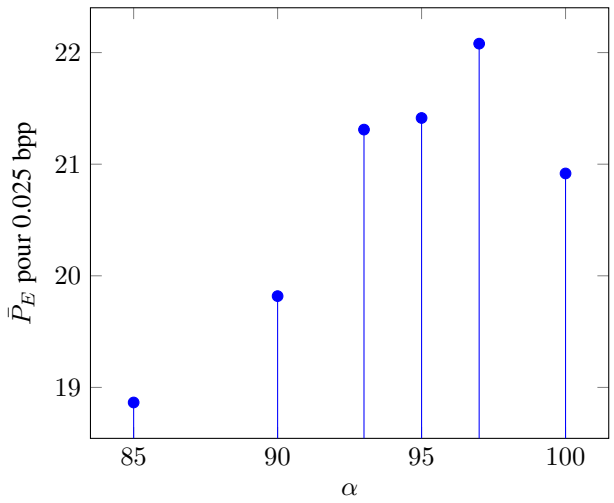


Figure 1 – *Détectabilite des variantes de JUNIWARD avec distribution du payload entre les canaux de couleur. Les courbes correspondent respectivement à une insertion de payload de 0.005 bpp, 0.010 bpp, 0.015 bpp, 0.020 bpp, 0.025 bpp et 0.030 bpp*

Nous pouvons tirer d'importantes observations des résultats présentés à la figure 1. De prime abord, les stratégies de distribution des payloads de 0.005 à 0.03 bpp montrent que l'insertion d'une portion du message dans les composantes de chrominances peut améliorer les performances du schéma stéganographique. La proportion adéquate α* à affecter à la luminance oscille entre 90% et 97%, rendant cornélien le choix de la variante de JUNIWARD à adopter. Nous observons d'autre part que les taux d'erreurs pour les payloads allant de 0.005 bpp à 0.015 bpp ne correspondent pas à des gammes d'erreur intéressantes, certaines probabilités de détection virant à l'aléatoire ($\approx 47\%$). En revanche, pour les autres payloads de 0.02 à 0.03 bpp, l'erreur de detection est maximale en 97%.
Ainsi, pour JUNIWARD à QF=75, une proportion α*=97% est un bon choix. L'insertion dans les chrominances fait donc gagner 1% en termes de sécurité. Ce gain est non négligeable (la variance étant de 10^{-6}) en stéganalyse.

5 Conclusion et perspectives

Au terme de cette étude, il apparaît qu'élaborer une stratégie de distribution du message entre les composantes couleurs dans le cadre de la stéganographie d'images JPEG couleur permet d'améliorer la sécurité du schéma d'environ 1% pour J-UNIWARD lorsque l'insertion est réalisée de maniére indépendante et à QF=75. La proportion de répartition du payload dans la luminance se situe autour de 97%. Toutefois, d'autres investigations prenant en compte d'autres facteurs de qualité, d'autres algorithmes d'insertion, d'autres bases, ainsi que d'autres approche de stéganalyse JPEG couleur nous permettront d'affiner nos conclusions.

Références

[1] Vojtech Holub et Jessica Fridrich. Digital image steganography using universal distortion. Dans *IH&MMSec 13 Proceedings of the first ACM workshop on Information hiding and multimedia security*, pages 59–68, Montpellier, France, Juin 2013.

[2] Linjie Guo, Jiangqun Ni, et Yun Qing Shi. Uniform embedding for efficient jpeg steganography. *IEEE Transactions on Information Forensics and Security*, 9(5) :814–825, Mai 2014.

[3] Xiaofeng Song, Fenlin Liu, Chunfang Yang, Xiangyang Luo, et Yi Zhang. Steganalysis of adaptive jpeg steganography using 2d gabor filters. Dans *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security*, pages 15–23, Portland, Oregon, USA, 2015.

[4] Vojtech Holub et Jessica Fridrich. Low complexity features for jpeg steganalysis using undecimated dct. *Information Forensics and Security, IEEE Transactions*, 10(2) :219–228, Juin 2015.

[5] Vojtech Holub et Jessica Fridrich. Phase-aware projection model for steganalysis of jpeg images. Dans *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics XVII*, pages 187–198, San Francisco, CA, Mars 2015.

[6] Hasan Abadulrahman, Marc Chaumont, Philippe Montesinos, et Baptiste Magnier. Color images steganalysis using rgb channel geometric transformation measures. *Wiley Journal on Security and Communication Networks (SCN) - Special Issue on Cyber Crime*, 9(15) :2945–2956, Février 2016.

[7] Hasan Abdulrahman, Marc Chaumont, Philippe Montesinos, et Baptiste Magnier. Color image steganalysis using correlations between rgb channels. Dans *Proceedings Int. Conf. Avail., Reliab., Security*, pages 448–454, Toulouse, France, Aout 2015.

[8] Miroslav Goljan, Jessica Fridrich, et Remi Cogranne. Rich model for steganalysis of color images. Dans *IEEE Int. Workshop Inf. Forensics Security*, pages 185–190, Atlanta, GA, USA, Decembre 2014.

