18éme édition

COmpression et REpresentation des Signaux Audiovisuels

CORESA'16

19 - 20 mai 2016

NANCY

UNIVERSITÉ DE LORRAINE

CRAN

cnrs

isis

jeudi 19 mai 2016

9h	Accueil - ouverture de la 18éme édition de CORESA
9h15	**Session orale – Reconnaissance** – ***Chairman : M. Yann GAUDEAU*** Oral 1 : Etude de la dynamique du visage en situation d'interaction naturelle – **B. Allaert, J. Menesso, N.M Bilasco, C. Djeraba** Oral 2 : Fusion et biométrie douce pour la dynamique de frappe au clavier – **Z. Syed Idrus, E. Cherrier, C. Rosenberger** Oral 3 : Analyse d'empreintes digitales à partir de paramètres structurels calculés sur une référence réduite de l'image – **B. Vibert, J.M. Le Bars, C. Charrier, C. Rosenberger**
10h15	Pause café
10h45	**Session orale - Image/Video** - ***Chairman : M. Florent DUPONT*** Oral 4 : Réseaux de neurones convolutionnels profonds pour la reconnaissance d'action dans les vidéos – **O. Seddati, S. Dupont, S. Mahmoudi** Oral 5 : Étude des réseaux de neurones sur la stéganalyse - **L. Pibre, M. Chaumont, D. Ienco, J. Pasquet** Oral 6 : Schéma de compression d'images intégrales basé sur l'extraction de vues - **A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, F. Dufaux** Oral 7 : Dans l'approximation de calcul pour la réduction de la complexité du codage et décodage des données multimédia pour des applications mobiles – **M. Jridi, A. Alfalou**
12h15	Repas
14h00	**Conférence invitée** : "Mathematical Microscopy" : **Pr. Jean-Christophe Olivo-Marin,** Institut Pasteur
15h00	Pause café
15h30	**Session orale – Vidéo** - ***Chairman : M. William PUECH*** Oral 8 : A Novel Video Coding Framework Based on Machine Learning - **D-K. Vo-Nguyen, M. Antonini and J. Jung** Oral 9 : Perceived audiovisual quality for videophone applications – **I. Saidi , L. Zhang, O. Deforges, V. Barriac** Oral 10 : Recherche Rapide du Mode de Prédiction Optimal basée Apprentissage via une Comparaison des Modes Intra/Inter en H.264/AVC - **M. Bichon, J. Le Tanoua, W. Hamidoucheb** Oral 11 : Performances de compression HEVC et H.264 dans le contexte du temps réel : Application en Télémédecine – **A. Chaabouni, Y. Gaudeau, J-M Moureaux**
18h15	Visite guidée de Nancy – Rendez-vous place Stanislas devant l'office du tourisme
20h00	Dîner de Gala dans les Grands salons de l'Hôtel de Ville de Nancy – Place Stanislas

vendredi 20 mai 2016

9h	**Conférence invitée** : “Musculoskeletal wide detector CT: principles, techniques and applications in clinical practice and research” : **Pr. Alain Blum,** CHRU de Nancy
10h	Pause café
10h30	Session orale - Médical - ***Chairman : M. Jean-Marie MOUREAUX*** Oral 12 : L’analyse de la voix comme outil de diagnostic précoce de la maladie de Parkinson : état de l’art - **L. Jeancolas, H. Benali, B.-E. Benkelfat, D. Petrovska-Delacrétaz** Oral 13 : Identification du scanner X à partir d’empreintes du capteur – **A. Kharboutly, W. Puech, G. Subsol, D. Hoa** Oral 14 : Evaluation par observateur numérique basé tâche de la qualité d’IRM compressées - **C. Cavaro-Menard, M. Schmidt, L. Zhang-Ge, J-Y Tanguy, P. Le Callet** Oral 15 : Evaluation statistique de la segmentation manuelle de données IRM de gliomes diffus de bas grade - **M. Ben Abdallah, M. Blonski, S. Mézières, Y. Gaudeau, L. Taillandier, J.M. Moureaux** Oral 16 : Modèles prédictifs pour les gliomes diffus de bas grade sous chimiothérapie - **M. Ben Abdallah, M. Blonski, S. Mézières, Y. Gaudeau, L. Taillandier, J.M. Moureaux** Oral 17 : Réduction de bruit multiplicatif dans les images ultrasons basée sur la Décomposition Multiplicative Multiresolution (MMD) - **M. Outtas, A. Serir, O. Deforges, L. Zhang**
12h30	Repas
14h00	Session orale - 3D - ***Chairman : Mme Luce MORIN*** Oral 18 : Reconnaissance d’actions basée sur la moyenne des articulations 3D issues des caméras RGB-D **- A. Ben Tamou, L. Ballihi, D. Aboutajdin** Oral 19 : Saillance visuelle multi-échelles des maillages 3D colorés **– A. Nouri, C. Charrier, O. Lézoray** Oral 20 : Segmentation de maillages 3D de pièces manufacturées numérisées : Application à la rétro-conception - **S. Gauthier, R. Bénière, W. Puech, G. Subsol** Oral 21 : Analyse de la symétrie en 4D, le cas de la paralysie facial – **P-A Desrosiers, B. Ben Amor, M. Daoudi, Y. Bennis, P. Guerreschi**
15h30	Clôture de la 18ème édition de CORESA

Étude des réseaux de neurones sur la stéganalyse

L. Pibre[1,3] M. Chaumont[1,2,3] D. Ienco[1,4] J. Pasquet[1,3]

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

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

[3] CNRS, UMR5506-LIRMM, F-34392 Montpellier Cedex 5, France

[4] IRSTEA, UMR TETIS, F-34093 Montpellier, France

{pibre, chaumont, pasquet, ienco}@lirmm.fr

Résumé

Des travaux récents ont montré que les réseaux de neurones ont un fort potentiel dans le domaine de la stéganalyse. L'avantage d'utiliser ce type d'architecture, en plus d'être robuste, est que le réseau apprend les vecteurs caractéristiques de manière automatique grâce aux couches de convolution. On peut dire qu'il crée des filtres intelligents. Dans cet article nous étudions le deep learning dans le domaine de la stéganalyse afin d'avoir une meilleure compréhension de son fonctionnement. Dans ce document nous présentons les travaux que nous avons effectués sur les réseaux de neurones convolutionnels. Tout d'abord, nous expliquons les aspects théoriques des réseaux de neurones, puis nous présentons nos protocoles expérimentaux et nous commentons les résultats obtenus.

Mots clefs

Stéganalyse, Deep Learning, réseaux de neurones, cover source mismatch.

1 Introduction

La stéganographie est l'art de cacher un message qui doit être indétectable à l'oeil nu dans un contenu. La stéganalyse, quant à elle a pour but de déceler la présence d'un message caché dans un document. Dans notre cas, nous nous sommes intéressé au scénario où le stéganographe utilise toujours la même clé d'insertion [1].

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

La première étape consiste à extraire les vecteurs de caractéristiques à partir des images. Cette étape permet de récupérer un maximum d'informations à partir des images afin de les modéliser et de différencier les images *stégos* [2] des images *covers* [3].

Durant la seconde étape, le stéganalyste utilise un classifieur pouvant être un Ensemble Classifieur [1], un SVM [2] ou bien un Perceptron [3] afin d'apprendre un modèle qui distingue les images *covers* des images *stégos*.

De nos jours, avec l'amélioration de la puissance de calcul des processeurs et des processeurs graphiques, le *Deep Learning* [4] est devenu une méthode beaucoup plus abordable. L'utilisation des réseaux de neurones a été un succès dans de nombreuses disciplines au vu des résultats obtenus, notamment celui de Krizhevsky utilisé sur les images des bases de données ImageNet [4] et Cifar [5] [5, 6].

Récemment, les premières architectures *Deep* ont obtenu des résultats très prometteurs dans le domaine de la stéganalyse [7]. En effet les auteurs de [7] ont montré des résultats proches d'une des méthodes les plus efficace du domaine de la stéganalyse, la classification par Ensemble Classifieur avec le Rich Models [8]. Afin de comparer les deux méthodes, ils ont utilisé les algorithmes d'insertion HUGO [9], WOW [10] et S-UNIWARD [11] sur la base de données BOSSBase [12]. Dans cet article, nous proposons une étude du fonctionnement des réseaux de neurones afin d'expliquer les résultats obtenus par Qian et al. [7].

Les avantages d'utiliser des réseaux de neurones convolutionnels pour la stéganalyse sont multiples. En plus d'être robustes, l'apprentissage est guidé par l'objectif de détection d'un algorithme d'insertion. La classification et le calcul des vecteurs caractéristiques sont faits en même temps, ce qui n'est pas le cas pour le Rich Models avec un Ensemble Classifieur [8] où les deux tâches sont dissociées.

Au cours de notre étude, nous avons voulu comparer une architecture *Deep* avec la méthode classique qui est l'utilisation du Rich Models avec un Ensemble Classifieur [8] qui jusqu'à présent a obtenu les meilleurs résultats sur le scénario de la clairvoyance [13]. De plus, nous avons analysé les résultats obtenus en utilisant le réseau de neurones convolutionnels afin de mieux comprendre son fonctionnement.

1. Ce scénario n'est pas recommandé puisqu'il affaibli la sécurité de l'algorithme d'insertion.
2. Une image *stégo* est une image qui contient un message caché.
3. Une image *cover* est une image qui ne contient pas de message caché.
4. http ://www.image-net.org/
5. http ://www.cs.toronto.edu/∼kriz/cifar.html

Nous présenterons en section 2 des rappels sur les concepts des réseaux de neurones convolutionnels. En section 3, nous verrons la structure de notre réseau. Puis, dans la section 4 nous expliquerons dans un premier temps la méthodologie utilisée pour réaliser nos expérimentations, et dans un second temps nous présenterons les résultats obtenus. Enfin, en section 5 et 6, nous discuterons des liens qui existent entre la construction de notre réseau et les travaux déjà effectués sur le sujet et nous conclurons.

2 Les réseaux de neurones convolutionnels

Un réseau de neurones est un modèle mathématique dont la conception est inspirée des neurones biologiques. Les représentations de ces réseaux sont fortement inspirées de l'article [14].

Ces représentations sont composées de trois parties que l'on appelle des couches qui sont elles même composées de neurones. Les neurones ont une fonction de propagation appelée fonction d'activation, un poids et un biais sur chacune de ses entrées. La première couche est la couche d'entrée dans laquelle on injecte les données que l'on souhaite analyser. La dernière couche est la couche de sortie, dans le cas d'une classification, elle retourne un numéro de classe. La partie intermédiaire est un ensemble de couches dites cachées.

L'opération réalisée par un neurone consiste à additionner le biais avec la somme de ses variables d'entrées pondérées par les poids de connexion. En pratique on réalise le produit scalaire entre le vecteur d'entrée et le vecteur poids (voir Eq. 1.a).

Dans un réseau de neurones convolutionnels, une couche est composée de trois étapes : la convolution, l'application d'une fonction d'activation et enfin le pooling. Le résultat de ces trois étapes est appelé une *feature map*.

Pour détailler les différentes étapes d'une couche, nous allons utiliser le réseau de Qian et al. [7] (voir figure 1). Pour ce réseau, on commence par filtrer une image de taille 256×256 avec un filtre passe-haut $F^{(0)}$. Ce pré-traitement est une spécificité de la stéganalyse, nous avons pu constater que sans ce filtre passe-haut, le réseau mettait beaucoup plus de temps pour converger. Les images filtrées de taille 252×252 sont ensuite données en entrée du réseau.

2.1 La convolution

La convolution de la première couche est une convolution classique. On applique la convolution entre l'image d'entrée et les filtres de la première couche.

Sur la deuxième couche du réseau, une convolution moins classique est utilisée. En effet lors de cette convolution, l'image résultante de cette convolution est la somme de $K^{(l-1)}$ convolutions, avec $K^{(l-1)}$ le nombre de sorties de la couche $l-1$ (voir Eq. 1.b).

$$\sigma_k^{(l)} = \begin{cases} \mathbf{x}_k^{(l)} \cdot \mathbf{w}_k^{(l)} + b_k^{(l)} & \text{si neurone “simple”(1.a)} \\ \sum_{i=1}^{i=K^{(l-1)}} \mathbf{x}_{k,i}^{(l)} \star w_{k,i}^{(l)} + \mathbf{b}_k^{(l)} & \text{si convolution(1.b)} \end{cases}$$

Avec $\sigma_k^{(l)}$ la valeur du neurone k de la couche l, $\mathbf{x}_k^{(l)}$ le vecteur d'entrée du neurone k, $\mathbf{w}_k^{(l)}$ le vecteur poids et $b_k^{(l)}$ le bias. Pour l'équation (2.b), les $\mathbf{x}_{k,i}^{(l)}$ sont les entrées du neurone k et $w_{k,i}^{(l)}$ les noyaux de convolution.

2.2 La fonction d'activation

Une fois la convolution effectuée, une fonction d'activation est appliquée sur toutes les valeurs de l'image filtrée.
Il existe beaucoup de fonctions d'activation, par exemple on peut utiliser le ReLU [15, 16] qui se définit comme $f(\sigma) = max(0, \sigma)$, la fonction $tanh()$ [17] ou bien la fonction sigmoïde [18]. La valeur de sortie $s_k^{(l)}$ d'un neurone k de la couche l dépend donc de sa fonction d'activation et est définie comme :

$$s_k^{(l)} = f(\sigma_k^{(l)}), \tag{2}$$

avec $s_k^{(l)}$ la valeur de sortie du neurone k de la couche l, f la fonction d'activation et $\sigma_k^{(l)}$ la valeur du neurone k.
Le choix de la fonction d'activation peut dépendre du problème de classification. Par exemple, Qian et al. [7] ont utilisé une fonction Gaussienne, ce qui est inhabituel dans les réseaux de neurones.

2.3 Le pooling

Le pooling est une spécificité des réseaux de neurones convolutionnels. Les deux méthodes les plus utilisées pour appliquer cette opération sont les suivantes, soit on fait la moyenne des valeurs de la zone (*pooling average*), soit on extrait uniquement la valeur la plus élevée (*pooling max*). Cette étape permet de faire une réduction de la dimension. L'opération de pooling est une étape de sous-échantillonnage. En pratique, le pooling permet de gagner en temps de calcul.
Le principal avantage du pooling average est qu'il est efficace lorsque l'on souhaite détecter des signaux faibles comme pour le cas de la stéganalyse. Le pooling max est quant à lui efficace lorsque l'on veut détecter des signaux forts, comme des objets par exemple, de plus il permet au modèle d'être invariant aux translations.
Après cette étape de sous-échantillonnage nous obtenons une *feature map* qui est définie comme :

$$I_k^{(l)} = pool(s_k^{(l)}), \tag{3}$$

avec $I_k^{(l)}$ une *feature map* de la couche l, $pool()$ l'opération de pooling et $s_k^{(l)}$ la valeur de sortie du neurone k de la couche l.
La succession de couches de convolutions se termine par un réseau neuronal *fully connected*. Celui-ci est composé

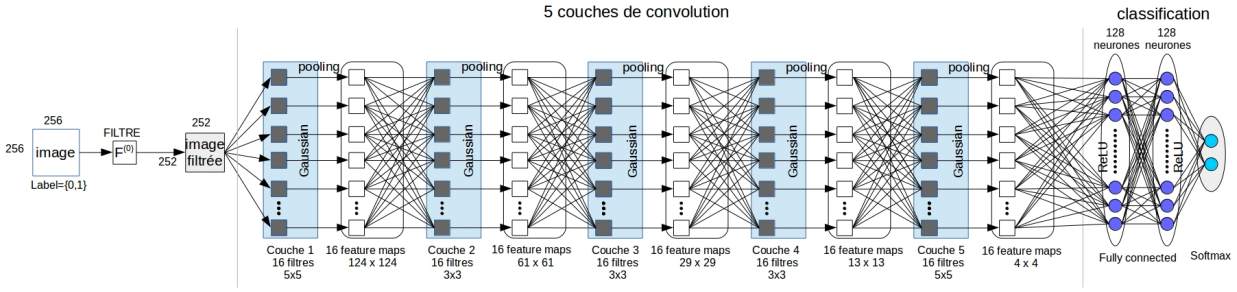


Figure 1 – *Schéma du réseau utilisé par Qian et al. [7].*

de deux parties, la première partie est constituée de couches que l'on appelle des couches *fully connected*. La particularité de cette couche est que tous ses neurones sont connectés avec tous les neurones de la couche précédente mais aussi avec tous les neurones de la couche suivante. La deuxième partie normalise les résultats, on connecte la dernière couche *fully connected* à un *softmax*. Cette fonction est utilisée pour produire une distribution de probabilité entre les différentes classes (chaque classe aura une valeur réelle comprise dans l'intervalle [0, 1]).

3 Configuration de notre réseau

Au cours de nos expérimentations, nous avons testé une quarantaine de réseaux différents. Pour construire notre réseau, nous avons choisi de suivre la même idée que le Rich Models, c'est-à-dire avoir une grande diversité. Contrairement au réseau de Qian et al. [7], nous avons utilisé peu de couches de convolutions, cependant nous avons mis un nombre important de filtres sur nos couches de convolution.

Le premier réseau que nous présentons figure 2 est composé de deux couches de convolution et de trois couches de *fully connected*.

L'image en entrée de taille 256×256 est d'abord filtrée par un filtre passe-haut $F^{(0)}$ de taille 5×5. La taille de l'image devient donc 252×252.

L'image filtrée passe ensuite à la première couche de convolution. Cette couche est composée de 64 filtres de taille 7×7. Notons que pour des contraintes matérielles, notamment à cause de la mémoire, nous appliquons la convolution avec un pas de deux pixels, c'est-à-dire que un pixel sur deux va être traité. Le fait d'appliquer la convolution un pixel sur deux réduit la taille de l'image qui devient 127×127.

Nous pensons que le fait d'utiliser un réseau en hauteur est une des raisons qui nous permis d'obtenir de meilleurs résultats. Augmenter le nombre de couches fait perdre de l'information, cela vient du sous-échantillonnage à l'étape du pooling qui a un effet néfaste sur les résultats du réseau. Nous avons donc décidé de supprimer l'étape de pooling de notre réseau.

Après les 64 convolutions, la fonction d'activation ReLU [15, 16] est appliquée, cette fonction d'activation force les neurones à retourner des valeurs positives. Après chaque couche de convolution, on applique une normalisation locale entre les neurones d'une même couche.

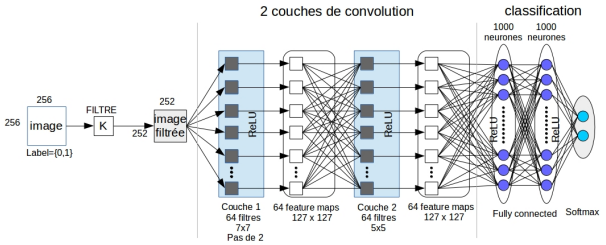


Figure 2 – *Schéma du réseau de neurones convolutionnels (CNN) utilisé.*

Ce type de normalisation se trouve être très efficace lorsque l'on utilise une fonction d'activation non bornée comme le ReLU [15, 16] car elle permet la détection de caractéristiques de hautes fréquences avec de grandes valeurs sur les sorties des neurones. Cette normalisation encourage la "concurrence" pour les grandes activités entre les différentes *feature maps*.

Ensuite, les 64 *feature maps* sont données en entrée de la deuxième couche de convolution qui est composée de 16 filtres. Sur cette couche, la convolution de l'Eq. 1.a est appliquée.

Chacune de nos couches de convolution est suivie d'une fonction d'activation ReLU [15, 16] et d'une normalisation. À la sortie de cette couche, nous avons 16 *feature maps* de taille 127×127. Le vecteur de caractéristiques issu des convolutions a une dimension de 258 064, ce qui est 7 fois plus que le Rich Models qui lui obtient des vecteurs de caractéristiques de taille 34 671.

Après ces deux couches de convolution, nous utilisons un réseau de neurones composé de trois couches *fully connected*. Les deux premières couches ont chacune 1 000 neurones où la fonction d'activation utilisée est le ReLU [15, 16], et la troisième couche est un *softmax* qui permet de calculer la distribution de probabilité des deux classes (*cover* et *stégo*).

4 Expériences

4.1 Protocole expérimental général

Afin de réaliser nos expérimentations, nous avons utilisé la base de données BOSSBase v1.0. Cette base de données est composée de 10 000 images en niveau de gris de taille 512×512 provenant de 7 appareils photo différents. Nous avons découpé les images en quatre afin d'obtenir 40 000 images de taille 256×256. Nous avons dû découper nos

images à cause des contraintes de mémoire GPU, cependant cela nous a permis d'utiliser une base d'apprentissage de grande taille. Qian et al. [7] lui aussi a dû réduire la dimension des images pour des contraintes de mémoire GPU, cependant il est à noté que Qian et al. [7] ont utilisé un redimensionnement et non pas un découpage.

Nous avons créé une deuxième base de données que nous avons appelé LIRMMBase [6]. Cette base de données est composée de 1 008 images en niveau de gris de taille 256×256. Pour créer cette base de données, nous avons utilisé six appareils photo, tous différents de ceux de BOSSBase, et 168 images par appareil photo. Cette base de données nous a permis d'évaluer la sensibilité au *cover source mismatch* de notre modèle.

Pour nos expériences, nous avons avons insérer les message en utilisant le simulateur de l'algorithme d'insertion S-UNIWARD [11] avec une taille de message de 0.4 bit par pixel en utilisant toujours la même clé d'insertion. Après l'insertion, nous avions une base de données de 80 000 images (40 000 images *covers* et 40 000 images *stégos*) pour BOSSBase, et 2 016 images (1 008 images *covers* et 1 008 images *stégos*) pour LIRMMBase.

Les résultats présentés dans les parties qui suivent sont, sauf mention du contraire, la moyenne de 10 exécutions sur des bases d'apprentissages et de tests tirées aléatoirement.

Les résultats donnés dans cette partie sont la probabilité d'erreur P_E qui est définie comme :

$$P_E = \min_{P_{FA}} \frac{1}{2}(P_{FA} + P_{MD}(P_{FA})) \tag{4}$$

avec P_{FA} la probabilité de fausse alarme et P_{MD} la probabilité d'erreur de détection [8].

Dans nos expérimentations, nous comparons trois approches de stéganalyse différentes, un réseau de neurones convolutionnels (CNN), un réseau de neurones entièrement connectés (FNN) représenté figure 3, et le Rich Models avec un Ensemble Classifieur (RM+EC). L'Ensemble Classifieur classe les vecteurs de caractéristiques des images que nous avons extraits avec l'algorithme SRM [8]. Le vecteur de caractéristiques d'une image a une dimension de 34 671.

4.2 Expérimentations sur le scénario de la *clairvoyance*

Dans ce premier test nous nous intéressons au scénario de la *clairvoyance* [13] où le stéganalyste connaît tous les paramètres publics, c'est-à-dire l'algorithme d'insertion, la taille du message inséré et possède une bonne connaissance de la distribution statistique des images du stéganographe. Nous faisons l'hypothèse que le stéganographe a toujours utilisé la même clé d'insertion et que le stéganalyste a accès à des couples d'images *cover/stégo* où les images *stégos* ont été générées avec la même clé.

6. www.lirmm.fr/∼chaumont/LIRMMBase.html

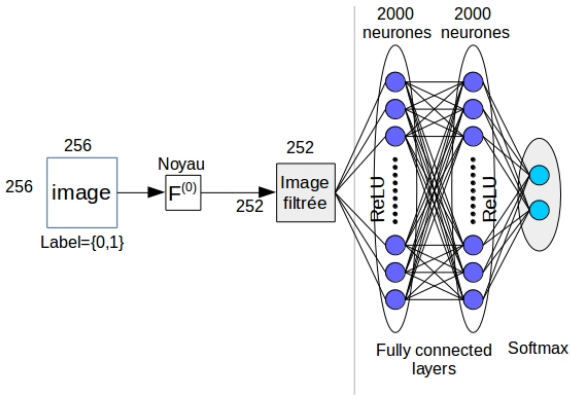


Figure 3 – *Schéma du réseau de neurones entièrement connectés (FNN) utilisé.*

Résultats. Les résultats des trois stéganalyses, par réseau convolutionnel (CNN), par réseau entièrement connecté (FNN) et par Rich Models avec Ensemble Classifieur (RM+EC) sont donnés dans la table 1.

Le Rich Models avec l'Ensemble Classifieur donne une probabilité d'erreur de 24.67% alors que le CNN donne une probabilité d'erreur de 7.4% et le FNN une probabilité d'erreur de 8.66%. Le CNN permet d'améliorer les résultats de plus de 17% et le FNN de plus de 15%.

Le CNN obtient de meilleurs résultats que le Rich Models avec l'Ensemble Classifieur, car les étapes d'extraction de caractéristiques et de classification sont faites en même temps. Avec le Rich Models et l'Ensemble Classifieur, ces deux tâches sont totalement dissociées.

Le nombre important de filtres permet d'avoir une grande diversité dans les vecteurs de caractéristiques. La deuxième couche de convolution semble rechercher la présence de signaux dans les bandes spatio-fréquentielles que nous avons obtenues grâce à la première couche de convolution.

De plus, le fait d'avoir supprimé l'étape de *pooling*, qui agit comme un sous-échantillonnage, nous permet de ne pas avoir une perte d'informations contraitrement au réseau utilisé par Qian et al. [7]. Notons aussi les résultats impressionnants que nous avons obtenus avec le FNN. Il est à noter que le nombre de paramètres inconnus du CNN est d'environ 259 millions alors que celui du FNN est de 131 millions.

Le fait que le stéganographe utilise toujours la même clé d'insertion, le chemin d'insertion est toujours le même, et le simulateur utilise toujours la même séquence de nombres pseudo-aléatoire pour générer la probabilité de modification de la valeur d'un pixel. Dans ce scénario le CNN et le FNN sont plus efficaces, car ils sont sensibles au contenu spatial alors que le Rich Models avec l'Ensemble Classifier est quant à lui sensible aux statistiques de l'image.

Notons que lorsque le stéganographe utilise une clé d'insertion différente pour chaque insertion les résultats chutes à une probabilité d'erreur moyenne de 45.31%. Lorsque la clé d'insertion change, le CNN n'est plus capable de trou-

ver des motifs *stégos*.

	CNN	FNN	RM+EC
Max	7.94%	8.92%	24.93%
Min	7.01%	8.44%	24.21%
Variance	0.12	0.16	0.14
Moyenne	**7.4%**	**8.66%**	24.67%

Tableau 1 – *Résultats des tests sur le scénario de la clairvoyance*

4.3 Expérimentations sur le scénario avec *cover source mismatch*

Le phénomène du *cover-source mismatch* apparaît lorsque l'on effectue l'apprentissage sur des images d'une base de données et que l'on fait les tests sur une autre base de données. La provenance des images apprises par le modèle étant différente de celle utilisée par le stéganographe, la distribution de probabilité des images des deux sources sont différentes et cela crée une difficulté supplémentaire pour le classifieur. Ce problème a été mis en avant lors du challenge BOSS [12]. Notons que [19, 3, 20] ont donné des pistes intéressantes pour lutter contre ce phénomène.
Dans notre cas, puisque les messages ont tous été insérés avec la même clé secrète, les expérimentations suivantes nous ont permis de vérifier que les réseaux apprennent bien la localisation des pixels modifiés.

Résultats. Les résultats des trois stéganalyses, par réseau convolutionnel (CNN), par réseau entièrement connecté (FNN) et par Rich Models avec Ensemble Classifieur (RM+EC) sont donnés dans la table 2.
Malgré la présence de *cover-source mismatch*, on peut constater que les réseaux CNN et FNN ne sont pas du tout impactés par ce phénomène puisqu'ils ont une probabilité d'erreur respectivement de 5.16% et 5.89%. En ce qui concerne le Rich Models avec l'Ensemble Classifieur, on peut voir que le *cover-source mismatch* a un effet très néfaste puisque le résultat est presque aléatoire.
Comme on peut le constater, nous avons obtenu de meilleurs résultats en effectuant nos tests sur LIRMM-Base que sur BOSSBase. Cela vient du fait que les images de BOSSBase sont plus texturées que celles de LIRMM-Base. LIRMMBase est donc plus facile à stéganalyser. Les réseaux de neurones étant invariant au *cover-source mismatch*, les résultats dépendent donc de la complexité des images de la base de données.
La robustesse du CNN peut s'expliquer par le fait que la première couche de convolution décompose les signaux *stégos* en une décomposition spatio-fréquentielle, et qu'ensuite la deuxième couche recherche des motifs particuliers dans les bandes spatio-fréquentielles. De plus, le CNN et le FNN sont spécialisés dans la recherche de motifs spatiaux où la probabilité de modification de la valeur d'un pixel est forte, ce qui est le cas lorsque l'on insère avec la même clé. Notons que dans le cas où la clé d'insertion est différente à chaque insertion, les résultats du CNN chutent à 42.07%.

	CNN	FNN	RM+EC
Max	5.90%	6.60%	49.85%
Min	4.00%	5.40%	47.20%
Variance	0.45	0.31	0.35
Moyenne	**5.16%**	**5.89%**	48.29%

Tableau 2 – *Résultats des tests sur le scnéario avec cover source mismatch*

5 Analyse et discussion

Dans cette section, nous présentons les liens entre le fonctionnement du réseau de neurones les recherches dans la domaine de la stéganalyse.
On peut retrouver dans des articles récents des éléments qui sont similaires au réseau, comme par exemple la projection d'un résidu sur une base de filtre du Rich Models par des projections [21] qui ressemble à ce qui est fait dès la première couche du réseau, ou bien par décomposition spatio-fréquentielle en utilisant des filtres de Gabor [22]. Ces filtres sont utilisés pour définir les projections qui seront utilisées pour calculer un histogramme qui va permettre d'obtenir les vecteurs de caractéristiques.
Sur la deuxième couche de convolution, on constate que l'opération qu'effectue cette couche est très inhabituelle. Nous pensons que c'est cette étape qui permet de rendre les vecteurs de caractéristiques invariants au *cover-source mismatch*. La somme de ces convolutions (voir Eq. 1.b) recherche la présence de motifs dans toutes les bandes issues de la première couche de convolution. La sortie de la deuxième couche de convolution permettrait de donner un indice sur la présence d'un signal *stégo*.
Il est aussi à noter que la normalisation que l'on effectue à la fin de chaque couche de convolution est une opération que l'on peut retrouver dans les articles [23, 24]. Cette normalisation permet à chaque neurone d'avoir des valeurs de sorties du même ordre. La fonction d'activation a elle aussi un impact important sur les performances du réseau, probablement car elle introduit de la non linéarité. Pour le moment l'impact de cette fonction n'est pas encore bien connu. On peut retrouver des opérations non linéaires dans l'Ensemble Classieur [1] avec le vote majoritaire, ou bien dans le Rich Models [8] avec les caractéristiques Min-Max.
Des travaux où la clé d'insertion est réutilisée sur plusieurs images par le stégnographe ont déjà été menés, notamment dans l'article de Ker [25]. Dans cet article Ker détermine les pixels qui ont été modifiés lors de l'insertion en utilisant la méthode de stéganalyse *Weighted Stego-image* (WS).

6 Conclusion

Dans cet article, nous avons poursuivi les travaux sur les réseaux de neurones convolutionnels dans le domaine de la stéganalyse réalisés par Qian et al. [7].
Nous avons évalué notre réseau sur deux scénarios différents. Les premiers tests ont été réalisés avec le scénario

de la *clairvoyance*. Nous avons utilisé la base de données BOSSBase, et inséré avec S-UNIWARD avec une taille de message de 0.4 bit par pixel en utilisant toujours la même clé d'insertion. Les résultats que nous avons obtenus avec le CNN surpassent de loin l'état de l'art puisque l'on a un gain de plus de 17%.
Les derniers tests que nous avons faits portent sur le scénario avec *cover-source mismatch*. Nous avons utilisé la base de données BOSSBase et S-UNIWARD en utilisant toujours la même clé d'insertion et avec une taille de message de 0.4 bit par pixel pour l'apprentissage du CNN. Nous avons ensuite fait nos tests sur la base de données LIRMM-Base. Alors que le Rich Models avec un Ensemble Classifieur obtient des résultats proches de 50%, le CNN est quant à lui insensible au phénomène du *cover-source mismatch* puisqu'il permet d'obtenir une probabilité d'erreur de classification de 5.16%.

Références

[1] J. Kodovský, J. Fridrich, et V. Holub. Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on Information Forensics and Security*, 7(2) :432–444, April 2012.

[2] C. Chang et C. Lin. Libsvm : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3) :27, 2011.

[3] I. Lubenko et A. Ker. Steganalysis with mismatched covers : Do simple classifiers help ? Dans *Proceedings of the on Multimedia and Security*, MM&Sec2012, pages 11–18, New York, NY, USA, 2012. ACM.

[4] Y. Bengio, I. Goodfellow, et A. Courville. Deep learning. Book in preparation for MIT Press, 2015.

[5] A. Krizhevský, I. Sutskever, et G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.

[6] A. Krizhevský et G. Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 2010.

[7] Y. Qian, J. Dong, W. Wang, et T. Tan. Deep learning for steganalysis via convolutional neural networks. Dans *Media Watermarking, Security, and Forensics 2015*, volume 9409 de *Proceedings of SPIE*, pages 94090J–94090J–10, San Francisco, CA, March 2015.

[8] J. Fridrich et J. Kodovský. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3) :868–882, June 2012.

[9] T. Pevný, T. Filler, et P. Bas. Using high-dimensional image models to perform highly undetectable steganography. Dans Rainer Böhme, Philip W.L. Fong, et Reihaneh Safavi-Naini, éditeurs, *Information Hiding*, volume 6387 de *Lecture Notes in Computer Science*, pages 161–177. Springer Berlin Heidelberg, 2010.

[10] V. Holub et J. Fridrich. Designing steganographic distortion using directional filters. Dans *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, pages 234–239, Dec 2012.

[11] V. Holub, J. Fridrich, et T. Denemark. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 2014(1) :1–13, 2014.

[12] P. Bas, T. Filler, et T. Pevný. "Break Our Steganographic System" : The Ins and Outs of Organizing BOSS. Dans *Information Hiding*, volume 6958 de *Lecture Notes in Computer Science*, pages 59–70, Czech Republic, Mai 2011.

[13] T. Pevný. Detecting messages of unknown length. Dans *Media Watermarking, Security, and Forensics III*, volume 7880 de *Proceedings of SPIE*, pages 78800T–78800T–12, Février 2011.

[14] J. Lettvin, H. Maturana, W. Mcculloch, et W. Pitts. What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, 47(11) :1940–1951, Nov 1959.

[15] K. Jarrett, K. Kavukcuoglu, M. Ranzato, et Y. LeCun. What is the best multi-stage architecture for object recognition ? Dans *Proceedings of the IEEE International Conference on Computer Vision*, pages 2146–2153, September 2009.

[16] V. Nair et G. Hinton. Rectified linear units improve restricted boltzmann machines. Dans *Proceedings of International Conference on Machine Learning*, pages 807–814, Haifa, Israel, June 2010.

[17] D. Nguyen et B. Widrow. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. Dans *International Joint Conference on Neural Networks*, pages 21–26 vol.3, June 1990.

[18] M. Norouzi, M. Ranjbar, et G. Mori. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. Dans *IEEE Transactions on Conference on Computer Vision and Pattern Recognition*, pages 2735–2742, June 2009.

[19] J. Pasquet, S. Bringay, et M. Chaumont. Steganalysis with cover-source mismatch and a small learning database. Dans *Proceedings of the European Signal Processing Conference*, pages 2425–2429. IEEE, September 2014.

[20] J. Pasquet, S. Bringay, et M. Chaumont. Des millions d'images pour la stéganalyse : inutiles. *COmpression et REprésentation des Signaux Audiovisuels (CORESA)*, page 53, 2013.

[21] V. Holub, J. Fridrich, et T. Denemark. Random projections of residuals as an alternative to co-occurrences in steganalysis. volume 8665, pages 86650L–86650L–11, 2013.

[22] X. Song, F. Liu, C. Yang, X. Luo, et Y. 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, New York, NY, USA, 2015. ACM.

[23] S. Kouider, M. Chaumont, et W. Puech. Adaptive steganography by oracle (aso). Dans *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6, July 2013.

[24] R. Cogranne, T. Denemark, et J. Fridrich. Theoretical model of the fld ensemble classifier based on hypothesis testing theory. Dans *Information Forensics and Security (WIFS), 2014 IEEE International Workshop on*, pages 167–172, Dec 2014.

[25] A. Ker. Locating steganographic payload via ws residuals. Dans *Proceedings of the 10th ACM Workshop on Multimedia and Security*, MM&Sec2008, pages 27–32, New York, NY, USA, 2008. ACM.

Reconnaissance d'actions basée sur la moyenne des articulations 3D issues des caméras RGB-D

A. BEN TAMOU[1] L. BALLIHI[1] D. ABOUTAJDINE[1]

[1] LRIT-CNRST URAC29,

Université Mohammed V de Rabat, Faculté des Sciences Rabat, Maroc

{ballihi,aboutaj}@fsr.ac.ma

abdelouahid.bentamou@gmail.com

Résumé

Dans ce travail, nous proposons une nouvelle approche de reconnaissance d'actions humaines en utilisant les articulations $3D$ récupérés des squelettes issues des caméras RGB-D. Nous concevons un descripteur basé sur la différence des coordonnées $3D$ des articulations, ce descripteur combine deux caractéristiques telles que la posture et le décalage. Ces deux caractéristiques permettent de coder les deux aspects spatiale et temporelle. Ensuite, nous appliquons une fonction moyenne sur ces caractéristiques afin de former le vecteur caractéristique pour le classifieur Random Forest utilisé pour la classification des actions. Les résultats expérimentaux obtenus sur la base MSR Daily Activité 3D montrent l'efficacité de notre approche par rapport aux méthodes de l'état de l'art en particulier les méthodes basées sur les squelettes.

Mots clefs

Reconnaissance d'actions, image de profondeur, squelette, Random Forest.

1 Introduction

La reconnaissance d'actions est un des domaines de vision par ordinateur, elle permet de regrouper l'ensemble des techniques visant à capturer des informations caractérisant une action et de reconnaître des actions inconnues dans une requête vidéo en se basant sur une collection de vidéos contenant des actions annotées. La reconnaissance d'actions est devenue un sujet très populaire grâce à un grand nombre d'applications possibles : Environnements de surveillance [1], environnements de divertissement [2], reconnaissance de la langue des signes [3], les systèmes de santé [4, 5].