[9] Miroslav Goljan et Jessica Fridrich. Cfa-aware features for steganalysis of color images. Dans *Proceedings SPIE 9409, Media Watermarking, Security, and Forensics*, page 94090V, San Francisco, Californie, Etats Unis, March 2015.

[10] Tomas Filler, Jan Judas, et Jessica Fridrich. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, 6(3) :920–935, Septembre 2011.

[11] Jan Kodovsky et Jessica Fridrich. Ensemble classifiers for steganalysis of digital media. *Information Forensics and Security, IEEE Transactions*, 7(2) :432–444, Juin 2012.

[12] Tomas Denemark et Jessica Fridrich. Improving steganographic security by synchronizing the selection channel. Dans *15th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–14, Portland, Oregon, USA, Juin 2015.

[13] Weixuan Tang, Bin Li, Weiqi Luo, et Jiwu Huang. Clustering steganographic modification directions for color components. *IEEE Signal Processing Letters*, 23(2) :197–201, Janvier 2016.

[14] Jessica Fridrich et Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3) :868–882, Juin 2012.

Chiffrement sélectif d'objets 3D

Sébastien BEUGNON[1,2] William PUECH[1] Jean-Pierre PEDEBOY[2]

[1] LIRMM, UMR 5506, Université de Montpellier
[2] STRATEGIES, Rungis, FRANCE
Mails : {sebastien.beugnon, william.puech}@lirmm.fr

Résumé

Contrairement aux méthodes de chiffrement classiques détruisant la structure interne des fichiers, ce papier présente une méthode de chiffrement sélective des objets 3D respectant les normes des formats de fichiers de maillage. Cette méthode chiffre des bits sélectionnés dans la géométrie du maillage pour protéger visuellement son contenu. Les distorsions géométriques générées permettent de cacher partiellement ou complètement le contenu, mais elles ne corrompent pas la scène 3D afin de rendre possible la lecture du document. Des résultats expérimentaux sont présentés et évalués afin de valider la méthode proposée.

Mots clefs

Chiffrement sélectif, Objet 3D, Visualisation sélective, Protection de contenu, Sécurité multimédia.

1 Introduction

Depuis ces dernières années, les contenus multimédia submergent Internet. Les modèles 3D deviennent de plus en plus utilisés pour de nombreuses applications comme la visualisation médicale, les outils de simulation, les jeux vidéo, les ventes en ligne, l'animation et les effets spéciaux au cinéma. De plus, aujourd'hui, les imprimantes 3D, permettant de fabriquer rapidement des objets pour le prototypage ou le divertissement, deviennent une tendance populaire très consommatrice de contenu 3D. Cette technologie est en train d'évoluer en une toute nouvelle économie [1]. Avec la démocratisation de l'impression 3D, des plateformes de téléchargement d'objets 3D se développent rapidement et l'utilisation de formats propriétaires devient très rapidement un frein à leur développement. Cependant, pour les créateurs de contenu, le besoin de gestion de leurs produits et droits se ressent cruellement. Car les objets 3D, au vu de leurs coûts de production, représentent des biens financiers pour leurs créateurs. Ces derniers souhaitent se protéger du piratage et des copies frauduleuses. Il existe deux catégories de méthodes pour protéger le contenu 3D : le chiffrement ou l'insertion de données cachées. Lors de ces dernières années, la littérature s'est concentrée principalement sur l'insertion 3D [2, 3, 4, 5]. L'insertion de données cachées rajoute des informations apportant de nouvelles fonctionnalités aux objets 3D telles que la détection de modifications, l'insertion de métadonnées de contenu ou la traçabilité. En parallèle, des méthodes de protection contre les accès non autorisés ont commencé à voir le jour [6, 7, 8, 9, 10]. L'approche traditionnelle du chiffrement ne prend pas en compte le contenu des données chiffrées. Ainsi, les images, le son, les vidéos ou les objets 3D sont traités comme des données binaires provoquant la destruction de la structure interne des fichiers lors du chiffrement. Koller *et al.* ont proposé de protéger les données en utilisant un système de rendu à distance [6], tandis que Cho *et al.* ont choisi de créer un tatouage 3D avec un chiffrement aléatoire des données de connectivité des objets 3D [7]. Gschwandtner et Uhl ont utilisé une représentation progressive de maillage possédant des couches de raffinement [8]. Cette structure est utilisée pour transmettre des objets 3D sur le réseau afin de profiter d'une prévisualisation en basse qualité. Ils ont choisi de chiffrer le contenu de ces couches transportant des parties de la géométrie, de la connectivité ou d'autres attributs des objets 3D. Plus tard, Éluard *et al.* ont présenté plusieurs méthodes de chiffrement préservant la géométrie utilisant des permutations de sommets ou de coordonnées pour protéger le contenu [9]. Leurs méthodes préservent certaines propriétés comme la boite englobante ou l'enveloppe convexe dans le but de minimiser le temps de rendu. Récemment, Yang et Zhang, avec une approche plus optique, ont proposé une méthode de chiffrement de nuage de points 3D par des images en niveau de gris et un patron de franges déformées en utilisant la transformation de Fresnel [10].

Dans ce papier, une nouvelle méthode de chiffrement des objets 3D est proposée. Cette approche est basée sur le chiffrement sélectif des bits des coordonnées des sommets. Grâce à ce chiffrement, préservant la structure des fichiers et se basant sur la représentation normalisée des valeurs flottantes, la méthode est capable de protéger les objets 3D en cachant partiellement ou complètement le contenu. Cette méthode permet aussi de choisir le niveau de chiffrement appliqué au maillage. En section 2, la méthode est présentée avec les spécifications nécessaires à sa compréhension. Puis, les résultats expérimentaux sont évalués en section 3 d'un point de vue statistique avec des métriques adaptées aux objets 3D et une analyse de la sécurité est réalisée en se basant sur la littérature. Enfin, nous concluons ce travail en section 4.

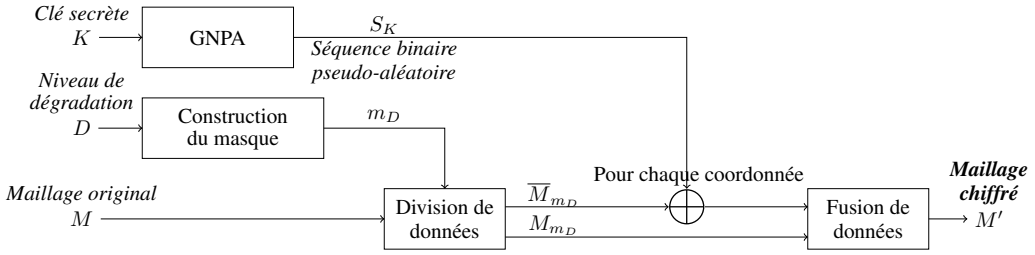


Figure 1 – *Vue d'ensemble de la méthode de chiffrement sélectif.*

2 Méthode proposée

Nous proposons une méthode chiffrant de manière sélective un objet 3D. La méthode permet de prendre le contrôle de l'impact géométrique du chiffrement en chiffrant sélectivement les bits des coordonnées des sommets du maillage. Comme illustré figure 1, notre méthode requiert 2 paramètres avec le maillage à chiffrer. Ces derniers sont la clé secrète K et le niveau de dégradation D. En section 2.1, nous présentons un résumé de la représentation binaire des valeurs flottantes utilisée dans les formats binaires d'objet 3D. Puis, les deux principales étapes constituant la méthode sont décrites. Ces dernières sont la sélection des données à chiffrer en section 2.2 et le chiffrement des coordonnées des sommets du maillage en section 2.3. Enfin, la section 2.4 présente l'étape de déchiffrement d'un objet 3D protégé par notre méthode.

2.1 Représentation de valeurs flottantes

Les coordonnées d'un sommet dans un maillage sont définies par des valeurs flottantes au sein des formats binaires de maillages par la norme *IEEE 754* [11]. Cette norme est la plus utilisée pour la représentation des valeurs sur les machines actuelles. Disposant de 32 bits, cette représentation contient 3 informations distinctes : le signe, l'exposant et la mantisse. Chaque information possède une certaine quantité de bits. Le bit le plus significatif (ou MSB) indique le signe de la valeur. Les 8 bits suivants représentent l'exposant, puis les 23 derniers bits correspondent à la mantisse. L'exposant et la mantisse permettent de représenter des valeurs absolues entre $1.175494e^{-38}$ et $3.402823e^{+38}$ avec une précision de 6 ou 7 chiffres significatifs après la virgule :

$$float = (-1)^s \times 2^{(e-127)} \times s.f. \quad (1)$$

2.2 Sélection des données à chiffrer

Cette étape comporte deux tâches spécifiques. Dans un premier temps, la méthode génère une séquence binaire pseudo-aléatoire S_K avec un générateur de nombres pseudo-aléatoires (GNPA) et la clé secrète K. Cette séquence sera utilisée pendant le chiffrement des coordonnées. Puis, dans un second temps, notre méthode construit un masque de chiffrement basé sur le niveau désiré de dégradation D. Cette variable définit la force du chiffrement sélectif. Plus le niveau de dégradation est bas, plus le maillage sera reconnaissable. Nous avons développé deux stratégies pour générer ce masque m_D. La première stratégie construit un masque, nommé masque D-LSB, avec une taille variable qui chiffre les D premiers bits de poids faible. La seconde stratégie, appelée masque D-SW, consiste à faire glisser une fenêtre sur les bits de la valeur flottante pour sélectionner ceux à chiffrer.