Les premières recherches dans ce domaine ont porté principalement sur l'apprentissage et la reconnaissance d'actions à partir des séquences d'images prises par des caméras RGB [6, 7, 8]. Cependant, ces caméras $2D$ ont plusieurs limitations, elles sont sensibles aux changements de couleur, d'éclairage, ainsi que le problème de fond mobile, des occlusions et la présence de bruit. Les principaux travaux basant sur les images RGB sont résumés dans les recherches d'Aggarwal et al. [9], Weinland et al. [10] et Poppe [11].

Avec l'avènement des capteurs de profondeur, des nouvelles données sont apparues, ces capteurs fournissent en plus des images RGB des cartes de profondeur, donnant la distance entre des objets présentées dans la scène et la caméra de profondeur, et l'estimation des coordonnées $3D$ des articulations de squelette récupérées à partir des cartes de profondeur grâce au travail de Shotton et al. [12]. La figure 1 montre un exemple de flux de données fournit par la caméra Kinect de Microsoft et il est composé de flux de couleur, flux de profondeur et d'un squelette.

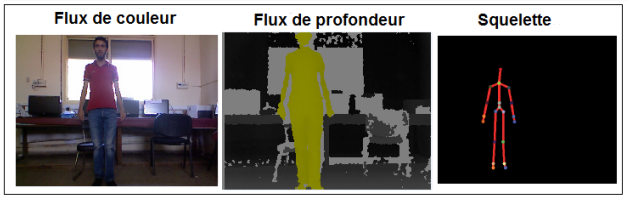


Figure 1 – *Flux de données fournit par la Kinect.*

Dans ce travail, nous proposons un nouveau descripteur de caractéristiques basé sur le calcul de la différence de coordonnées $3D$ des articulations et d'appliquer la fonction moyenne. Nous utilisons ensuite le classifieur Forêts Aléatoires (*Random Forest*) pour la classification des actions. Le reste de ce travail est organisée comme suit : Dans la section 2, nous étudierons les méthodes de la reconnaissance d'actions existantes dans la littérature. Ensuite, dans la section 3, nous présenterons l'aperçu générale de notre approche, et nous allons la détailler dans la section 4, puis dans la section 5, nous allons tester notre approche sur la base *MSR Daily Activity 3D* [13]. Finalement, une conclusion et perspectives dans la section 6.

2 État de l'art

Dans cette section, nous allons présenter l'état de l'art de deux types d'approches de reconnaissance d'actions. Tout d'abord, nous commençons par les approches basées sur la

profondeur, ces approches utilisent en général les images de profondeur afin d'extraire les caractéristiques pour préparer les données d'entrée de leurs classifieur. Vieira et al. [14] ont représenté chaque séquence de profondeur comme une grille $4D$ en divisant les axes d'espace et de temps en plusieurs segments afin d'extraire des caractéristiques de motif d'occupation spatio-temporelles (*SpatioTemporal Occupancy Pattern STOP*). Wang et al. [15] ont considéré la séquence de profondeur comme une forme $4D$ partitionné en sous-volumes $4D$ aléatoires avec différentes tailles et à différents emplacements. Cette fonction est appelée motif d'occupation aléatoire (*Random Occupancy Pattern ROP*).

Yang et al. [16] ont utilisé des caractéristiques d'histogrammes de gradients orientées (*Histograms of Oriented Gradients HOG*) calculées sur des cartes de profondeur de mouvement (*Depth Motion Maps DMM*) comme la représentation d'une séquence d'action. Ils ont projeté chaque carte de profondeur sur les trois plans cartésiens. Chaque carte projetée est normalisée et une carte binaire est générée par le calcul de différence de deux images consécutives. On obtient alors pour chaque carte projetée son DMM. HOG est ensuite appliqué sur la carte DMM pour extraire les caractéristiques de chaque vue. La concaténation de HOG de trois vues forme un descripteur DMM-HOG.

Oreifej et al. [17] ont présenté un histogramme des normales en $4D$ comme un descripteur. Les données initiales sont capturées via une caméra à capteur de profondeur (type Kinect). Les normales sont calculées dans un espace à $4D$ (3 dimensions spatiales + 1 dimension temporelle). Un ensemble de projecteurs sont définis, il s'agit des 120 sommets d'un polytope régulier à 600-cellules. Une optimisation est ensuite effectuée sur ces projecteurs dans le but de capturer au mieux la distribution des normales de façon discriminante. Enfin, un vecteur de 120 valeurs - l'histogramme - est obtenu en sommant pour chacun des 120 projecteurs les projections de chaque normale avec lui-même.

Enfin, nous présentons quelques approches basées sur le squelette, ces approches utilisent le squelette extrait des images de profondeur afin de former le vecteur caractéristique du classifieur. Yang et al. [18] ont proposé d'appliquer Analyse en Composantes Principales ACP sur trois caractéristiques extraites à partir des séquences de positions des articulations. Ces caractéristiques comprennent des caractéristiques de posture, de mouvement et de décalage, afin d'obtenir un descripteur des articulations propres (*Eigen Joints*).

Ofli et al. [19] ont proposé un descripteur basé sur la sélection des articulations $3D$ contenant les plus d'informations (*Most Informative Joints*) par le calcul de quantité d'information associée à chaque articulation, ensuite, ils les ont ordonné par quantité d'information décroissante et finalement ils ont sélectionné les k articulations les plus informatives.

Hussein et al.[20] ont proposé un descripteur basé sur la matrice de covariance (*Covariance of 3D Joints Cov3DJ*). Les données initiales sont des articulations d'un squelette dont les positions sont données dans l'espace $(x; y; z)$ évoluant dans le temps. En pratique, le descripteur proposé est la matrice de covariance de l'ensemble des coordonnées de toutes les articulations. Afin de conserver l'aspect temporel du mouvement, une hiérarchie temporelle inspirée des travaux de Lazebnik et al. [21] composée de sous-séquences temporelles est construite.

Reyes et al. [22] ont effectué le *Dynamic Time Warping DTW* sur un vecteur de caractéristique définie par 15 articulations d'un squelette humain $3D$ obtenue en utilisant Prime-Sense Nite. De même, Sempena et al. [23], ont calculé les quaternions du modèle de squelette humain $3D$ pour former un vecteur de caractéristique de 60 éléments.

Xia et al. [24] ont proposé des histogrammes des positions $3D$ des articulations (*Histograms Of 3D Joint HOJ3D*) qui encodent principalement l'occupation spatiale des articulations par rapport au centre de la silhouette (hanche). En effet, les articulations sont projetées dans un espace sphérique partitionné en n-$bins$. Ensuite, une quantification vectorielle est réalisée à l'aide de k-$means$ pour construire les vecteurs de primitives.

Approche Sac-de-mots (*Bag-Of-Words BOW*), provenait de recherche de récupération de texte, est adoptée pour la reconnaissance d'action en utilisant des squelettes, tel que proposé par Seidenari et al. [25]. L'idée principale de cette approche est d'utiliser des positions des articulations pour aligner plusieurs parties du corps humain en utilisant une solution Sac-de-poses appliquée dans un cadre plus proche voisin.

3 Aperçu générale de notre approche

Le système de reconnaissance d'actions de notre approche est basé sur trois étapes importantes comme illustré dans la figure 2 :

- **Construction de la matrice des caractéristiques** : Après la récupération du squelette à partir des images de profondeur, l'étape de construction de la matrice des caractéristiques permet de calculer pour chaque image sélectionnée deux caractéristiques : la première dite la caractéristiques de posture f_{cc} et la deuxième est la caractéristique de dynamique globale f_{ci}, ces deux caractéristiques sont concaténées afin d'obtenir la caractéristique f_c. La matrice des caractéristiques F_c correspond à la concaténation de toutes les caractéristiques f_c calculées sur plusieurs images.
- **Calcul le vecteur des moyennes M :** À cette étape on applique la fonction moyenne sur chaque ligne de la matrice des caractéristiques F_c.
- **Classification par Forêts Aléatoires :** Le vecteur des moyennes M correspond à une action sera l'en-

trée du classifieur Forêts Aléatoires afin de reconnaître l'action.

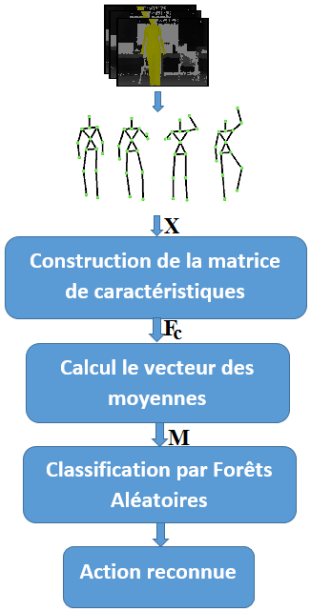


Figure 2 – *L'aperçu générale des différentes étapes l'approche proposée.*

4 Moyenne des Articulations 3D

L'approche Moyenne des Articulations $3D$ (*Mean of 3D Joints :* ***Mean3DJ***) est une approche basée sur le squelette qui permet de construire un descripteur qui calcule les moyennes de différences de coordonnées $3D$ des articulations.

L'approche contient deux étapes : la première consiste à construire une matrice des caractéristiques F_c par le calcul de deux caractéristiques décrivant la différence de coordonnées $3D$ des articulations qui sont la caractéristique de posture statique f_{cc} qui encode l'aspect spatiale et la caractéristique de dynamique globale f_{ci} qui encode l'aspect temporelle dans chaque image c. Ces deux caractéristiques sont inspirées de [18]. La seconde étape permet de calculer la moyenne de chaque ligne de la matrice F_c. La figure 3 résume les étapes pour calculer la matrice des caractéristiques F_c.

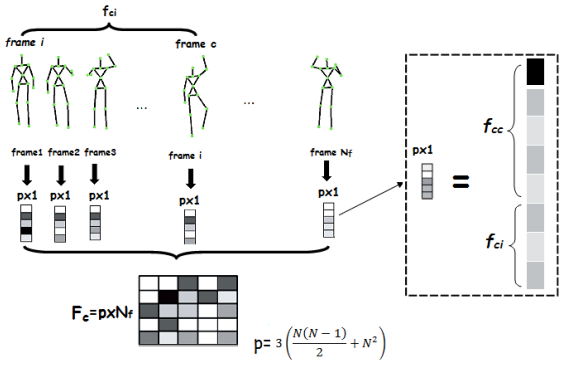


Figure 3 – *Les différentes étapes pour calculer la matrice des caractéristiques F_c.*

Chaque image fournit un ensemble X des articulations $3D$ extraites à partir des cartes de profondeur [12], à savoir $X = \{p_1, p_2, \ldots, p_N\}$ avec $X \in R^{3N}$ où N désigne le nombre des articulations et $p_i = (x_i, y_i, z_i)$ la position en $3D$ de l'articulation i.
Pour caractériser l'aspect spatiale représentée par la posture statique de l'image actuelle c, nous calculons la différence de coordonnées $3D$ des articulations au sein de l'image courante :

$$f_{cc} = \{p_i - p_j | i = 1, 2, ..., N; j > i\} \quad (1)$$

Pour caractériser l'aspect temporelle représentée par la caractéristique de décalage ou la dynamique globale de l'image actuelle c par rapport à l'image initiale i, nous calculons la différence de coordonnées $3D$ des articulations entre l'image c et l'image i :

$$f_{ci} = \{p_i^c - p_j^i | p_i^c \in X_c; p_j^i \in X_i\} \quad (2)$$

La combinaison de ces deux caractéristiques constitue la représentation de la caractéristique préliminaire de chaque image $f_c = [f_{cc}; f_{ci}]$.
Le capteur de profondeur génère dans chaque image N positions $3D$ des articulations, ce qui aboutir à une dimension égale dans f_{cc} à $N(N-1)/2$ et dans f_{ci} à N^2, et puisque chaque positions contient trois coordonnées (x, y, z) alors la dimension finale de f_c pour chaque image est $3[N(N-1)/2 + N^2]$. Par exemple, dans le cas d'utilisation de la caméra Kinect, qui extrait 20 articulations du squelette dans chaque image, f_c est de dimension 1770.
En concaténant les caractéristiques f_c calculées par chaque image sélectionnée on obtient la matrice des caractéristique F_c. Finalement on obtient le vecteur des moyennes M en appliquant la fonction moyenne sur la matrice de caractéristique F_c selon l'équation suivante :

$$M_i = \frac{1}{N_f} \sum_{j=0}^{N_f - 1} F_c(i, j) \; ; \; i = 0, ..., p-1 \quad (3)$$

Où N_f est le nombre des images sélectionnées et $p = 3[N(N-1)/2 + N^2]$ est le nombre de ligne où N est le

nombre des articulations. Le vecteur M est le descripteur final de caractéristiques d'action.
Nous déduisons à partir de l'équation 3 que le descripteur final M aura toujours une dimension de $1 \times p$ quelques soit le nombre des images sélectionnées, et cela parmi les avantages de cette méthode : la faible dimension et par conséquent la rapidité.

5 Résultats Expérimentaux

Dans cette section, nous allons tester notre approche sur la base *MSR Daily Activity 3D* [13] et nous allons comparer notre approche avec les approches de l'état de l'art qui sont basées sur le squelette [13, 18, 25]. On note que nous avons codé notre méthode en langage de programmation C++, en utilisant la bibliothèque OpenCV 2.4.10 et le Microsoft SDK 1.8, en utilisant le classifieur Random Forest.

5.1 Base de données MSR Daily Activity 3D

Le jeu de données *MSR Daily Activity 3D* [13] présente 16 actions faites dans un salon, qui sont : *Boire, Manger, Lire un livre, Téléphoner, Écrire, Utiliser un portable, Utiliser un aspirateur, Encourager, Rester assis, Jeter un papier, Jouer à un jeu, S'allonger sur le canapé, Marcher, Jouer de la guitare, Se lever et S'assoir*. Chaque action est faite par 10 acteurs deux fois : une fois assis et une fois debout, par conséquent la base contient $16 \times 10 \times 2 = 320$ séquences. Les données sont composées d'une acquisition couleur, d'une acquisition de profondeur et d'une capture de mouvements. Elles ont été enregistrées avec une $Kinect^{TM}$. La figure 4 présente quelques exemples d'actions de cette base en couleur, en profondeur et en squelette.

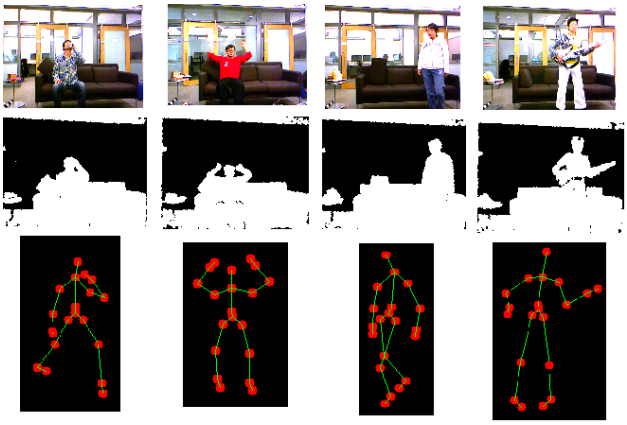


Figure 4 – *Quelques exemple des actions de la base MSR Daily Activity 3D [13].*

5.2 Protocole d'évaluation

Les méthodes de reconnaissance d'actions sont évaluées principalement par leur taux de reconnaissance qui est le pourcentage d'actions qui sont reconnus correctement et par la matrice de confusion. Dans notre teste, nous allons évaluer notre méthode par la matrice de confusion et la technique *Leave-One-Out Cross-Validation* ($LOOCV$) où nous allons laisser à chaque itération un acteur en dehors. Nous pouvons l'appeler aussi *Leave-One-Acteur-Out Cross-Validation* ($LOAOCV$).
La figure 5 montre les résultats d'évaluation de $Mean3DJ$ pour chaque itération de $LOAOCV$ et la figure 6 montre la matrice de confusion de l'ensemble des actions de la base MSR Daily Activity $3D$ [13].

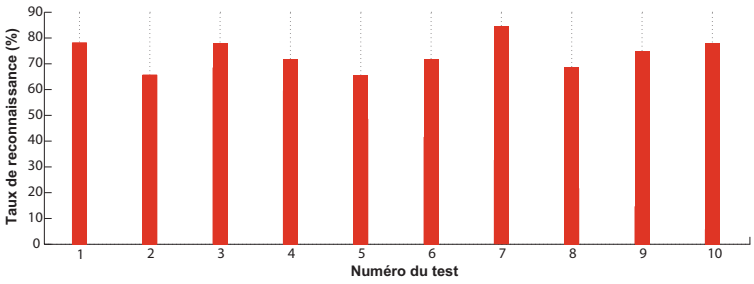


Figure 5 – *Évaluation de $Mean3DJ$ sur la base MSR Daily Activity $3D$.*

Boire	60	20	5	0	0	0	5	5	0	0	0	0	0	0	5	0
Manger	15	65	0	0	0	0	0	0	0	0	0	0	0	0	15	5
Lire	0	5	55	5	15	5	0	0	0	0	0	0	0	0	15	0
Téléphoner	20	10	0	25	0	0	10	0	0	0	0	5	15	10	5	0
Ecrire	0	5	15	0	60	5	0	0	0	0	0	0	5	0	10	0
PC	0	0	5	0	5	50	0	5	0	0	0	5	5	10	5	10
Encourage	0	0	0	0	0	0	95	0	0	0	0	5	0	0	0	0
Marcher	0	0	0	0	0	0	0	95	0	0	0	0	0	0	0	5
Se lever	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
S'assoire	0	0	0	0	0	0	0	0	5	85	5	0	5	0	0	0
Aspirateur	0	0	0	0	0	0	0	0	0	0	95	0	0	5	0	0
Guitar	0	0	0	0	0	5	5	0	0	5	0	80	0	0	5	0
Rester	0	0	5	0	0	0	0	0	0	0	0	0	90	0	5	0
Jetter	20	10	0	0	0	5	0	0	0	0	0	0	0	20	25	20	0
Jeux	0	5	0	0	5	0	0	0	0	0	0	0	15	0	75	0
S'allonger	0	0	0	0	0	0	0	0	0	5	5	0	0	0	0	90

Figure 6 – *Matrice de confusion de l'ensemble des actions de la base MSR Daily Activity $3D$ [13].*

Nous remarquons dans la figure 5 que les taux de reconnaissance de notre approche sont entre 65% et 85%, en calculant la moyenne on trouve un taux de reconnaissance moyenne égale à 73.75%. Nous remarquons à partir de la matrice de confusion illustré à la figure 6 que dans les actions *Encourager, Marcher, Se lever, S'assoir, Rester sans bouger et S'allonger sur le canapé* l'algorithme $Mean3DJ$ donne des meilleurs résultats, et nous remarquons que ces actions n'ont pas des interactions entre l'acteur et un objets.