Masque D-LSB. Dans cette première approche, le niveau de dégradation indique le nombre de bits à chiffrer en commençant par celui de poids faible de la valeur flottante comme illustré figure 2. Dans ce masque, la valeur

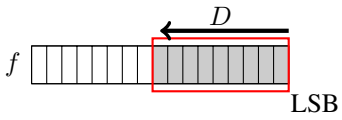


Figure 2 – *Sélection des bits par la méthode de chiffrement en fonction du niveau de dégradation D pour le masque D-LSB.*

de D est comprise dans l'intervalle $\{1, 31\}$ et à chaque incrémentation, le niveau de dégradation augmente de 1 le nombre de bits à chiffrer. Le chiffrement de ce bit supplémentaire accentue les distorsions géométriques au sein du maillage chiffré.

Masque D-SW. La seconde stratégie D-SW (ou *D-Sliding Window*) réutilise l'idée de chiffrer un certain nombre de bits. Cependant, au lieu de sélectionner seulement le nombre de bits à chiffrer D, le niveau de dégradation pour cette approche D' définit aussi la position p du premier bit du masque sur la valeur flottante où $p \in \{D, 31\}$. Le niveau de dégradation D' se note sous la forme de la paire : $D' =< p, D >$. Comme illustré en figure 3,

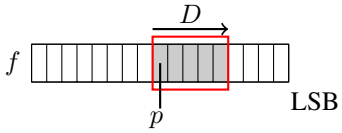


Figure 3 – *Sélection des bits par la méthode de chiffrement en fonction du niveau de dégradation D pour le masque D-SW.*

cette fenêtre glissante permet à l'utilisateur de sélectionner

plus précisément les bits à chiffrer. Il est possible avec le masque D-SW de générer un masque D-LSB en choisissant le niveau de dégradation D' correspondant à la paire $< p, D >$ avec $p = D$ où $D \in \{1, 31\}$.

2.3 Chiffrement sélectif des coordonnées de sommets

Durant le processus de chiffrement du maillage, illustré figure 1, pour chaque coordonnée de chaque sommet du maillage à chiffrer, la coordonnée courante est divisée en deux valeurs grâce au masque m_D, par exemple pour x_i :

— $x_{i_{m_D}}$, est la partie conservée de la coordonnée x_i telle que : $x_{i_{m_D}} = x_i \wedge \neg m_D$;

— $\overline{x_{i_{m_D}}}$, est la partie à chiffrer de la coordonnée x_i telle que : $\overline{x_{i_{m_D}}} = x_i \wedge m_D$.

En utilisant la séquence pseudo-aléatoire S_K générée précédemment par un GNPA, on extrait une valeur *nonce* r et le masque m_D est appliqué dessus :

$$r_{m_D} = r \wedge m_D. \tag{2}$$

La valeur intermédiaire $e = r_{m_D} \oplus \overline{x_{i_{m_D}}}$ est calculée. Ce résultat est ensuite fusionné avec $x_{i_{m_D}}$ pour donner la coordonnée chiffrée sélectivement x'_i tel que $x'_i = e \vee x_{i_{m_D}}$. Nous répétons le même processus pour y_i et z_i. Ces différentes opérations, peuvent être réduites à :

$$c'_i = ((c_i \oplus r) \wedge m_D) \vee (c_i \wedge \neg m_D), \tag{3}$$

où c_i est la variable à remplacer par les coordonnées du $i^{\text{ème}}$ sommet.

2.4 Déchiffrement des objets 3D

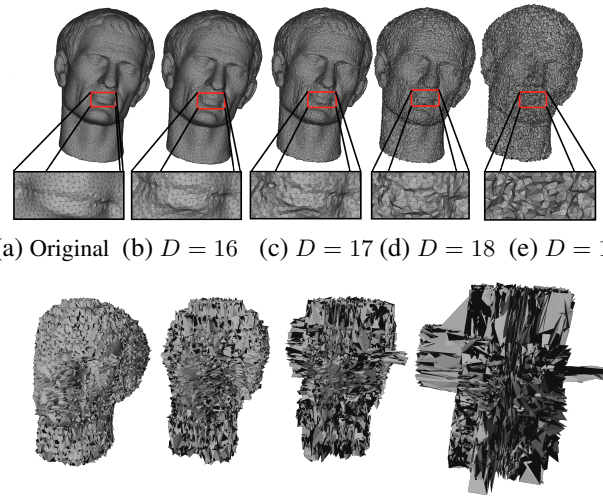
Pour déchiffrer l'objet 3D traité par notre méthode, il suffit d'appliquer à nouveau l'algorithme de chiffrement présenté en figure 1 sur l'objet avec la même clé secrète et le même niveau de dégradation D (ou D').

3 Résultats expérimentaux

Cette section présente les résultats expérimentaux de notre méthode de chiffrement sélectif d'objets 3D. Tout d'abord, en section 3.1, nous présentons une application de notre méthode sur un maillage pour les deux stratégies de construction de masque : masque D-LSB et masque D-SW. Puis, nous réalisons en section 3.2 une analyse statistique, une comparaison des résultats en fonction des masques et une analyse de la sécurité de notre méthode.

3.1 Application de notre méthode

La figure 4 illustre l'application de notre méthode de chiffrement avec un masque D-LSB pour différents niveaux de dégradation dans l'intervalle $\{16, 23\}$. La figure 4.a représente le maillage original, tandis que les figures 4.b-i sont des maillages chiffrés sélectivement avec un niveau de dégradation D spécifique. Nous observons que notre méthode permet de chiffrer modérément un objet 3D dans le but de

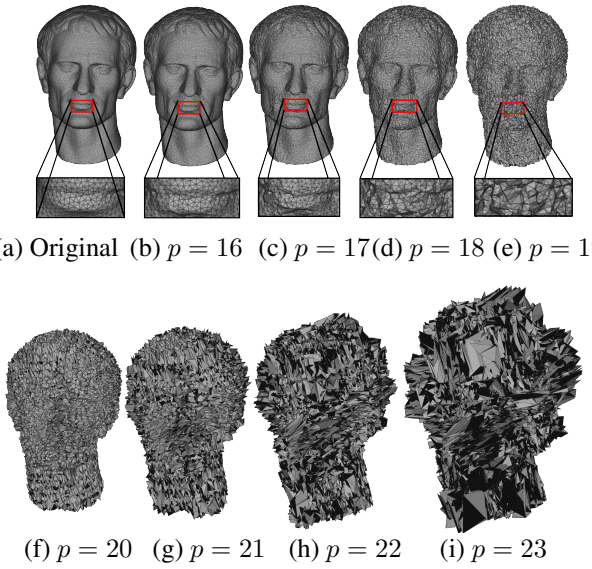


(a) Original (b) $D = 16$ (c) $D = 17$ (d) $D = 18$ (e) $D = 19$

(f) $D = 20$ (g) $D = 21$ (h) $D = 22$ (i) $D = 23$

Figure 4 – *Maillages chiffrés par notre méthode en fonction du niveau de dégradation D.*

le rendre plus ou moins reconnaissable, mais géométriquement impacté. Nous notons aussi qu'à partir du niveau de dégradation $D = 21$, il n'est plus possible de reconnaître le contenu de l'objet 3D, à partir du système visuel humain. La figure 5 présente notre méthode utilisant un masque D-SW pour $D' =< p, D >$ avec $D = 1$ et $p \in \{16, 23\}$, où p est la position du premier bit à chiffrer.



(a) Original (b) $p = 16$ (c) $p = 17$ (d) $p = 18$ (e) $p = 19$

(f) $p = 20$ (g) $p = 21$ (h) $p = 22$ (i) $p = 23$

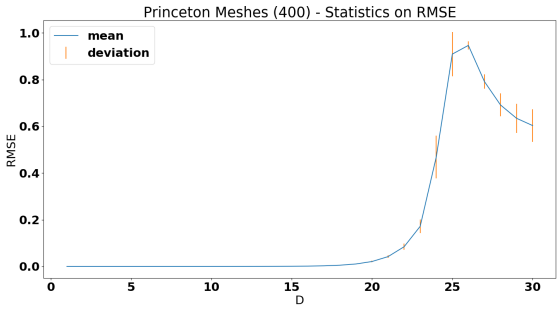
Figure 5 – *Maillages chiffrés par notre méthode en fonction du niveau de dégradation $D' =< p, 1 >$.*

Nous remarquons que notre méthode, avec cette stratégie qui consiste à chiffrer un seul bit, montre des résultats très similaires visuellement à ceux obtenus par l'approche précédente.

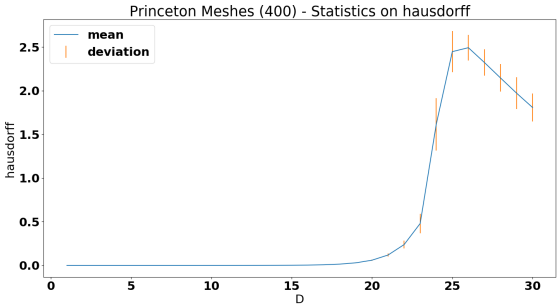
3.2 Analyse statistique

Pour cette expérimentation, nous comparons les maillages chiffrés pour les deux stratégies au maillage original. Nous avons utilisé 400 objets 3D de la base de données *Princeton Mesh Segmentation Project* [12]. Les métriques utilisées pour comparer les maillages sont la métrique *RMSE* et la *distance de Hausdorff* [13].

Masque D-LSB. Avec le masque D-LSB présenté en section 2.2, les résultats pour les métriques choisies sont illustrés en figures 6.a et 6.b. La métrique *RMSE* et la *distance de Hausdorff* montrent les mêmes dynamiques en figures 6.a et 6.b. Les deux métriques ont leurs valeurs proches de 0 quand $D \in \{1, 18\}$. *RMSE* et la *distance de Hausdorff* augmentent rapidement entre $D = 20$ et $D = 25$. La valeur moyenne de la *distance de Hausdorff* atteint un maximum autour de 2.5 à $D = 26$. Puis, cette valeur chute rapidement quand la méthode commence à chiffrer les bits de l'exposant. Le même comportement est visible pour la métrique *RMSE*.