5.3 Étude comparative

Le tableau 1 montre une étude comparative entre notre approche $Mean3DJ$ et les approches de l'état de l'art appliquées sur la base *MSR Daily Activity 3D* [13].
Puisque notre approche basée seulement sur le squelette, le tableau 1 ne compare que notre approche avec les approches basées aussi sur le squelette, donc nous avons également comparé notre approche avec les solutions rapportées dans [13], [18] et [25]. Pour [13] nous présentons

Tableau 1 – *Comparaison des taux de reconnaissance de quelques approches de l'état de l'art sur la base MSR Daily Activity 3D.*

Méthode	Taux de reconnaissance
Only LOP features [13]	42.5 %
NBNN [25]	53 %
EigenJoints [18]	58.10 %
NBNN + time [25]	60 %
NBNN + parts [25]	60 %
Only Joint Position features [13]	68 %
NBNN + Parts + Time [25]	70 %
Mean3DJ	**73.75 %**

les résultats obtenus en utilisant uniquement les positions des articulations, comme indiqué par les auteurs. On note qu'une précision de 88.20% a été signalée dans [26], cependant, quatre activités avec moins de mouvement (par exemple, rester assis, lire des livres, écrire sur du papier, et utiliser un ordinateur portable) ont été enlevées dans leur expérience.

Nous déduisons à partir du tableau 1 que notre approche proposée est performante et efficace par rapport aux autres approches.

5.4 Analyse des résultats

Dans la matrice de confusion illustrée dans la figure 6, nous constatons que la plupart des actions où notre algorithme est échoue il y a une interaction entre l'acteur et un objet comme canette, livre, cahier, stylo... ceci est expliqué que la méthode $Mean3DJ$ ne prend pas en considération cette interaction avec des objets. Nous remarquons qu'il y a une confusion entre les actions qui ont pratiquement les mêmes mouvements, en effet, l'emplacement des articulations de ces actions sont pratiquement identique, par conséquent, le vecteur des moyennes est presque le même.

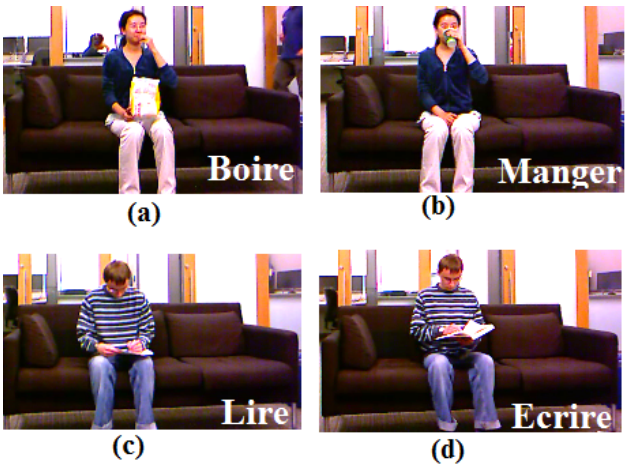


Figure 7 – *Quelques exemples d'actions non reconnues.*

Nous remarquons dans la figure 7 que notre approche a classé la figure (a) comme *Boire* alors que c'est une action *Manger*, contrairement dans (b). Cela est causé par le même mouvement de la main, par conséquent, les mêmes positionnement des articulation. Même remarque pour les actions dans (c) et (d).

6 Conclusion et Perspectives

Dans ce travail, nous avons proposé une nouvelle approche de reconnaissance d'actions humaines basée sur les squelettes. Cette approche est basé sur la différence des coordonnées $3D$ des articulations issues des squelettes. Elle permet de calculer des caractéristiques de posture et de décalage entre l'image courante et l'image initiale. Ensuite elle calcule la moyenne de ces caractéristiques afin d'obtenir le descripteur $Mean3DJ$, ce descripteur est caractérisé par sa dimension fixe $1 \times p$ quelques soit le nombre d'images utilisés. Les résultats expérimentaux obtenus sur la base *MSR Daily Activité 3D* [13] ont montré que notre approche est efficace par rapport aux méthodes de l'état de l'art.

Comme expliqué dans la section 5.4, nous avons constaté que le défis majeur de notre approche est quand il y a une interaction entre l'acteur et un objet de la scène. Dans les perspectives nous allons travailler sur une approche hybride, qui permet de fusionner entre les trois informations (RGB, profondeur et squelette) afin d'améliorer le taux de classification.

Références

[1] Weiyao Lin, Ming Ting Sun, Radha Poovandran, et Zhengyou Zhang. Human activity recognition for video surveillance. Dans *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pages 2737–2740, 2008.

[2] Sean Ryan Fanello, Ilaria Gori, Giorgio Metta, et Francesca Odone. Keep it simple and sparse : Real-time action recognition. *The Journal of Machine Learning Research*, 14(1) :2617–2640, 2013.

[3] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, et Peter Presti. American sign language recognition with the kinect. Dans *Proceedings of the 13th international conference on multimodal interfaces*, pages 279–286, 2011.

[4] Dominick Sanchez, Monica Tentori, et Jesús Favela. Activity recognition for the smart hospital. *Intelligent Systems, IEEE*, 23(2) :50–57, 2008.

[5] Erik E Stone et Marjorie Skubic. Fall detection in homes of older adults using the microsoft kinect. *Biomedical and Health Informatics, IEEE Journal of*, 19(1) :290–301, 2015.

[6] Roberto Vezzani, Davide Baltieri, et Rita Cucchiara. Hmm based action recognition with projection histogram features. Dans *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 286–293. 2010.

[7] Roberto Vezzani, Massimo Piccardi, et Rita Cucchiara. An efficient bayesian framework for on-line action recognition. Dans *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3553–3556, 2009.

[8] Simone Calderara, Andrea Prati, et Rita Cucchiara. Markerless body part tracking for action recognition. *International Journal of Multimedia Intelligence and Security*, 1(1) :76–89, 2010.

[9] Jake K Aggarwal et Michael S Ryoo. Human activity analysis : A review. *ACM Computing Surveys (CSUR)*, 43(3) :16, 2011.

[10] Daniel Weinland, Remi Ronfard, et Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2) :224–241, 2011.

[11] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6) :976–990, 2010.

[12] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, et Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1) :116–124, 2013.

[13] Jiang Wang, Zicheng Liu, Ying Wu, et Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. Dans *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297, 2012.

[14] Antonio W Vieira, Erickson R Nascimento, Gabriel L Oliveira, Zicheng Liu, et Mario FM Campos. Stop : Space-time occupancy patterns for 3d action recognition from depth map sequences. Dans *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 252–259. 2012.

[15] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, et Ying Wu. Robust 3d action recognition with random occupancy patterns. Dans *Computer vision–ECCV 2012*, pages 872–885. Springer, 2012.

[16] Xiaodong Yang, Chenyang Zhang, et YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. Dans *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060, 2012.

[17] Omar Oreifej et Zicheng Liu. Hon4d : Histogram of oriented 4d normals for activity recognition from depth sequences. Dans *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 716–723, 2013.

[18] Xiaodong Yang et YingLi Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. Dans *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 14–19, 2012.

[19] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, et Ruzena Bajcsy. Sequence of the most informative joints (smij) : A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 25(1) :24–38, 2014.

[20] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, et Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. Dans *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2466–2472, 2013.

[21] Svetlana Lazebnik, Cordelia Schmid, et Jean Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. Dans *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178, 2006.

[22] Miguel Reyes, Gabriel Domínguez, et Sergio Escalera. Featureweighting in dynamic timewarping for gesture recognition in depth data. Dans *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1182–1188, 2011.

[23] Samsu Sempena, Nur Ulfa Maulidevi, et Peb Ruswono Aryan. Human action recognition using dynamic time warping. Dans *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, pages 1–5, 2011.

[24] Lu Xia, Chia-Chih Chen, et JK Aggarwal. View invariant human action recognition using histograms of 3d joints. Dans *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 20–27, 2012.

[25] Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Del Bimbo, et Pietro Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. Dans *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 479–485, 2013.

[26] Lu Xia et JK Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. Dans *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2834–2841, 2013.

Evaluation statistique de la segmentation manuelle de données IRM de gliomes diffus de bas grade

M. Ben Abdallah[1] M. Blonski[1,2] S. Mézières[3] Y. Gaudeau[1,4] L. Taillandier[1,2] J.M. Moureaux [1]

[1] Université de Lorraine, Centre de Recherche en Automatique de Nancy (CRAN), CNRS UMR 7039, Faculté de Médecine - Bât D - BP 184, Vandoeuvre-lès-Nancy, 54505, France

[2] Service de Neuro-Oncologie, Centre Hospitalier Régional Universitaire de Nancy, Avenue du Maréchal Lattre de Tassigny,54035 Nancy, France

[3] Université de Lorraine, Institut de Mathématiques Elie Cartan, INRIA BIGS CNRS UMR 7502, BP 239, F-54506 Vandoeuvre-lès-Nancy Cedex, France

[4] Université de Strasbourg, 30 Rue du Maire André Traband, Haguenau, 67500, France

Résumé

Les gliomes diffus de bas grade sont des tumeurs cérébrales primitives rares des adultes. La segmentation manuelle est essentielle pour le suivi des patients atteints de cette tumeur et pour le choix du traitement optimal. Cette méthode étant chronophage, il semble difficile de l'inclure dans la routine clinique. La segmentation automatique apparaît donc comme une solution potentielle pour répondre à cette problématique. Cependant, les algorithmes actuels de segmentation automatique n'ont pas encore prouvé leur efficacité pour les gliomes diffus de bas grade [1][2][3][4][5][6] en raison de la spécificité de ce type de tumeurs. De ce fait, la segmentation manuelle demeure, aujourd'hui, la seule vérité terrain dans ce domaine. Une alternative pour contourner la perte en temps liée à la segmentation manuelle serait de partager la tâche entre différents praticiens, à condition que cette dernière soit reproductible. Le but de notre travail est d'évaluer la reproductibilité de la segmentation manuelle des examens IRM de gliomes diffus de bas grade, en fonction des praticiens, de leur expérience et de leur spécialité. Dans ce travail, nous avons conduit une étude statistique sur les volumes tumoraux d'un panel de 14 experts ayant manuellement segmenté 12 examens IRM de gliomes diffus de bas grade en utilisant le logiciel OsiriX. La plupart des études de segmentation de tumeurs cérébrales publiées mélangent différents types de tumeurs et comparent la segmentation automatique à la segmentation manuelle. Notre étude, au contraire, se focalise uniquement sur les gliomes diffus de bas grade et sur leur segmentation manuelle, car ce sont les plus difficiles à délimiter en raison de leur nature invasive. Une analyse statistique a fourni des résultats prometteurs en démontrant que les facteurs praticien, spécialité médicale et nombre d'années d'expérience n'ont pas d'impact significatif sur les valeurs moyennes de la variable volume tumoral.

Mots clefs

Segmentation, IRM, Gliomes Diffus de Bas Grade.

1 Introduction

Les Gliomes Diffus de Bas Grade (GDBG) sont des tumeurs cérébrales primitives rares des adultes. Ces tumeurs progressent continuellement au cours du temps pour se transformer en tumeurs de grade supérieur dont la malignité est associée à différentes anomalies neurologiques et dont l'issue est fatale. La taille de la tumeur est l'un des plus importants facteurs pronostiques statiques [7]. La régression linéaire à base de modèles mixtes a mis en avant un taux d'augmentation moyen du diamètre tumoral de 4,1 mm par an pour ce type de tumeurs [8]. La stratégie thérapeutique est basée sur une approche personnalisée en plusieurs étapes et avec, à long terme, une adaptation au cours du temps en fonction des changements volumétriques et cliniques. La chirurgie fonctionnelle précoce est généralement le premier traitement proposé lorsque cela est possible. La chimiothérapie peut être utilisée en tant que traitement adjuvant mais aussi, parfois, en traitement néoadjuvant avant la résection chirurgicale. La radiothérapie est habituellement réservée en cas de progression après une chimiothérapie de tumeurs non résécables ou au moment de la transformation anaplasique [9]. Pour le suivi des patients, il est essentiel d'appréhender l'évolution volumétrique dans des conditions cliniques habituelles (pendant les consultations) afin d'adapter de façon optimale le traitement en temps réel [10]. La comparaison qualitative simple de deux examens IRM séparés de 4 à 6 mois d'intervalle ne permet pas habituellement d'objectiver la crois-

sance. Il a été proposé à l'origine de mesurer les 3 plus grands diamètres dans les 3 plans de l'espace, D1, D2, D3, et ensuite d'extrapoler le volume avec la formule suivante : D1*D2*D3/2 [11]. Une méthode de segmentation manuelle sur logiciel a été développée dans [12] et est devenue, depuis, la technique de référence pour la majorité des experts. Comme cette technique est chronophage, une segmentation par différents cliniciens permettrait d'accélérer la prise en charge thérapeutique des patients. Cependant, à notre connaissance, la reproductibilité de la segmentation manuelle sur les images IRM de GDBG n'a pas encore été évaluée [13]. En effet, les principales recherches actuelles englobent plusieurs types de tumeurs cérébrales dans la même étude et se concentrent plutôt sur une comparaison de performance entre segmentation automatique et manuelle. Si la segmentation automatique peut être d'un grand intérêt [14] [15], nous affirmons que, dans le cas des GDBG, la segmentation manuelle reste non seulement la seule vérité terrain mais également la meilleure méthode pour déterminer le volume de ces tumeurs pour la majorité des équipes spécialisées. En effet, la segmentation automatique ne semble pas encore être assez fiable pour distinguer les anomalies de signal de la tumeur d'autres causes d'anomalies de signal (modifications post-chirurgicales ou post-radiothérapie, leucoencéphalopathie de diverses étiologies, etc.). Comme la segmentation manuelle est chronophage, la méthode moins précise des 3 diamètres est généralement préférée pour le suivi du volume tumoral en pratique hospitalière régulière. Le travail que nous proposons ici aborde la question de la reproductibilité de la segmentation manuelle en étudiant l'incidence du praticien sur l'estimation du volume du GDBG. En effet, ce dernier peut fortement influencer le choix d'une thérapie ainsi que les instants de son début et de son arrêt. Ces répercussions motivent la conduite d'une étude subjective de la cohérence de la segmentation manuelle au sein d'un groupe d'experts de GDBG. Une telle cohérence est la clé de la fiabilité et de la reproductibilité du diagnostic clinique et, par conséquent, de la sélection de la thérapie appropriée. Le reste du papier est organisé comme suit. Dans le paragraphe 2, la méthodologie et les outils utilisés dans cette étude sont décrits. Dans le paragraphe 3, les techniques statistiques d'évaluation sont détaillées. Le paragraphe 4 présente les résultats de l'étude statistique. Le paragraphe 5 résume ces résultats et discute de leur conséquences pour la pratique médicale.

2 Matériels et méthodes

Les tests subjectifs permettent d'évaluer la qualité des images et des vidéos, comme par exemple en compression de données à différents taux de compression [16][17]. Ces tests prennent en considération les pratiques médicales les plus courantes et sont réalisés dans un environnement strictement contrôlé. Les résultats des tests subjectifs sont quantifiés par des métriques objectives et se basent, pour leur interprétation, sur une vérité terrain pré-définie. Dans cette étude, la moyenne des volumes a été choisie comme vérité terrain en raison de l'absence d'une vérité terrain absolue. Un expert neuroradiologue a sélectionné 12 examens IRM de 9 patients diagnostiqués avec un GDBG et n'ayant subi aucun traitement préalable. Cet expert ne fait, bien entendu, pas partie du panel de testeurs pour la suite. Les données ont toutes été fournies en coupes axiales, avec une pondération en FLAIR, sauf un examen, qui présentait une pondération en T2. Les examens en pondération FLAIR sont actuellement utilisés en pratique clinique pour le suivi des GDBG et en l'absence de cette séquence IRM, on recourt aux séquences en T2. L'IRM en pondération FLAIR, comme dans l'exemple de la Figure 1, fait apparaître la tumeur en hypersignal, sans le Liquide Céphalo-Rachidien (LCS), et ce, contrairement à la séquence en T2, qui rehausse le LCS en plus de la tumeur, augmentant ainsi la difficulté de la segmentation. Par ailleurs, dans les examens sélectionnés, il y avait 3 examens IRM en Cube FLAIR et 9 examens en acquisition FLAIR usuelle. L'étude de la reproductibilité de la segmentation manuelle a été réalisée dans le Living Lab PROMETEE (PeRceptiOn utilisateur pour les usages du MultimEdia dans les applications mEdicalEs : User perception for multimedia medical usages). [1] Cette plate-forme, située à TELECOM Nancy, école d'ingénieurs de l'Université de Lorraine, est une plate-forme d'innovation permettant l'étude et la bonne gestion de la qualité technique de vidéos et d'images pour un usage médical. Cette plate-forme est également dédiée au post-traitement d'images et vidéos pour une utilisation multimédia dans le domaine médical. Elle est bien équipée et aménagée, offrant un environnement conforme aux conditions générales de visualisation pour ce genre de tests, telles que fixées par la recommandation ITU-BT.500-13 [18]. L'éclairage de la pièce a été contrôlé afin de ne pas produire de réflexion sur l'écran. Un panel de 14 experts a segmenté manuellement l'ensemble des données avec OsiriX comme illustré dans l'exemple de la Figure 1.

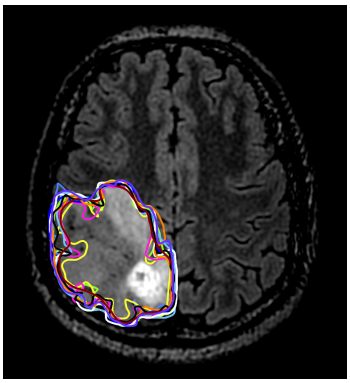


Figure 1 – *Exemple de la segmentation manuelle d'une coupe IRM avec OsiriX. Chaque courbe colorée correspond à la segmentation effectuée par un participant.*

OsiriX est un logiciel libre de visualisation de fichiers

1. http ://telecomnancy.univ-lorraine.fr/fr/recherche/living-lab.

Tableau 1 – *Distribution des participants par spécialité médicale.*

Spécialité médicale	Neurologie	Radiologie	Radiothérapie
Nombre de participants	6	4	3

Tableau 2 – *Distribution des participants par années d'expérience.*

Années d'expérience	$]0\,;10]$	$]10\,;+\infty[$
Nombre de participants	8	5

Dicom sous Apple Macintosh [19] [20]. C'est l'un des meilleurs logiciels de segmentation manuelle d'accès libre, qui est largement utilisé au sein de la communauté de neuro-oncologie. La version 32 bits d'OsiriX a été adoptée pour cette étude. Les participants ont commencé par un test visuel sur une tablette afin de détecter les participants ayant des problèmes d'acuité visuelle. Ensuite, ils ont segmenté un examen d'apprentissage, qui n'a pas été inclus dans les résultats de l'étude, de manière à se familiariser avec l'outil de segmentation OsiriX. L'instruction consistait à délimiter manuellement les contours de la tumeur sur des coupes contenant une prise de contraste liée à un GDBG. Pour être en accord avec la pratique médicale, le fenêtrage radiologique et le nombre de coupes à segmenter n'ont pas été précisés. Les participants ont, d'abord, segmenté la moitié de l'ensemble des examens, ont pris 5 minutes de pause, puis ont complété la segmentation de la deuxième moitié des données. A la fin, ils ont rempli un questionnaire sur leur spécialité médicale et leurs années d'expérience depuis le résidanat. Suite aux premiers tests de cohérence, il s'est avéré que l'un des participants avait des résultats incohérents. Ainsi, tous les résultats décrits par la suite sont basés sur les évaluations réalisées sur les 13 participants cohérents. Pour l'étude de la variabilité introduite par la spécialité médicale sur la tumeur, trois catégories ont été définies : neurologues, radiologues et radiothérapeutes. Pour les années d'expérience, deux groupes ont été créés : $]0\,;10]$ and $]10\,;+\infty[$. La distribution des spécialités médicales et des années d'expérience est répertoriée, respectivement, dans les tableaux 1 et 2. A l'issue du test, les contourages manuels ont été sauvegardés et le volume tumoral, pour chaque examen IRM, a été calculé sous OsiriX en utilisant la méthode de reconstruction des triangulations de Delaunay.

3 Analyse statistique

Les données retenues pour l'étude comportent 12 volumes tumoraux pour chacun des 13 participants : $(x_{i,j})_{i=1...13,j=1...12}$. Le but de l'analyse statistique est d'étudier la variabilité introduite par le facteur praticien sur la variable volume tumoral afin d'examiner l'influence du praticien sur les volumes tumoraux obtenus. Cette étude vise, également, à analyser la relation entre la spécialité médicale du participant ainsi que ses années d'expérience et les volumes tumoraux. Pour l'étude de la variabilité introduite par le praticien, une analyse de la variance à un facteur (ANOVA) [21] a été appliquée aux volumes tumoraux. Afin de quantifier statistiquement la variabilité introduite par les années d'expérience et la spécialité médicale sur les volumes tumoraux, un volume standard, $y_{i,j}$, a été calculé comme suit :

$$y_{i,j} = \left(\frac{x_{i,j} - \overline{x}_j}{\sigma_j} \right) \tag{1}$$

où $\overline{x}_j$ est le volume moyen, et σ_j est l'écart-type d'un volume. En centrant $x_{i,j}$ autour de la valeur moyenne des volumes pour un ensemble de données et en divisant par son écart-type, on tient compte de la difficulté de segmentation. L'écart-type, σ_{y_i}, de $y_{i,j}$, a ensuite été calculé et un test exact de Fisher [22] a été appliqué sur σ_{y_i} pour les deux études.

Afin d'évaluer la variabilité inter-observateurs, le coefficient de variation (COV) [2][4] en volume a été utilisé. Ce coefficient mesure la variation en volume des objets segmentés, et est défini comme suit :

$$COV_j = \frac{\sigma_j}{\overline{x}_j}$$

Une autre métrique qui est utilisée pour évaluer la variabilité inter-participants est l'indice d'accord (AI) [3]. Cette métrique donne l'accord inter-participants, par paires de participants, pour chaque volume $j = 1, ..., 12$:

$$AI_{(i,i'),j} = 1 - \frac{2\,|x_{i,j} - x_{i',j}|}{x_{i,j} + x_{i',j}}$$

pour toutes les paires de participants $(i, i'); i \neq i'$;
$i, i' \in \{1, ..., 13\}$.

4 Résultats

L'analyse statistique a été réalisée avec le logiciel R. Tout d'abord, les volumes des différents participants ont été tracés avec la moyenne des volumes qui est la vérité terrain sélectionnée. Sur la Figure 2, nous observons de faibles variations de volume, même pour les examens IRM en Cube FLAIR, qui sont censés être plus difficiles à segmenter que des examens FLAIR usuels, et l'ensemble des courbes se confond bien avec la courbe de la vérité terrain. Ce premier résultat est confirmé par la boîte à moustaches sur la Figure 3, où la dispersion autour de la moyenne et de la médiane des volumes tumoraux par examen est faible. Cela suggère que les praticiens ont tendance à segmenter les GDBG de manière similaire. Afin de confirmer ce résultat visuel, une ANOVA a été réalisée sur l'ensemble des données. Avec un seuil de signification de 5%, nous avons pu conclure que le facteur praticien n'a pas d'influence significative sur les valeurs moyennes de la variable volume

tumoral.

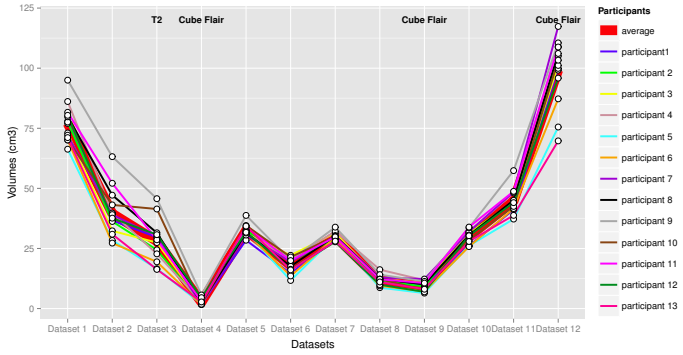


Figure 2 – *Variation du volume tumoral pour tous les participants et de la moyenne des volumes en fonction des examens IRM.*

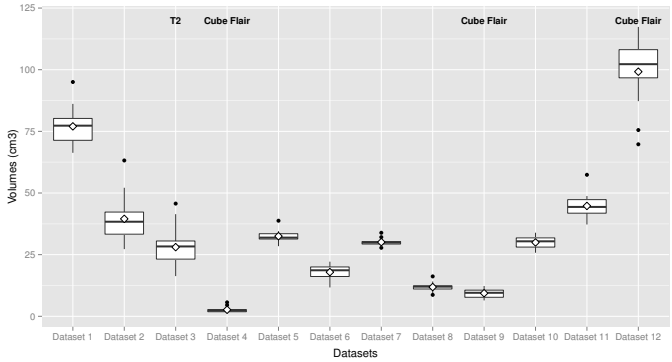


Figure 3 – *Boîtes à moustache des volumes tumoraux pour tous les examens.*

En ce qui concerne la variabilité introduite par la spécialité médicale sur la variable volume tumoral, nous avons appliqué le test exact de Fisher avec un seuil de signification de 5% et nous avons conclu, pour une valeur-p égale à 0,604, que la spécialité médicale n'avait pas d'impact significatif sur l'estimation de la valeur du volume tumoral. Le tableau 3 confirme cette assertion pour la spécialité médicale. Cela montre que la moyenne des AI et que la moyenne des COV sont très proches entre les différentes spécialités. En outre, l'application des tests de Kolmogorov-Smirnov sur le COV par paires de groupes (avec un seuil de signification de 5%) a confirmé ce résultat. A noter que les volumes tumoraux varient de $1,67\ cm^3$ à $117,35\ cm^3$ pour les différents examens. On a, donc, une grande variabilité de la taille du volume. Le COV est évidemment plus sensible aux faibles volumes.
Quant à la variabilité produite par les années d'expérience sur la variable volume tumoral, l'application du test exact de Fisher donne une valeur-p de 0,8961, ce qui indique clairement que l'expérience n'a pas d'influence significative sur le volume segmenté. Ce résultat est, de plus, confirmé par les résultats du tableau 4.

Tableau 3 – *COV et AI par spécialité médicale.*

Spécialité médicale	Neurologie	Radiologie	Radiothérapie
COV (Moyenne±ET)	17.99 ±12.44	16.56 ±10.11	14.48 ±12.32
AI (Moyenne±ET)	0.74 ±0.28	0.73 ±0.27	0.74 ±0.27

Tableau 4 – *COV et AI par années d'expérience.*

Années d'expérience	$]0\,;10]$	$]10\,;+\infty[$
COV (Moyenne±ET)	16.58 ±11.09	14.86 ±11.88
AI (Moyenne±ET)	0.75 ±0.28	0.73 ±0.27

5 Discussions et conclusions

Dans ce travail, nous avons évalué la reproductibilité de la segmentation manuelle du volume tumoral sur des IRM pour différents praticiens, en fonction de leur expérience et de leur domaine d'expertise. Cette étude montre que, en moyenne, ni le praticien médical, ni la spécialité ou l'expérience n'a d'impact significatif sur le volume tumoral. Ce dernier résultat est plutôt surprenant : en effet, on s'attendrait plutôt à ce que l'expérience soit un facteur discriminant. Une autre observation surprenante est que les plus grandes différences en volume ont été remarquées sur les 3 premiers examens, qui sont censés être faciles à contourer. La répartition autour de la médiane de leurs boîtes à moustache (voir Figure 3) est grande par rapport, par exemple, aux examens 4 et 11, qui sont des examens Cube Flair et qui sont censés être plus compliqués à segmenter. Cela pourrait s'expliquer par la nouveauté de l'outil de segmentation utilisé, OsiriX, pour certains participants. Donc, sur les premiers examens, les participants n'avaient pas encore maîtrisé cet outil. Parmi les examens en Cube Flair, l'examen 14 semble avoir été plus difficile à contourer que les examens 4 et 11. Ceci pourrait s'expliquer par l'effet de la fatigue à la fin du test. Mais cela ne semble pas affecter le résultat global de l'étude. Sur la base de plusieurs critères couramment utilisés dans la littérature, tels que le coefficient de variation et l'indice d'accord, l'analyse statistique réalisée pour cette étude a démontré l'absence d'impact, à la fois de la spécialité médicale et des années d'expérience, sur le volume tumoral segmenté, et ce, quelle que soit la difficulté du jeu de données (Cube vs IRM classique). Ce résultat encourageant favorise la collaboration interdisciplinaire entre les cliniciens, en particulier pour les consultations alternées qui sont très fréquentes entre neurochirurgiens et neuro-oncologues. Et même si ce résultat doit être confirmé par des études plus larges, il ouvre la porte à des perspectives intéressantes dans le contexte difficile de la prise en charge des GDBG, où la segmentation automatique ne semble pas encore être en mesure d'offrir une solution totalement fiable. Comme conséquence directe des résultats de nos travaux, le processus de segmen-

tation manuelle pourrait être accéléré, car de nombreux cliniciens seraient en mesure de contourer les GDBG de différents patients, même ceux qu'ils ne suivent pas. Le suivi de l'évolution de la tumeur serait amélioré puisque moins de temps serait consacré par les praticiens à la segmentation manuelle et plus de temps pourrait être dédié aux décisions cliniques concernant le traitement approprié à prescrire à différents stades de la maladie.

Remerciements

Nous adressons nos remerciements à tous les cliniciens du CHRU de Nancy et du CHR de Metz qui ont participé à cette étude. A savoir, nous tenons à remercier : Pr Serge Bracard, Pr Luc Taillandier, Dr Marie Blonski, Dr Basile Wittwer, Dr Guillaume Vogin, Dr Christian Delgoffe, Dr Claire Griffaton-Taillandier, Dr Marie-Alexia Ottenin, Dr Sophie Planel, Dr Fabien Rech, Dr Valérie Bernier, Camille Dahan, Dr Emanuelle Schmitt, Dr Lavinia Jager Simon et Dr Philippe Quetin.

Références

[1] G. P. Mazzara, R. P. Velthuizen, J. L. Pearlman, H. M. Greenberg, et H. Wagner. Brain tumor target volume determination for radiation treatment planning through automated mri segmentation. *Int. J. Radiation Oncology Biol. Phys.*, 59, No. 1 :300–312, May 2004.

[2] C. Weltens, J. Menten, M. Feron, E. Bellon, P. Demaerel, F. Maes, W. Van den Bogaert, et E. van der Schueren. Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. *Radiotherapy & Oncology*, 60 :49–59, Juillet 2001.

[3] K. Xie, J. Yang, Z. G. Zhang, et Y. M. Zhu. Semi-automated brain tumor and edema segmentation using mri. *European Journal of Radiology*, 56 :12–19, Oct 2005.

[4] M. R. Kaus, S. K. Warfield, A. Nabavi, P. M. Black, F. A. Jolesz, et R. Kikinis. Automated segmentation of mr images of brain tumors. *Radiology*, 218 :586–591, Feb 2001.

[5] K. E. Emblem, B. Nedregaard, J. K. Hald, T. Nome, P. Due-Tonnessen, et A. Bjornerud. Automatic glioma characterization from dynamic susceptibility contrast imaging : brain tumor segmentation using knowledge-based fuzzy clustering. *Journal of Magnetic Resonance Imaging*, 30 :1–10, Jul 2009.

[6] I. Njeh, L. Sallemi, I. ben Ayed, K. Chtourou, S. Lehericy, D. Galanaudb, et A. Ben Hamidaa. 3d multimodal mri brain glioma tumor and edema segmentation : a graph cut distribution matching approach. *Computerized Medical Imaging and Graphics*, 40 :108–119, March 2015.

[7] L. Capelle, D. Fontaine, E. Mandonnet, L. Taillandier, J. L. Golmard, L. Bauchet, J. Pallud, P. Peruzzi, M. H. Baron, M. Kujas, J. Guyotat, R. Guillevin, M. Frénay, S. Taillibert, P. Colin, V. Rigau, F. Vandenbos, C. Pinelli, H. Duffau, et French Réseau d'Étude des Gliomes.. Spontaneous and therapeutic prognostic factors in adult hemispheric world health organization grade ii gliomas : a series of 1097 cases : clinical article. *Journal of Neurosurgery*, 118 :1157–1168, June 2013.

[8] E. Mandonnet, J.-Y. Delattre, M. L. Tanguy, K. R. Swanson, A. F. Carpentier, H. Duffau, P. Cornu, R. Van Effenterre, E. C. Jr Alvord, et L. Capelle. Continuous growth of mean tumor diameter in a subset of grade ii gliomas. *Annals of Neurology*, 53 :524–528, April 2003.

[9] H. Duffau et L. Taillandier. New concepts in the management of diffuse low-grade glioma : Proposal of a multistage and individualized therapeutic approach. *Neuro-Oncology*, August 2014.

[10] J. Pallud, L. Taillandier, L. Capelle, D. Fontaine, M. Peyre, F. Ducray, H. Duffau, et E. Mandonnet. Quantitative morphological magnetic resonance imaging follow-up of low-grade glioma : a plea for systematic measurement of growth rates. *Neurosurgery*, 71 :729–739, sep 2012.

[11] J. Pallud, M. Blonski, E. Mandonnet, E. Audureau, D. Fontaine, N. Sanai, L. Bauchet, P. Peruzzi, M. Frénay, P. Colin, R. Guillevin, V. Bernier, M. H. Baron, J. Guyotat, H. Duffau, L. Taillandier, et L. Capelle. Velocity of tumor spontaneous expansion predicts long-term outcomes for diffuse low-grade gliomas. *Neuro-Oncology*, 15 :595–606, Feb 2013.

[12] E. Mandonnet, J. Pallud, D. Fontaine, L. Taillandier, L. Bauchet, P. Peruzzi, J. Guyotat, V. Bernier, M. H. Baron, H. Duffau, et L. Capelle. Inter- and intrapatients comparison of who grade ii glioma kinetics before and after surgical resection. *Neurosurgical Review*, 33 :91–96, 2010.

[13] M. C. Chamberlain. Is the volume of low-grade glioma measurable and is it clinically relevant ? *Neuro-Oncology*, 16 :1027–1028, June 2014.

[14] Z. Akkus, J. Sedlar, L. Coufalova, P. Korfiatis, T. L. Kline, J. D. Warner, J. Agrawal, et B. J. Erickson. Semi-automated segmentation of pre-operative low grade gliomas in magnetic resonance imaging. *Cancer Imaging*, 15(1) :1–10, 2015.

[15] N. Porz, S. Bauer, A. Pica, P. Schucht, J. Beck, R. K. Verma, et R. Wiest. Multi-modal glioblastoma segmentation : Man versus machine. *PLoS ONE*, 9(5), May 2014.

[16] Y. Gaudeau, J. Lambert, N. Labonne, et J-M. Moureaux. Compressed image quality assessment : application to an interactive upper limb radioloy atlas. Dans *IEEE International Conference on Image Processing,ICIP 2014*, Paris, France, Octobre 2014.

[17] A. Chaabouni, Y. Gaudeau, J. Lambert, J.-M. Moureaux, et P. Gallet. Subjective and objective quality assessment for h264 compressed medical video sequences. Dans *4th International Conference on Image Processing Theory, Tools and Applications,IPTA'14*, Paris, France, Octobre 2014.

[18] ITU-R Rec-BT.500. Recommandation 500-13, methodology for the subjective assessment of the quality of television pictures. 2012.

[19] A. Rosset, L. Spadola, et O. Ratib. Osirix : An open-source software for navigating in multidimensional dicom images. *Journal of Digital Imaging*, 17 :205–216, Septembre 2004.

[20] Osirix downloads. http ://www.osirix-viewer.com/Downloads.html.

[21] G. E. P. Box, W. G. Hunter, et J. S. Hunter. *Statistics for Experimenters : an introduction to Design, Data Analysis,and Model Building*. Wiley Series in Probability and Mathematical Statistics, 1978.

[22] D. G. Altman. *Practical statistics for medical research*. Chapman & Hall, 1991.

A Novel Video Coding Framework Based on Machine Learning

D-K. Vo-Nguyen[1,2], M. Antonini[1] and J. Jung[2]

[1]Laboratoire I3S (Université de Nice Sophia-Antipolis - CNRS)
2000 route des Lucioles - Les Algorithmes - bât. Euclide B
06900 Sophia Antipolis, France
dang-khoa.vo-nguyen@etu.unice.fr, am@i3s.unice.fr

[2]Orange Labs
38-40 rue du Général Leclerc
92794 Issy-les-Moulineaux, France
joelb.jung@orange.com

Abstract

Until now, there are a few works that propose to exploit machine learning in video coding. Existing approaches mainly focus on reducing complexity while sacrificing compression quality as trade-off. This paper presents a novel video coding framework with assisting supervised learning algorithms that aim to efficiently compress video by providing bitrate savings. The idea is to predict optimal coding mode of the current block using classification techniques based on already reconstructed frames. This classification process is conducted at the encoder and also at the decoder which is provided with more processing capacity than a conventional decoder. Bitrate savings are eventually obtained from correct predictions.
Implemented in HEVC test model software HM12, a simple practical application using Support Vector Machine as classification algorithm reports an interesting average bitrate savings of 0.8% on a wide set of test sequences.

Keywords

HEVC, Machine Learning, Video Coding.

1 Introduction

In the past decade, machine learning techniques have made great progress. However, still very few works propose to exploit machine learning in video coding. Besides, those approaches mainly aim to reduce complexity and sacrifice the compression quality as trade-off. Inspired by the recent advancement on learning techniques, we propose a novel video coding framework that, unlike existing methods, efficiently compresses video by providing bitrate reduction. The approach is modeled as a classification problem where blocks are classified by their coding mode based on their causal characteristics at both encoder and decoder sides. Bitrate saving is made from correct predictions.
The rest of this paper is organized as follows. The state of the art is first presented, providing an overview on existing video coding approaches based on machine learning. Then the general outline of the proposed coding framework is described, followed by an illustrating practical application. Next, experimental results are presented. Finally, conclusion with possible perspectives are given.