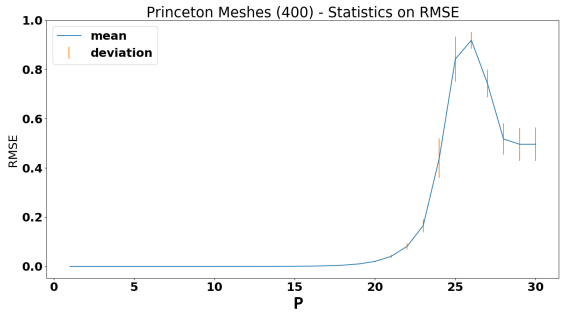
(a) RMSE



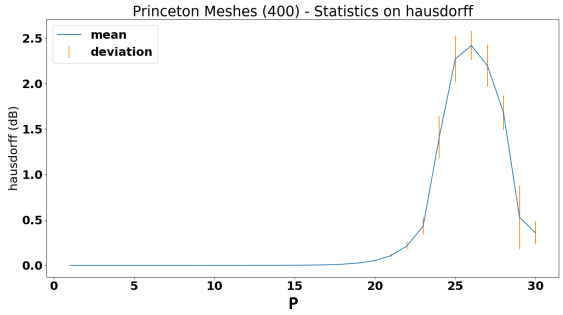
(b) Distance de Hausdorff

Figure 6 – *Résultats pour les métriques RMSE et distance de Hausdorff en fonction du niveau de dégradation D.*

Masque D-SW. Les figures 7.a et 7.b montrent les résultats pour les mêmes métriques avec le masque D-SW, avec les paramètres $D = 1$ et p, la position du premier bit à chiffrer dans l'intervalle $\{1, 31\}$. Tel qu'observé en figure 7, même si notre méthode chiffre qu'une petite partie de la géométrie de l'objet 3D, soit 3 bits sur les 96 que constitue un sommet, nous remarquons des similarités visuelles avec les résultats obtenus en figure 6 de la précédente stratégie.



(a) RMSE



(b) Distance de Hausdorff

Figure 7 – *Résultats pour les métriques RMSE et distance de Hausdorff en fonction du niveau de dégradation $D' =< p, 1 >$.*

Comparaison de stratégies. En utilisant un masque D-SW, nous remarquons l'importance du chiffrement des bits de poids significatif sur les impacts géométriques. Ces derniers sont par ailleurs identiques pour le système visuel humain lorsque les deux masques chiffrent le même bit de poids fort indépendamment du nombre de bits chiffrés.

3.3 Analyse de la sécurité

Dans cette section, nous discutons de la sécurité de notre méthode. Dans un premier temps, nous analysons la sensibilité du niveau de dégradation. Puis, nous rappelons la fragilité liée aux méthodes de chiffrement partiel face à des attaques par force brute ciblées. Nous présentons ensuite des attaques basées sur des algorithmes de traitement de maillages. Et enfin, une amélioration de notre méthode est proposée pour augmenter la complexité de notre méthode face aux attaques par force brute.

Sensibilité du niveau de dégradation. Quand un utilisateur essaie de déchiffrer un maillage avec la bonne clé secrète K, mais un mauvais niveau de dégradation D_{wrong}, l'objet 3D n'est pas révélé. Dans le cas d'un masque D-LSB, quand $D_{wrong} \leq D$, où D est le bon niveau de dégradation pour déchiffrer, D_{wrong} bits sont déchiffrés, mais pas les plus significatifs. Or, comme expliqué en section 3.2 avec un masque D-SW, le chiffrement d'un seul bit

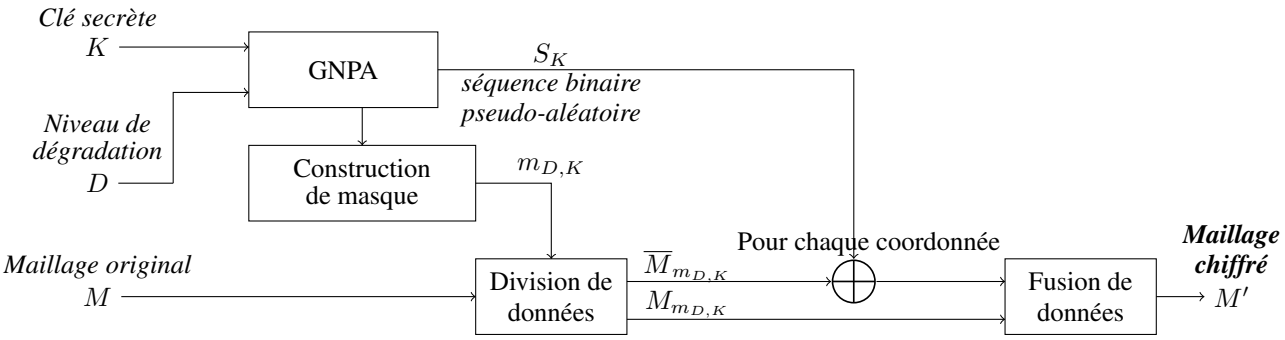


Figure 8 – *Vue d'ensemble de la méthode de chiffrement avec l'amélioration de complexité.*

est suffisant pour protéger le contenu autant qu'un masque D-LSB complet. De plus si $D_{wrong} \geq D$, le masque D-LSB est déchiffré, mais $D_{wrong} - D$ bits plus significatifs sont chiffrés rendant l'objet 3D encore plus méconnaissable. La même chose se produit pour le masque D-SW.