2 State of the art

The idea behind using machine learning in video coding is to exploit structural similarities in video in order to make an accurate prediction of the optimal coding mode of a block. In conventional codecs (AVC, HEVC) the mode used to encode a block is signaled in the bitstream. In general, this mode is computed at the encoder using the rate-distortion (R-D) criterion: $J = D + \lambda R$ with J the R-D cost, D the distortion between reconstructed and original blocks, R the estimated encoding rate, and λ the Lagrange multiplier that depends on the quantization parameter. On the contrary, in approaches based on machine learning, classification techniques are exploited to predict the optimal mode based on characteristics (also called *features*) of the block.
Unlike conventional block based coding schemes, in [1], an approach which compresses texture images on the frame level is proposed. The encoder signals a small number of representative pixels (RP) instead of performing conventional frequency transformation. A semi-supervised learning process is then performed at the decoder to reconstruct the entire image by predicting the rest of the pixels based on the signaled RPs. However, initial results show a poor performance compared with existing codecs.
Other approaches use machine learning techniques mainly to reduce the encoding runtime by creating fast mode decisions. Indeed, block mode decision is a computationally expensive process due to the exhaustive R-D competition of all available options, such as block coding modes, block partitions, Intra directions or Inter motion estimation. Therefore, machine learning is exploited to give relatively accurate decisions on some block parameters, allowing to

skip the time consuming R-D competition. Different classification algorithms are proposed as follows:

— In [2, 3, 4], decision trees are used to find decision rules for the optimal partition size or the optimal Intra direction. Predetermined features, such as the mean and variance values of the current block, are used for the classification process.

— In [5, 6, 7], Bayesian classifier is used in fast mode decision methods to predict block parameters using the information of already encoded neighboring blocks. In [8] and [9], K-means clustering and Support Vector Machine (SVM) are respectively exploited.

All those approaches can be assimilated to the classic classification problem which consists of two main processes: a *training step* and a *classification step*, depicted in figure 1.

(a) Training step: compute a block classifier based on the training data.

A training set is created and includes all blocks (i.e. *training samples*) located in already decoded frames. Each block is characterized by its feature and its R-D optimal coding mode. A classification algorithm (e.g. decision tree, SVM, K-nearest neighbors) is conducted on the set of couples *(Feature, Mode)* of those blocks to infer a block classifier. In the end, the optimal classifier that best classifies training blocks by their coding mode based on their feature is obtained.

(b) Classification step: predict the optimal R-D coding choice of the current block using the block classifier.

For the current block being encoded, its feature is first computed. The block classifier obtained in the training step is then used to predict the R-D optimal coding choice for the current block based on its feature.

Notice that the mentioned methods related to fast mode decision do not reduce the signaling overhead. They only help the encoder to make a fast choice based on observations from the past. That choice must still be signaled in the bitstream. None of these methods uses the training step or the classification step at the decoder side: the decoder simply reads the instruction in the bitstream and then applies the corresponding mode to decode a block, remaining thus a standard decoder such as for HEVC.

In the next section, we will present a novel approach that differs from methods previously presented by providing a reduction in signaling overhead. According to our understanding, there is still no efficient approach that successfully exploits machine learning to increase compression ratio and hence our motivation.

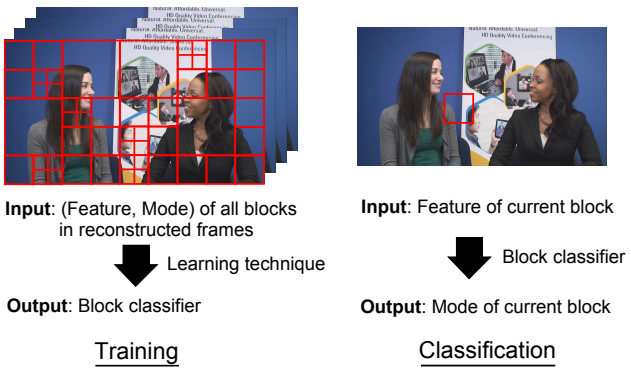


Figure 1 – *Classification problem applied to video coding.*

3 Proposed coding framework

3.1 General outline of the decoding scheme

Being based on the classic classification problem, the proposed approach includes both the training and classification steps. The major difference compared to the existing machine learning based coding approaches is that the proposed method does not have to signal the mode predicted by the classification process, reducing thus the signaling overhead. Indeed, that mode is computed by the decoder itself which can perform training and classification steps identically as the encoder.

We present the proposed method by describing the decoding process. Following notations are used:

— P: the current block to encode/decode

— ϵ: the residual texture

— M_P^*: the optimal coding mode of the block P

— M_P^{**}: the probable coding mode of the block P

— f: the block classifier, modeled as a function that takes characteristics of a block as arguments and gives as output the most probable coding mode. This classifier is predetermined a priori through a training step using machine learning techniques (e.g. SVM, decision tree, etc.) conducted on a set of training block samples independent to the current block to encode

— X_P: *causal* characteristics of the block P, calculated on already decoded regions of picture (e.g. on already reconstructed neighboring blocks of P) and thus also available at the encoder. They can be for example the average gray level, the block size, etc. of the surrounding blocks.

The decoder corresponding to the proposed approach, illustrated in figure 2, includes the following steps:

Step 1: Classifying P

— Calculation of causal characteristics X_P of P

— Deduction of probable coding modes of P, by applying the classifier f, defined a priori by learning techniques from decoded blocks as training samples, on X_P. The result $f(X_P)$ of this classification includes in general the list of several probable coding modes of P.

Step 2: Processing the classification result $f(X_P)$ that includes the list of probable coding modes of P and their probability, then extracting a coding mode M_P^{**} from $f(X_P)$.

Step 3: Decoding P with the computed mode M_P^{**} and the texture residual ϵ parsed from the bitstream. The index of M_P^{**} is optionally parsed depending on the processing and extraction module as mentioned below.

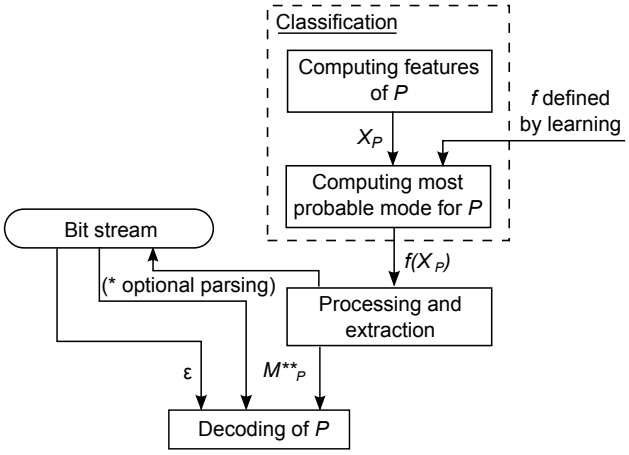


Figure 2 – *Proposed decoding scheme based on machine learning.*

Given that the classification provides in general the list of several probable coding modes of P, the processing and extraction module at the decoder side allows:

(a) Either to modify the list of probable modes by reordering them in function of their probability (based on the result of the classification), and to extract modes that are most probable,

(b) Or to modify the list of probable modes by removing modes with least probability, and to keep modes that are most probable,

(c) Or to extract the best mode with highest probability.

In the two cases (a) and (b), the classification optimizes the list of candidates modes. The index of the optimal mode (based on the modified list) is parsed from the bitstream. In the case (c), the most probable mode is selected; thus no index needs to be read from the bitstream. One among these three configurations is selected beforehand and known by both the encoder and the decoder.

The features of a block being used must be causal, i.e. constructed based on decoded data, and are similar at both the encoder and the decoder so that similar learning process can be performed at both sides.

The classifier f, defined by learning techniques, can be built a priori and independently of the current sequence to encode, or on-the-fly to adapt to the content of the current sequence. These two types of classifier are referred to respectively as static and dynamic classifier.

3.2 Proposed features for block classification

Concerning the block features to be exploited, for the sake of simplification, we propose the following histograms to construct the feature descriptor of the current block:

— Grayscale histogram: sampled with 128-bins.

— Histograms based on Gabor filters: consisting of four concatenated grayscale histograms computed after applying four Gabor filters [10] with angles respectively of 45°, 90°, 135°, 180° as illustrated in figure 3. The concatenated histogram has 512 bins.

— Histogram of Oriented Gradients (HOG): a 36-bins histogram corresponding to 9 directions of gradient.

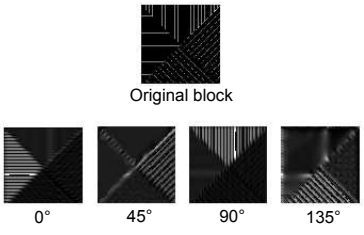


Figure 3 – *Different Gabor filters applied on a block.*

HOG histograms are calculated using *VLFeat* toolbox [11]. The algorithm decomposes the image into square cells of a given size (typically 8×8 pixels), computes a histogram of oriented gradient in each cell, and then renormalizes the cells by looking into adjacent blocks.

— Optical flow (OF) based histograms: consisting of a 256-bins histogram of motion vectors (MV) magnitude coupled with a 360-bins histogram of motion orientations. MVs are computed on pixel level using OF technique applied on causal regions in both previously decoded frame and current frame. The algorithm to calculate the OF can be found in [12].

For the current block, all those histograms are computed based on data of five causal neighboring blocks (Left, Above, AboveLeft, LeftBottom, AboveRight) and are then concatenated to produce a vector corresponding to the feature of the current block, with the vector size of 6460.

4 Proposed practical application

4.1 Encoding scheme

We present a practical application of the general outline of the proposed framework. The specific encoding scheme of this application is illustrated in the figure 4. For simplification purpose, the processing and extraction module considers only a single coding mode which is the most probable mode predicted by the classification (case (c) among three possibilities related to the processing and extraction module). The R-D competition is still performed without any modification, along with the introduced classification process. If the mode predicted by classification does not match the optimal mode computed by the R-D competition, their difference is signaled in the bitstream.

We implemented the proposed scheme in the HEVC software test model version 12 (HM12). For the classification, we choose SVM, one of the most efficient machine learning algorithms. Being a powerful classifier, SVM can be a useful tool when the data are not regularly distributed or have an unknown distribution. It is also often used to classify problems with arbitrary complexity. We choose $\mathrm{SVM}^{multiclass}$ [13] as the SVM implementation to be the core of our classification module. It is integrated directly in the test model to allow on-the-fly prediction.

4.2 Coding modes to be classified

We limit the number of modes to be classified to three for simplification purpose. A study is made with the objective

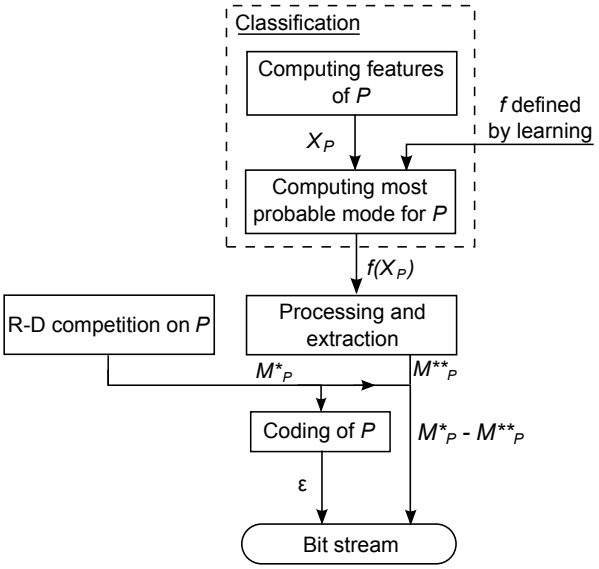


Figure 4 – *A speficic encoding scheme for the proposed practical application.*

to determine the coding modes that are suitable to be predicted by the proposed approach. We compute in table 1 the signaling cost of different syntax elements mentioned in figure 5 which represents the HEVC signaling scheme of different coding modes.

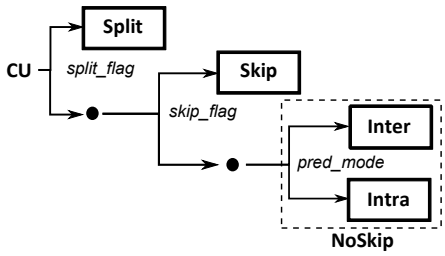


Figure 5 – *HEVC signaling scheme for different coding modes Split/Skip/Inter/Intra*

We observe that if a block is not split, indicating whether the block is Skip or NoSkip by signaling $skip\_flag$ requires an average signaling overhead of 3.3% in the total bitstream. Signaling cost of $split\_flag$ and $pred\_mode$ is respectively 2.2% and 0.9%. All those three syntax elements are encoded using an entropy coding method called Context-Adaptive Binary Arithmetic Coding (CABAC), where the probability models are selected adaptively based on local context, allowing better modelling of probabilities. The significant proportion of $skip\_flag$ in the total bitstream suggests us to predict Split/Skip/NoSkip since the potential impact on the coding performance is significant.

Sequences	***split_flag***	***skip_flag***	***pred_mode***
BasketBallPass	2.6	4.1	0.8
BQSquare	2.2	3.6	0.1
BlowingBubbles	2.0	2.6	1.2
RaceHorses	2.0	3.0	1.5
Average	**2.2**	**3.3**	**0.9**

Table 1 – *Cost of some syntax elements signaling coding modes in the total bitstream (%) in HEVC for some video sequences. Test conducted in LD-P-Main configuration.*

4.3 Signaling scheme

Two syntax elements are introduced as follows:

— *learn_opt_flag*: to signal whether or not the choice predicted by the classifier matches the optimal choice computed by the R-D competition,

— *learn_correction*: to correct the decision of the classifier in case *learn_opt_flag* is false by signaling the difference between the optimal choice and the choice predicted by the classifier.

With those syntax elements, the signaling scheme adapted for this proposed practical application is illustrated in figure 6, replacing the conventional scheme. If the R-D optimal mode cannot be predicted correctly by the classifier, it is enough to signal it among two remaining modes using *learn_correction* with only one bit. Compared with the HEVC signaling scheme, the gain comes from a Skip/NoSkip block that is correctly predicted (gain of one bit), and the loss comes from a Split block that is incorrectly predicted (loss of one bit).

We design CABAC contexts for the two syntax elements *learn_opt_flag* and *learn_correction* as follows:

— *learn_opt_flag*: for each block size of 64×64, 32×32 or 16×16, 3 contexts are defined based on *learn_opt_flag* of the neighboring blocks, in a similar way as for *split_flag* or *skip_flag* in HM12 codec:

- None of the Left and Above blocks is encoded with *learn_opt_flag* $= 1$ or is available
- Only one block among the Left and Above blocks is encoded with *learn_opt_flag* $= 1$
- Both the Left and Above blocks are encoded with *learn_opt_flag* $= 1$

— *learn_correction*: for each block size, 10 contexts are defined based on the combination of cases where coding modes of the neighboring Left and Above blocks take values among {Split/Skip/NoSkip/NotAvailable}.

For evaluation purposes, the proposed scheme is compared against the reference HEVC. Furthermore, we give, as side information, the compression performance of proposed method in raw binary bits in comparison with HEVC. For that, we need to disable CABAC contexts for related syntax elements, i.e. *learn_opt_flag*, *learn_correction* in the proposed scheme and *split_flag*, *skip_flag* in HEVC. Otherwise, CABAC contexts mechanism will further compress bits related to those syntax elements by exploiting local correlation with probability modelling.

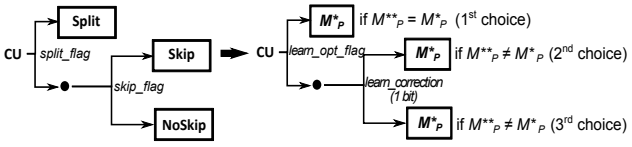


Figure 6 – *HEVC signaling scheme for Split/Skip/NoSkip (left) replaced by proposed machine learning based signaling scheme (right). Each of three choices takes a value among Split/Skip/NoSkip.*

	Grayscale	Gabor	HOG	OF
Grayscale	52.8%	56.8%	54.5%	56.2%
Gabor		53.7%	54.3%	53.7%
HOG			50.7%	53.3%
OF				50.7%
Combined use of all 4 histograms: 59.2%				

Table 2 – *Classifier accuracy with different histograms as block features for classifying Split/Skip/NoSkip modes.*

4.4 Selection of features to be exploited

We perform a test to evaluate the efficiency of four histograms mentioned in previous section as block features. In table 2, each column and row corresponds to a type of histogram. Therefore, diagonal values represent the classifier accuracy when only one type of histogram is used. This accuracy value represents the performance of the block classifier, indicating how often the classifier makes correct prediction of the optimal coding mode for tested blocks. Other values correspond to the combination of two different histograms, where both are concatenated to produce a vector corresponding to the feature of the current block. All histograms are normalized with a L2-norm for concatenation. We observe that any case at position (i, j) will have better accuracy value than cases at positions (i, i) and (j, j) on the diagonal, proving that combining different histograms improves block classifier accuracy regarding the use of a unique histogram. Eventually, a combination of all four histograms yields best accuracy of 59.2%.

4.5 Number of frames in training set

To adapt to the content of the tested sequences, we propose to use the dynamic classifier which is built based on a dynamic training set, consisted of all blocks in the reconstructed frames located within a sliding time window. Different number of reconstructed frames in the training set (i.e. window width) is tested: 1, 2, 4, 8 or 16 frames. The results are given in table 3, where CABAC contexts are disabled as previously mentioned in order to evaluate the raw bitrate saving. The combination of four proposed histograms is used as block feature.

We observe that increasing the number of reconstructed frames in the training set up to 8 frames improves the classifier accuracy which results in more coding gain. On the set of tested sequences, the learning algorithm effectively exploits structural similarities in previously reconstructed frames to make predictions. However, taking too many frames in the training set (16 frames) reduces in turn the gain because the data from more distant frames is less correlated to the current frame. Note that most test sequences have frame rate of 50 Hz.

Sequences	**1 fr.**	**2 fr.**	**4 fr.**	**8 fr.**	**16 fr.**
BasketBallPass	-0.5	-1.2	-2.1	-2.1	-1.6
BQSquare	-0.4	-0.5	-0.9	-1.2	-1.2
BlowingBubbles	-0.1	-0.2	-0.0	-0.1	0.5
RaceHorses	-0.3	-0.1	-0.2	-0.3	-0.4
Average	**-0.3**	**-0.5**	**-0.8**	**-0.9**	**-0.7**
Classifier accuracy	56%	58%	59%	59%	59%

Table 3 – *B-D rate savings (%) of proposed method for classifying Split/Skip/NoSkip modes with different widths of sliding training window (CABAC contexts disabled).*

5 Experimental results

All coding gains are measured using Bjøntegaard Delta (B-D) rate [14] which represents the average difference between two R-D curves on the considered QP range. Standard QP values of 22, 27, 32 and 37 are used. Only the configuration LowDelay-P-Main is tested.

Coding performance of the proposed method is evaluated against the reference HEVC and is given in table 4. Dynamic classifier with 8-frames sliding window is used. The comparison in raw bitrate reduction is also provided. There are thus two columns corresponding to cases where CABAC is respectively enabled and disabled for related syntax elements in both the tested version and the reference version. The theoretical maximum gain is also provided in brackets, which corresponds to the case where the classifier always predicts correctly the optimal coding mode of the current block (i.e. no correcting information is signaled).

When comparing with the reference HEVC, the performance of the proposed approach is interesting, providing an average gain of -0.8%. Gain up to -3.6% is achieved.

If we observe the comparison in terms of raw bitrate reduction, i.e. when CABAC contexts are disabled, the proposed approach can efficiently save up the signaling overhead dedicated for Split/Skip/NoSkip modes compared to HEVC. A significant coding gain of -1.8% is achieved in average. Gain up to -6.3% is reported. This result also proves that our CABAC contexts designed for *learn_opt_flag* and *learn_correction* in the proposed scheme do not perform as good as the contexts for *split_flag* and *skip_flag* in HEVC. Indeed, contexts for HEVC syntax elements are highly optimized to achieve such level of compression.