Fragilité des méthodes de chiffrement partiel. En tant que méthode de chiffrement partiel, la quantité de bits chiffrés est bien inférieure à celle d'un chiffrement classique. Cela rend la méthode plus rapide pour le chiffrement et le déchiffrement, mais aussi sensible à des attaques. Au lieu de chercher la clé secrète K, un attaquant va chercher à reconstruire le contenu chiffré. Attaquer naïvement notre méthode, requiert de trouver les bonnes valeurs pour les bits chiffrés dans chaque coordonnée de chaque sommet du maillage. Ceci correspond à trouver la bonne combinaison parmi $2^{3\times N\times D}$.

Comme expliqué dans la littérature [14], l'auteur montre clairement la faiblesse de ce genre d'approche de chiffrement. Il propose un système de cryptanalyse des images chiffrées sélectivement, utilisant les informations autour de celles chiffrées. Si nous prenons en considération son approche pour notre méthode en utilisant le niveau de dégradation ou la topologie du maillage, un attaquant peut imaginer des heuristiques guidant le choix des combinaisons. Ainsi, au lieu de chercher la combinaison globale des bits, on peut chercher à calculer la bonne combinaison pour le premier bit de chaque coordonnée pour réduire le niveau de dégradation, soit parmi $2^{3\times N}$, où N est le nombre de sommets. Évidemment, réaliser une telle attaque demande un investissement en temps et puissance de calcul de plus en plus important en fonction du nombre de sommets.

Attaques par traitement de maillage. Une autre attaque possible est d'essayer de traiter directement l'objet 3D chiffré. Comme notre méthode peut générer des maillages laissant le contenu reconnaissable, cela la rend sensible à des attaques à base de lissage ou de reconstruction. Ainsi, pour certains niveaux de dégradation où D ou $p \in \{1, 19\}$, un lissage laplacien [15] peut récupérer un maillage suffisamment proche de l'original. Bien que reconnaissables, les maillages lissés perdent énormément d'informations sur la courbure de l'objet original. De plus, lorsque ces attaques sont portées sur des maillages chiffrés avec un niveau de dégradation $D \in \{20, 29\}$, le maillage lissé diverge comme illustré figure 9. La *distance d'Hausdorff* révèle aussi cette information ; en effet le maillage lissé à une distance d'environs 8.416×10^{-2} au maillage original, tandis que le chiffré a une valeur 1.689×10^{-2} pour cette même métrique. Des méthodes de reconstruction peuvent aussi être utilisées pour récupérer le maillage secret. Cependant, notre méthode ne préservant pas les propriétés comme la boîte englobante ou l'enveloppe convexe, des reconstructions comme le Marching Cube [16] échouent quand le niveau de dégradation D est élevé comme $D \in \{20, 29\}$.

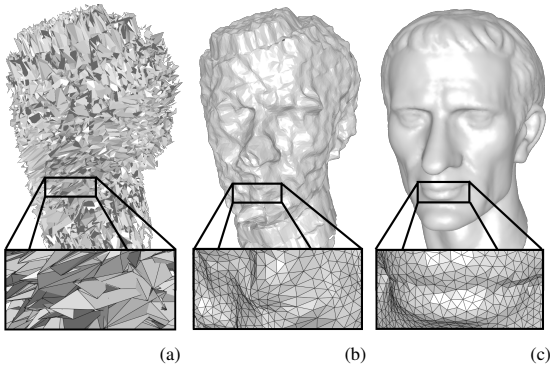


Figure 9 – *Attaque par lissage laplacien : a) Maillage chiffré avec $D = 21$, b) Résultat après lissage ($\lambda = 0.3$ et 100 itérations), c) Maillage original.*

Amélioration de la complexité. Dans le but d'améliorer la complexité de notre méthode, nous proposons une modification où la méthode génère pseudo-aléatoirement un masque D-SW pour chaque coordonnée du maillage. Comme illustré figure 8, au lieu d'utiliser le même masque, l'utilisateur choisit s'il souhaite que le maillage chiffré laisse le contenu reconnaissable ou non. Puis, avec le GNPA, la méthode construit pseudo-aléatoirement un ensemble de niveaux de dégradation $S_{K,D'}$ qui sont attribués à chaque coordonnée. Le nombre de bits à chiffrer varie entre 1 et p dans chaque masque. Quand l'utilisateur choisit de laisser reconnaissable le maillage à chiffrer, la mé-

thode utilise l'intervalle $\{17, 21\}$ pour p et quand il souhaite cacher complètement le contenu $p \in \{22, 26\}$. Avec cette amélioration, la complexité pour reconstruire le secret a augmenté. Si l'attaquant connait l'intervalle utilisé pour cacher le contenu, chaque coordonnée a une probabilité d'être trouvée de $\frac{1}{(5\times 2^{D_i})^N}$, où D_i est le nombre de bits chiffrés pour la $i^{\text{ème}}$ coordonnée.