	CABAC ctx. enabled	**CABAC ctx. disabled**
BasketBallPass_wvga	-0.3 (*-2.6*)	-2.1 (*-2.5*)
BQSquare_wvga	-1.0 (*-4.1*)	-1.2 (*-2.8*)
BlowingBubbles_wvga	-0.0 (*-2.5*)	-0.1 (*-1.4*)
RaceHorses_wvga	-0.2 (*-3.4*)	-0.3 (*-1.3*)
Anemone_wvga	-0.5 (*-7.6*)	-0.2 (*-3.7*)
Book_wvga	-0.1 (*-6.6*)	-2.5 (*-1.2*)
Ducks_wvga	-0.2 (*-5.3*)	-1.1 (*-2.2*)
Keiba3_wvga	-0.2 (*-3.6*)	-0.2 (*-1.8*)
Flower4_qwvga	-3.0 (*-3.8*)	-6.3 (*-7.1*)
Keiba3_qwvga	-0.5 (*-2.9*)	-0.4 (*-1.4*)
Nuts3_qwvga	-0.0 (*-3.8*)	-2.6 (*-4.8*)
HallMonitor_qwvga	-0.8 (*-3.7*)	-2.1 (*-2.8*)
Irene_qwvga	-0.4 (*-3.1*)	-1.7 (*-2.9*)
Marc_qwvga	-1.0 (*-5.2*)	-0.6 (*-2.6*)
Modo_qwvga	-3.6 (*-9.2*)	-6.0 (*-6.1*)
NewsCar_qwvga	-0.1 (*-3.7*)	-1.7 (*-3.2*)
Average	**-0.8 (*-4.4*)**	**-1.8 (*-3.0*)**

Table 4 – *B-D rate savings (%) of proposed method classifying Split/Skip/NoSkip modes with CABAC contexts enabled/disabled for related syntax elements. Theoretical maximum gain is given in brackets.*

Predicted modes	% blocks	Prediction accuracy
Split	46.7%	77.6%
Skip	24.0%	63.8%
NoSkip	29.3%	26.0%
All	100%	59%

Table 5 – *Percentage of blocks and classifier accuracy for each predicted mode when classifying Split/Skip/NoSkip modes (CABAC contexts disabled).*

We also provide in table 5 the classifier accuracy specifically when predicting each mode among Split/Skip/NoSkip in case CABAC contexts are disabled. It is reported that the proposed classifier predicts less accurately NoSkip mode than Split and Skip modes. Effectively, NoSkip includes Inter and Intra modes which are often selected on blocks with less spatio-temporal correlation from previously reconstructed data than Skip mode. The prediction, which is based on blocks correlation, is thus more prone to error. Furthermore, compared with other approaches exploiting machine learning in video coding, the classifier accuracy is not as good since features of current block are computed solely on its causal regions, thus providing less precision. Indeed, in all other approaches, data of the current block itself is exploited as features, which is not available to be computed by the decoder when using this proposed method. The overall low accuracy of the classifier leads to the disparity between B-D rate savings and theoretical maximum gain.

Finally, we note that the runtime in this practical application is very high and not yet suitable for realistic video coding due to following reasons:

— $\text{SVM}^{multiclass}$: this SVM implementation is very slow when used for classifying multiple classes. For example, it takes several hours to compute the model of the classifier based on training blocks.

— Computation of histograms as block features: the construction of different histograms (Gabor, HOG and OF) also requires significant processing time, mainly by the use of third party libraries.

6 Conclusion

In this paper, a novel framework that applies machine learning algorithms in video coding is presented. Unlike other existing works that mainly focus on reducing complexity, proposed approach allows improving compression ratio by reducing signaling overhead. A practical application is also described, using SVM classifier to predict the coding mode of the current block. The features descriptor consists solely of different histograms constructed based on causal neighboring blocks. Experimental results show interesting bitrate savings, with 0.8% of gain in average compared to the reference HEVC.

Several perspectives can be envisaged. First, since the accuracy of the block classifier remains low, research to improve the classifier performance is to be conducted. Furthermore, it is planned to further extend proposed framework to predict more coding modes to increase the saving bitrate. Improving CABAC contexts for newly introduced syntax elements is also an interesting perspective. Finally, it is considered to reduce the runtime so that realistic video coding applications can be made.

References

[1] B. Bai, L. Cheng, C. Lei, P. Boulanger, and J. Harms. Learning-based multiview video coding. *Picture Coding Symposium (PCS)*, 2009.

[2] D. Han, T. Purushotham, K.V.S. Swaroop, and K.R. Rao. Low complexity h.264 encoder using machine learning. *Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA)*, 2010.

[3] P. Carrillo, T. Pin, and H. Kalva. Low complexity h.264 video encoder design using machine learning techniques. *International Conference on Consumer Electronics(ICCE)*, 2010.

[4] W. Ma, S. Yang, L. Gao, C. Pei, and S. Yan. Fast mode selection scheme for h.264/avc inter prediction based on statistical learning method. *IEEE International Conference on Multimedia and Expo*, 2009.

[5] C. H. Lampert. Machine learning for video compression: Macroblock mode decision. *Proceedings of the 18th International Conference on Pattern Recognition*, 2006.

[6] H. R. Tohidypour, M. T. Pourazad, and P. Nasiopoulos. A low complexity mode decision approach for hevc-based 3d video coding using a bayesian method. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.

[7] X. Zhou, C. Yuan, C. Li, and Y. Zhong. Fast mode decision for p-slices in h.264/avc based on probabilistic learning. *11th International Conference on Advanced Communication Technology*, 2009.

[8] C. Di and C. Yuan. A novel fast mode decision algorithm for h.264/avc based on probabilistic learning. *3rd International Congress on Image and Signal Processing*, 2010.

[9] C-.K. Chiang, W.-H. Pan, C. Hwang, S.-S. Zhuang, and S.-H. Lai. Fast h.264 encoding based on statistical learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2011.

[10] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biological Cybernetics*, 1989.

[11] A. Vedaldi and B. Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. *Proceedings of the international conference on Multimedia*, pages 1469–1472, October 2010.

[12] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, May 2009.

[13] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.)*, 1999.

[14] G. Bjøntegaard. Calculation of average PSNR differences between RD curves. *VCEG-M33, ITU-T VCEG, 13th Meeting, Austin, TX, USA*, April 2001.

SAILLANCE VISUELLE MULTI-ÉCHELLE DES MAILLAGES 3D COLORÉS

Anass Nouri, Christophe Charrier, Olivier Lézoray

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

RÉSUMÉ

La détection de la saillance visuelle est une étape de prétraitement importante pour plusieurs applications traitant des données 3D. Ce papier propose une nouvelle approche utilisant un descripteur local sous forme de patch de taille adaptative pour le calcul de la saillance visuelle des maillages 3D. Ce descripteur est utilisé pour une mesure de similarité qui sera intégrée par la suite dans le calcul multi-échelle de la saillance. Les résultats expérimentaux montrent que l'approche proposée permet d'obtenir des résultats compétitifs avec la littérature mais surtout innovants quant à la détection de la saillance des maillages colorés.

Mots-clés— Saillance visuelle, maillages 3D colorés, patch, multi-échelle.

1. INTRODUCTION

La saillance visuelle peut être définie par l'information perceptuelle permettant de faire ressortir des objets ou des régions de leur voisinage et qui, par conséquent, attirent l'attention visuelle humaine. Plusieurs modèles de saillance ont été proposés pour les images 2D [1]. Dans le même temps, les technologies d'acquisition de données 3D ont connu des développements importants, ce qui a contribué à l'acquisition de grandes quantités de données sous forme de maillages 3D.

Etant donné que les modèles de saillance 2D ont facilité l'émergence d'un grand nombre d'applications basées sur la saillance telles que la détection des objets saillants, la segmentation, et le recadrage d'images, il est à la fois naturel et nécessaire de disposer d'une mesure de saillance pour les maillages 3D. En effet, la valeur ajoutée de la saillance est notable dans diverses applications du domaine de la vision par ordinateur, par exemple : la compression adaptative [2, 3], le lissage [4], le redimensionnement [5], et la sélection des points de vues optimaux [6, 7].

La saillance visuelle peut également être vue comme la mesure permettant de quantifier l'importance d'un nœud sur une surface d'un maillage 3D similairement à la vision humaine. Dans la littérature, on ne trouve pas d'approches traitant la saillance des maillages 3D colorés mais uniquement celles traitant la saillance géométriques et qui restent moindres par rapport aux méthodes estimant la saillance 2D. Nous citons dans ce papier quelques travaux importants sur la saillance géométrique. Lee *et al* [2] calculent la saillance d'un maillage 3D en utilisant un opérateur centré sur les courbures gaussiennes dans un espace d' échelles DoG (Difference Of Gaussians). Leifman *et al* [6] definissent les zones d'interêts sur une surface en combinant la distinction d'un nœud (similarité entre des descripteurs Spin-Images [8]) et les extrémités du maillage. Wu *et al.* proposent une approche de détection de la saillance considérant à la fois le contraste local (similarités multi-échelle entre des cartes de profondeurs [9]) et la rareté globale (en utilisant un clustering sur les caractéristiques du contraste local). Tao *et al.* [10] proposent une approche basée sur la segmentation de la surface du maillage en régions distinctes en utilisant des cartes de profondeurs locales [9] et sur le classement des patchs non saillants pour détecter les régions saillantes. Song *et al.* [11] estiment la saillance dans le domaine spectral en analysant les déviations et caractéristiques du spectre du log-Laplacien.

Dans ce papier, nous proposons une nouvelle méthode multi-échelle pour l'estimation de la saillance des maillages 3D. Celle-ci est basée sur un nouveau descripteur local sur chaque nœud sous la forme de patch de taille adaptative. Il sera utilisé lors du calcul des similarités entre patchs et intégré dans le calcul multi-échelle de la saillance. Par ailleurs, nous appliquons cette méthode pour détecter la saillance des maillages 3D colorés qui n'a jamais été traitée auparavant. Notre approche est comparée à l'état de l'art et présente des résultats compétitifs.

2. LA SAILLANCE VISUELLE DES MAILLAGES 3D

Afin de définir notre mesure de saillance des maillages 3D, nous commençons par construire, sur chaque nœud du maillage un descripteur local (sous forme de patch) caractérisant sa configuration géométrique (ou colorimétrique) locale. Une fois les patchs locaux définis, les similarités entre les caractéristiques locales des nœuds adjacents sont calculées et affectées à leurs arêtes incidentes. Ainsi, le degré d'un nœud permet de fournir sa saillance mono-échelle. En variant la taille du voisinage sphérique de chaque nœud lors de la construction des patchs, plusieurs échelles sont obtenues et fusionnées en utilisant leurs mesures d'entropie. Nous représentons un maillage $\mathcal{M}$ par un graphe non-orienté [12] $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ où $\mathcal{V} = \{v_1, \ldots, v_N\}$ est l'ensemble des nœuds N et $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ l'ensemble des arêtes. L'ensemble des arêtes

est obtenu à partir des faces connectant les nœuds du maillage. Pour chaque nœud du maillage sont associés 3 coordonnées 3D $\mathbf{p}_i = (x_i, y_i, z_i)^T \in \mathbb{R}^3$ représentant sa position dans l'espace et éventuellement sa couleur $\mathbf{c}_i = (r_i, g_i, b_i)^T \in \mathbb{R}^3$. La notation $v_i \sim v_j$ est utilisée pour représenter 2 nœuds adjacents dans $\mathcal{G}$ (i.e, $(v_i, v_j) \in \mathcal{E}$).

2.1. Déscripteur local de la surface sous forme de patch adaptatif

Nous définissons un nouveau concept d'adaptabilité des patchs permettant de mieux considérer la configuration locale du voisinage. Ce patch est une imagette représentant le champ des hauteurs associé au plan tangent sur le nœud. Ceci peut être considéré comme une extension des travaux [13, 9, 14] avec l'adaptabilité de la taille afin de mieux refléter les propriétés géométriques locales. Le choix d'utiliser un patch 2D comme descripteur local d'une surface 3D est justifié par le succès de l'auto-similarité définie dans le traitement des images [15]. Afin de construire un patch sur un nœud cible, nous estimons le plan tangent et construisons une sous-region de ce dernier. Ensuite, les noeuds voisins appartenant à une sphère centrée en ce nœud cible sont projetés sur le plan tangent. Ces projections serviront à la fois à définir la taille du patch et à remplir ses cellules par la somme de leurs hauteurs. Les détails de ce processus sont détaillés ci dessous.

2.1.1. Estimation du plan tangent

Pour chaque nœud v_i sur la surface du maillage, nous calculons son vecteur normal $\mathbf{z}(v_i)$ et ses vecteurs directionnels $\mathbf{x}(v_i)$ et $\mathbf{y}(v_i)$ estimant le plan tangent 2D. Classiquement, l'ACP de la matrice de covariance des nœuds voisins à v_i est considérée [16]. Les nœuds appartenant à une sphère $S_\varepsilon(v_i) = \left\{v_j \mid ||\mathbf{p}_j - \mathbf{p}_i||_2^2 \leq \varepsilon\right\}$ centrée en v_i avec un rayon ε défini d'une manière empirique sont pris en compte. Evidemment, un rayon ε défini automatiquement en fonction des aspects intrinsèques de la surface traitée serait plus adéquat. Ceci constitue une perspective de nos futurs travaux. Soient $\bar{\mathbf{p}} = \frac{1}{|S_\varepsilon(v_i)|} \sum_{v_j \in S_\varepsilon(v_i)} \mathbf{p}_j$ et $C(v_i) = \frac{1}{|S_\varepsilon(v_i)|} \sum_{v_j \in S_\varepsilon(v_i)} (\mathbf{p}_j - \bar{\mathbf{p}})(\mathbf{p}_j - \bar{\mathbf{p}})^T$ respectivement le centre de gravité de $S_\varepsilon(v_i)$ et sa matrice de covariance. A partir de cette matrice, les vecteurs propres sont déduits et utilisés pour estimer le vecteur normal $\mathbf{z}(v_i)$ et les deux vecteurs directeurs $\mathbf{x}(v_i)$ et $\mathbf{y}(v_i)$ formant une base orthogonale du plan tangent $\mathbf{P}(v_i)$. Par ailleurs, les vecteurs représentant les normales pourraient avoir des directions différentes (vers l'extérieur de la surface ou vers l'intérieur). Afin de rendre l'orientation des normales uniforme sur tout le maillage, nous propageons l'orientation d'une normale, choisie arbitrairement, aux normales des points voisins par l'intermédiaire du parcours de l'arbre de poids minimal généré pour le maillage [17].

2.1.2. Construction du patch

Afin de construire le patch adaptatif, les nœuds $\mathbf{p}_j$ contenus dans la sphère $S_\varepsilon(v_i)$ sont projetés sur le plan 2D $\mathbf{P}(v_i)$ défini par les vecteurs directionnels associés, donnant lieu ainsi à des nœuds projetés $\mathbf{p'}_j$. Reste à définir la taille du patch. Pour cela, nous proposons une configuration dynamique dans laquelle la taille du patch n'est pas liée au rayon ε de la sphère $S_\varepsilon(v_i)$, mais plutôt à la distance maximale entre toutes les projections 2D des nœuds suivant les axes x et y. Les dimensions du patch sont déterminées par

$$\mathrm{T}_d(v_i) = \max_{(\mathbf{p'}_j, \mathbf{p'}_k) \in \mathbf{P}(v_i)} (||\mathbf{p'}_j^d - \mathbf{p'}_k^d||_2^2) \qquad (1)$$

où d représente la coordonnée x ou y, $\mathbf{p'}_j^d$ la d-ème coordonnée de $\mathbf{p'}_j$, et $||.||_2$ la norme Euclidienne. Ainsi le patch en v_i est représenté par un rectangle (ou carré selon la configuration du voisinage) de taille $\mathrm{T}_x(v_i) \times \mathrm{T}_y(v_i)$ centré en v_i. Il est à noter que classiquement, un patch est représenté par un carré de taille fixe [13, 9]. Ce patch est ensuite divisé en $l \times l$ cellules et chaque nœud projeté est affecté à la cellule dont le centre lui est proche. Finalement on affecte à chaque cellule $\mathcal{P}_i$ du patch la valeur absolue de la somme des hauteurs de projections $\sum_{\mathbf{p'}_j \in \mathcal{P}_i} ||(\mathbf{p}_j - \mathbf{p'}_j)||_2^2$ des nœuds qui lui sont associés. Nous aboutissons alors à une imagette reflétant la géométrie avoisinante du nœud v_i : chaque nœud est alors décrit par un vecteur caractéristique $\boldsymbol{\mathcal{P}}(v_i)$ de toutes les cellules du patch. Cette approche est robuste pour des surfaces à faible résolutions.

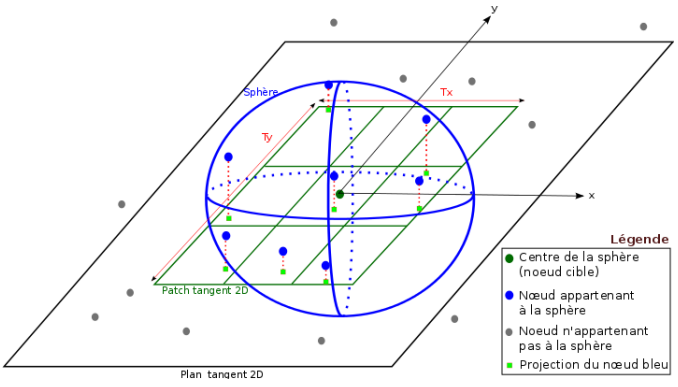


Fig. 1. Patch local adaptatif : les points bleus représentent les nœuds situés dans la sphère S_ε dont le centre est représenté par un nœud cible en vert foncé. Les nœuds en gris n'appartiennent pas à la sphère S_ε. Ainsi seuls les nœuds en bleu sont projetés sur le patch 2D. Les projections sous forme de carrés fluorescents représentent les vecteurs 2D projetés.