4 Conclusion

Dans ce papier, nous proposons une méthode de chiffrement préservant le format qui protège sélectivement un objet 3D en chiffrant partiellement les bits des coordonnées des sommets en utilisant une clé secrète K. Notre méthode permet de choisir un niveau de dégradation donnant le contrôle des distorsions géométriques créées sur le maillage. Les maillages chiffrés peuvent être affichés dans des scènes 3D, car la structure interne des fichiers d'objet 3D a été préservée. Deux stratégies ont été proposées pour sélectionner les bits à chiffrer à savoir le masque D-LSB sélectionnant les D bits de poids faible, ou bien le masque D-SW utilisant une fenêtre glissante pour choisir les bits à chiffrer. Les résultats calculés pour les deux stratégies montrent qu'en chiffrant le même bit de poids fort, nous obtenons des maillages identiques en terme de qualité pour le système visuel humain. Chacun de ces masques apporte des propriétés utiles, une plus grande complexité calculatoire pour le masque D-LSB, tandis que le masque D-SW peut produire le même résultat visuel de chiffrement en chiffrant un seul bit par coordonnée, soit $3,125\%$ de la géométrie du maillage. Cette méthode se basant sur la norme standard des valeurs flottantes, peut être adaptée pour n'importe quelle précision du standard utilisé dans les formats binaires d'objets 3D. Nous présentons aussi le niveau de sécurité de notre méthode face à diverses attaques et proposons une amélioration pour augmenter sa complexité face à des attaques par force brute. La primitive cryptographique employée ici peut être remplacée par une autre.

Références

[1] M Baumers, P Dickens, C Tuck, et R Hague. The cost of additive manufacturing : machine productivity, economies of scale and technology-push. 102 :193 – 201.

[2] Jean-Luc Dugelay, Atilla Baskurt, et Mohamed Daoudi. *3D Object Processing : Compression, Indexing and Watermarking*. John Wiley & Sons.

[3] Vincent Itier, William Puech, et J.-P. Pedeboy. High-capacity data-hiding for 3d meshes based on static arithmetic coding. Dans *IEEE International Conference on Image Processing (ICIP), 2015*, pages 4575–4579. IEEE.

[4] Kai Wang, Guillaume Lavoué, Florence Denis, et Atilla Baskurt. A comprehensive survey on three-dimensional mesh watermarking. *IEEE Transactions on Multimedia*, 10(8) :1513–1527, 2008.

[5] T. Harte et A. G. Bors. Watermarking 3d models. Dans *Proceedings. International Conference on Image Processing*, volume 3, pages 661–664 vol.3.

[6] David Koller, Michael Turitzin, Marc Levoy, Marco Tarini, Giuseppe Croccia, Paolo Cignoni, et Roberto Scopigno. Protected interactive 3d graphics via remote rendering. Dans *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, pages 695–703. ACM.

[7] Misung Cho, Seokryul Kim, Maenghee Sung, et Giwon On. 3d fingerprinting and encryption principle for collaboration. Dans *Automated Production of Cross Media Content for Multi-Channel Distribution, 2006. AXMEDIS'06. Second International Conference on*, pages 121–127. IEEE.

[8] Michael Gschwandtner et Andreas Uhl. *Protected Progressive Meshes*, pages 35–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[9] M. Éluard, Y. Maetz, et G. Doërr. Impact of geometry-preserving encryption on rendering time. Dans *IEEE International Conference on Image Processing (ICIP),2014*, pages 4787–4791. IEEE, 2014.

[10] Xin Yang et Hongbo Zhang. Encryption of 3d point cloud object with deformed fringe. 2016.

[11] IEEE. Ieee standard for floating-point arithmetic. *IEEE Std 754-2008*, pages 1–70, Aug 2008.

[12] Xiaobai Chen, Aleksey Golovinskiy, et Thomas Funkhouser. A benchmark for 3D mesh segmentation. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), Août 2009.

[13] N. Aspert, D. Santa-Cruz, et T. Ebrahimi. Mesh : Measuring errors between surfaces using the hausdorff distance. Dans *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 705–708. IEEE, 2002.

[14] Amir Said. Measuring the strength of partial encryption schemes. Dans *IEEE International Conference on Image Processing (ICIP), 2005*, volume 2, pages II–1126. IEEE.

[15] L.R. Herrmann. Laplacian-isoparametric grid generation scheme. *Journal of the Engineering Mechanics Division*, 102(5) :749–907, 1976.

[16] William E. Lorensen et Harvey E. Cline. Marching cubes : A high resolution 3d surface construction algorithm. *SIGGRAPH Computer Graphics*, 21(4) :163–169, Août 1987.

Liste des auteurs