2.2. Saillance visuelle mono-échelle

Pour calculer la saillance mono-échelle d'un nœud v_i, une mesure de similarité entre son patch et les patchs associés

à ses voisins adjacents est requise. La similarité affectée au poids de l'arête $(v_i, v_j) \in \mathcal{E}$ est définie par :

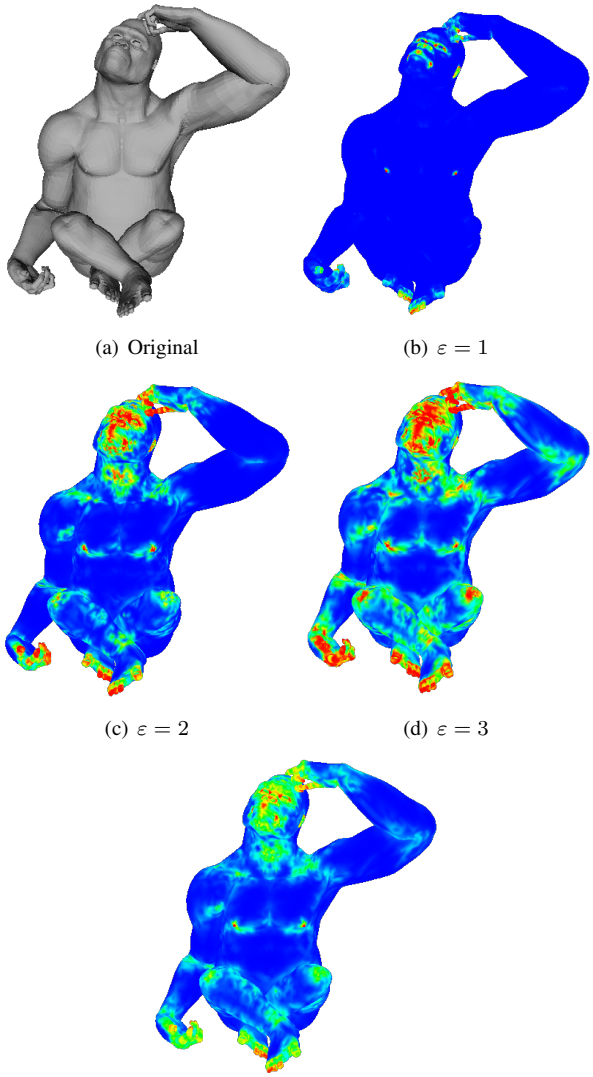
$$w_{\mathcal{P}}(v_i, v_j) = \exp\left[-\frac{\kappa(v_j) \cdot ||\boldsymbol{\mathcal{P}}(v_i) - \boldsymbol{\mathcal{P}}(v_j)||_2^2}{\sigma_{\mathcal{P}}(v_i) \cdot \sigma_{\mathcal{P}}(v_j) \cdot ||\mathbf{p}_i - \mathbf{p}_j||_2^2}\right] \quad (2)$$

où $\boldsymbol{\mathcal{P}}(v_i) \in \mathbb{R}^{l \times l}$ représente le vecteur des hauteurs de projections accumulées 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 and v_j. Nous proposons de calculer le paramètre d'échelle localement $\sigma_{\mathcal{P}}(v_i)$. En effet, l'utilisation d'un paramètre d'échelle spécifique pour chaque nœud du maillage nous permet de mieux prendre en considération la distribution locale autour de chaque nœud. Le paramètre d'échelle est défini par $\sigma_{\mathcal{P}}(v_i) = \max_{v_k \sim v_i}(||\mathbf{p}_i - \mathbf{p}_k||_2^2)$. Nous avons également testé la différence entre les patchs, cependant, nous avons remarqué que l'utilisation de la différence entre les coordonnées 3D mène à de meilleurs résultats. Finalement, la saillance visuelle mono-échelle d'un nœud v_i est définie par son degré dans $\mathcal{G}$ par :

$$\text{ss-saillance}_{\mathcal{P}}(v_i) = \frac{1}{|v_j \sim v_i|} \sum_{v_i \sim v_j} w_{\mathcal{P}}(v_i, v_j) \quad (3)$$

2.3. Saillance visuelle muti-échelle

Afin de mieux considérer l'aspect hiérarchique de la vision humaine, nous proposons d'estimer la saillance sur différentes échelles. En effet, sur les petites échelles, la mesure de saillance détectera les détails acérés et fins, tandis que sur les grandes échelles, elle fera ressortir les larges zones saillantes (voir Figure 2 (b-d)). Notre objectif est la détection de la saillance sur différentes échelles afin d'être robuste face au bruit qui est perceptible uniquement sur quelques échelles. Pour intégrer l'aspect multi-échelle dans notre approche, nous avons observé que la taille du voisinage sphérique $S_\varepsilon(v_i)$ peut être utilisée en tant que paramètre d'échelle et que l'élargissement du voisinage est similaire à un lissage [16]. Ainsi, nous varions le rayon ε de la sphère S_ε pour définir le patch local et considérons 3 différents voisinages (i.e. rayons). Le calcul de la saillance mono-échelle est effectué pour chaque voisinage. Avant de fusionner les cartes de saillance obtenues, nous calculons l'entropie sur chacune des cartes afin de mesurer le désordre et la disparité de l'information relative à la saillance. Pour cela, sur une échelle donnée k, nous calculons l'histogramme H^k des valeurs de la saillance des nœuds afin de quantifier la probabilité d'obtenir la valeur de la saillance $i : P_i^k = H^k(i)/|V|$, où H_i^k représente le nombre de nœuds ayant la valeur de saillance i sur une échelle k. Ensuite, l'entropie par échelle est donnée par $E_k = -\sum_i P_i^k * \log P_i^k$. En pondérant la saillance de chaque nœud par l'entropie lors de la fusion des différentes échelles, nous obtenons une carte de saillance multi-échelle robuste qui considère la disparité de la saillance sur chaque échelle. La saillance multi-échelle est



(a) Original (b) $\varepsilon = 1$ (c) $\varepsilon = 2$ (d) $\varepsilon = 3$ (e) Saillance multi-échelle

Fig. 2. Influence du rayon ε sur le calcul de la saillance visuelle mono-échelle(b-d) multi-échelle (e).

obtenue par :

$$\text{ms-saillance}_{\mathcal{P}}(v_i) = \frac{\sum_{k=1}^{K} \text{ss-saillance}^k(v_i) \cdot E_k}{\sum_{k=1}^{K} E_k} \quad (4)$$

où k est l'indice de l'échelle, et K représente le nombre d'échelles.

2.4. Extension aux maillages 3D colorés

Dans les précédentes sections, nous avons considéré uniquement les coordonnées $\mathbf{p}_i$ qui sont associés aux nœuds v_i. Cependant, avec les récents scanners 3D, il est dorénavant possible d'acquérir simultanément une couleur par nœud et donc un vecteur de couleurs RGB $\mathbf{c}_i$ est alors associé aux dif-

férents nœuds. Pour étendre notre mesure de saillance multi-échelle pour les maillages 3D colorés [18], nous procédons de la manière suivante. Un patch est construit similairement aux maillages non-colorés, cependant ses cellules $\mathcal{P}_i$ sont remplies avec la moyenne RGB des couleurs des nœuds projetés $\frac{1}{|\mathbf{p'}_j \in \mathcal{P}_i|} \sum_{\mathbf{p'}_j \in \mathcal{P}_i} \mathbf{c'}_j$, definissant alors un vecteur-couleur représentant le patch local sur chaque nœud. Les arêtes sont pondérées par :

$$w_{\mathcal{C}}(v_i, v_j) = \exp\left[-\frac{||\boldsymbol{\mathcal{C}}(v_i) - \boldsymbol{\mathcal{C}}(v_j)||_2^2}{\sigma_{\mathcal{C}}(v_i) \cdot \sigma_{\mathcal{C}}(v_j) \cdot l^2}\right] \quad (5)$$

avec $\sigma_{\mathcal{C}}(v_i) = \max_{v_k \sim v_i}(||\boldsymbol{\mathcal{C}}_i - \boldsymbol{\mathcal{C}}_k||_2)$. Nous pouvons alors définir de la même manière que précédemment une mesure de saillance mono-échelle pour les maillages colorés :

$$\text{ss-saillance}_{\mathcal{C}}(v_i) = \frac{1}{|v_j \sim v_i|} \sum_{v_i \sim v_j} w_{\mathcal{C}}(v_i, v_j) \quad (6)$$

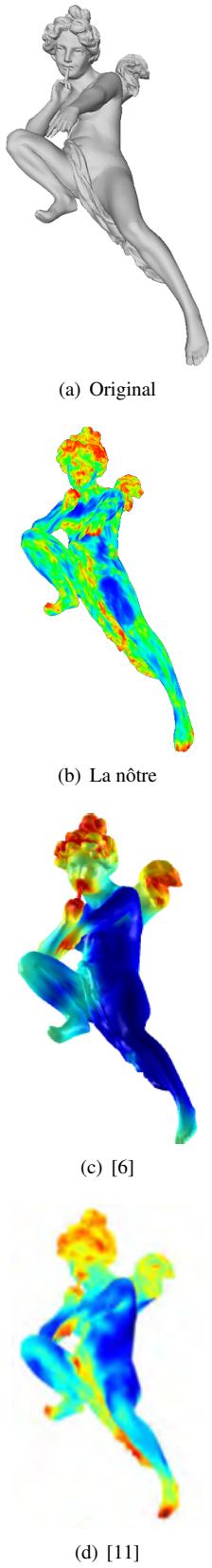
et déduire $\text{ms-saillance}_{\mathcal{C}}(v_i)$ qui représentera la valeur de saillance colorimétrique sur un nœud.

3. RÉSULTATS EXPERIMENTAUX

Dans cette section, nous présentons les résultats de notre mesure de saillance sur les maillages 3D non colorés et comparons ces derniers avec l'état de l'art. Les résultats pionniers de la détection de la saillance sur les maillages colorés sont également analysés. Sur toutes les expérimentations, un patch est divisé en 27×27 cellules. Le nombre d'échelles K dans le calcul de la saillance multi-échelle est fixé à 3. Ces 3 échelles sont obtenues avec les sphères $S_\varepsilon(v_i)$ de rayons ε_0, $2\varepsilon_0$, et $3\varepsilon_0$. Pour visualiser la saillance détectée, nous utilisons une palette de couleurs. Les couleurs chaudes (rouges et jaunes) représentent une forte saillance et les couleurs froides (vertes et bleues) une faible saillance.

Premièrement, nous commençons par présenter les résultats de la saillance mono-échelle. La figure 2 présente les saillances calculées avec trois différentes valeurs de ε pour un maillage original. Comme prévu, lorsque le voisinage est réduit (ε petit), les regions restreintes et très différentes de leurs voisinage local sont considérées comme saillantes, alors que lorsque ε est important, la saillance détectée est plus dispersée et de larges régions sont jugées saillantes. Cependant, nous pouvons remarquer que la saillance est saturée au niveau des régions très saillantes (les yeux du maillage Gorille). L'aspect multi-échelle permettra de remédier à ce problème en fusionnant judicieusement les cartes de saillance mono-échelles.

Deuxièmement, nous avons comparé notre approche avec l'approche référence de [6] et avec la plus récente de [11]. La figure 3 montre les résultats de cette comparaison. La surface du maillage 3D Angel est complexe dans la mesure où elle contient plusieurs extrémités. Elle comprend également à



(a) Original

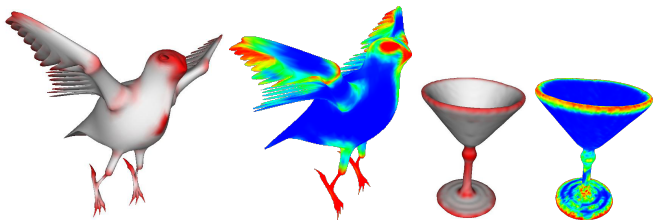
(b) La nôtre

(c) [6]

(d) [11]

Fig. 3. Comparaison de notre saillance visuelle détectée sur le maillage 3D Ange avec celles de l'état de l'art.

la fois des régions rugueuses et lisses. Nous pouvons remarquer l'extrémité de l'écharpe présentant beaucoup de fluctuations. Celle-ci est considérée comme saillante par l'approche de [11] et la nôtre alors que la méthode de [6] l'a considère comme non saillante. Les discontinuités au niveau des yeux, des bras, de la hanche et du ventre sont jugées saillantes par la méthode de [11] et la nôtre contrairement à celle de [6]. Cette comparaison montre que notre approche fourni une détection de la saillance plus précise et détaillée que les méthodes de référence qui estiment la saillance d'une manière assez grossière. Nous pensons que ce constat est un apport considérable de notre approche et que des résultats similaires aux méthodes de la littérature peuvent être facilement obtenus par un simple lissage de notre carte de saillance.



(a) (b) (c) (d)

Fig. 4. Comparaison avec la vérité terrain [19]. (a),(c) saillance visuelle issue de la vérité terrain, (b),(d) notre approche.

Troisièmement, nous évaluons l'exactitude et la finesse de notre approche de détection de la saillance par une comparaison de nos résultats obtenus sur les maillages 3D non colorés de la base SHREC 2007 Watertight Models [20] avec leurs pseudo vérité terrain obtenues à partir d'un modèle analytique issu d'experimentations visuelles. La figure 4 présente cette comparaison. Nous pouvons remarquer visuellement que la saillance détectée par notre approche correspond bel et bien à la vérité terrain. Ceci confirme encore une fois la précision de notre approche.

Quatrièmement, nous présentons les résultats innovants de notre approche de détection de la saillance sur des maillages colorés, ce qui, à notre connaissance, n'a jamais été traité auparavant. La figure 5 présente ces résultats sur un scan 3D d'une tête humaine et un canard empaillé. Les figures 5(b),(g) présentent les saillances géométriques obtenues qui prennent en compte uniquement les coordonnées des nœuds pour construire les patchs. Nous pouvons remarquer que les régions planes apparaissent non saillantes alors que les régions fluctuantes sont jugées saillantes. Par ailleurs, les figures 5(d),(i) présentent les saillances colorimétriques se basant uniquement sur les couleurs des nœuds lors de la construction des patchs. Ces résultats sont très différents des saillances géométriques. En effet, les régions ayant des variations de couleurs importantes apparaissent saillantes (le collier sur le coup du canard est dorénavant considéré très saillant). Le même résultat est présenté sur au niveau de la chemise (figure 5(d)). En effet, la chemise comprend des parties blanches dans un fond noir. Les figures 5(e),(j) présentent une combinaison des deux résultats associés à la saillance géométrique et à la saillance colorimétrique.

4. CONCLUSION

Dans ce papier, nous avons présenté une nouvelle approche répondant au problème difficile de l'estimation de la saillante des maillages 3D. Afin de décrire les régions saillantes, des descripteurs locaux sous forme de patchs adaptatifs reflétant les caractéristiques du voisinage géométrique ou colorimétrique sont mis en oeuvre. Ces descripteurs sont utilisés comme une base pour la mesure de la similarité et intégrés dans un calcul multi-échelle pondéré. L'approche va au delà de l'état de l'art en quantifiant la saillance des maillages 3D *colorés*, qui en l'occurrence, n'a jamais été traitée auparavant.

5. REFERENCES

[1] Zhi Liu, Wenbin Zou, and Olivier Le Meur, "Saliency tree : A novel saliency detection framework," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 1937–1952, 2014.

[2] Chang Ha Lee, Amitabh Varshney, and David W. Jacobs, "Mesh saliency," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 659–666, 2005.

[3] Yitian Zhao, Yonghuai Liu, Ran Song, and Min Zhang, "A saliency detection based method for 3d surface simplification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 889–892.

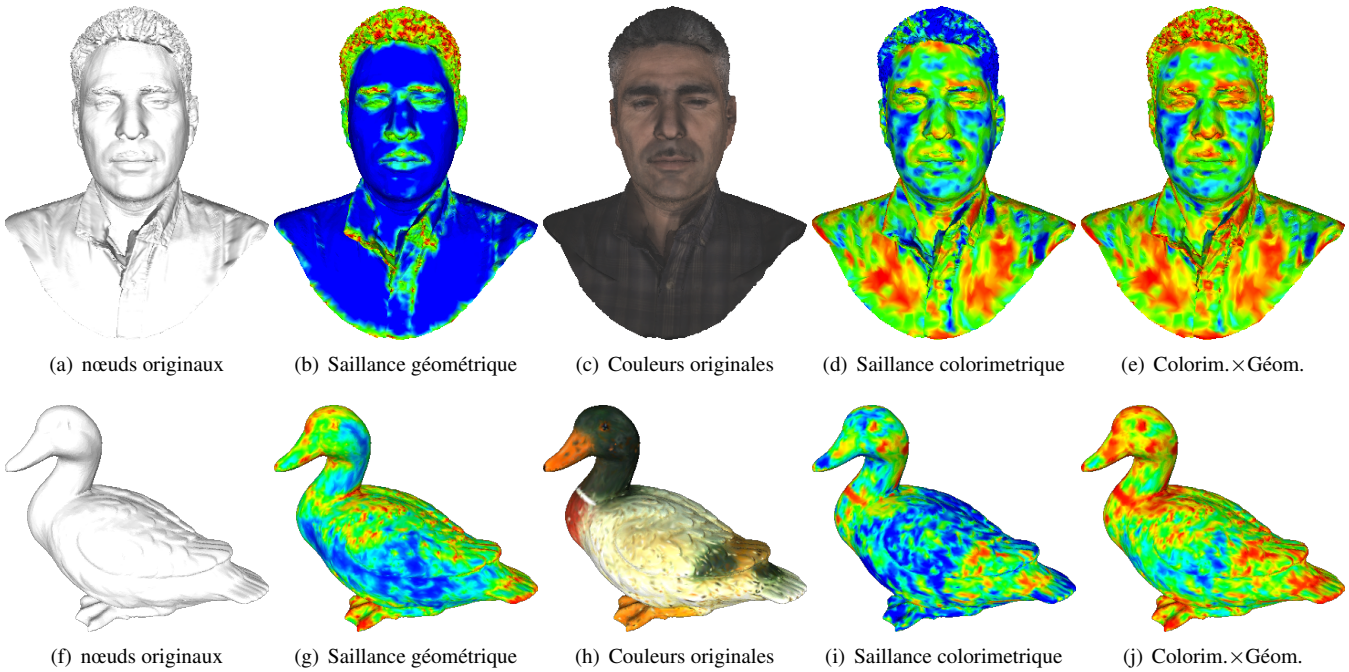
[4] Jinliang Wu, Xiaoyong Shen, Wei Zhu, and Ligang Liu, "Mesh saliency with global rarity," *Graphical Models*, vol. 75, no. 5, pp. 255 – 264, 2013.

[5] Shixiang Jia, Caiming Zhang, Xuemei Li, and Yuanfeng Zhou, "Mesh resizing based on hierarchical saliency detection," *Graphical Models*, vol. 76, no. 5, pp. 355 – 362, 2014.

[6] George Leifman, Elizabeth Shtrom, and Ayellet Tal, "Surface regions of interest for viewpoint selection.," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 414–421.

[7] 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.

[8] Andrew Edie Johnson and Martial Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions onPattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999.



(a) nœuds originaux (b) Saillance géométrique (c) Couleurs originales (d) Saillance colorimetrique (e) Colorim.×Géom.

(f) nœuds originaux (g) Saillance géométrique (h) Couleurs originales (i) Saillance colorimetrique (j) Colorim.×Géom.

Fig. 5. Saillance visuelle des maillages 3D colorés.

[9] A. Maximo, R. Patro, A. Varshney, and R. Farias, "A robust and rotationally invariant local surface descriptor with applications to non-local mesh processing," *Graphical Models*, vol. 73, no. 5, pp. 231 – 242, 2011.

[10] Pingping Tao, Junjie Cao, Shuhua Li, Xiuping Liu, and Ligang Liu, "Mesh saliency via ranking unsalient patches in a descriptor space," *Computers & Graphics*, vol. 46, no. 1, pp. 264 – 274, 2015.

[11] Ran Song, Yonghuai Liu, Ralph R. Martin, and Paul L. Rosin, "Mesh saliency via spectral processing," *ACM Transactions on Graphics*, vol. 33, no. 1, pp. 6 :1–6 :17, 2014.

[12] Olivier Lézoray and Leo Grady, *Image Processing and Analysis with Graphs : Theory and Practice*, Digital Imaging and Computer Vision. CRC Press / Taylor and Francis, 2012.

[13] Julie Digne, "Similarity based filtering of point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 73 –79.

[14] Francois Lozes, Abderrahim Elmoataz, and Olivier Lezoray, "Partial difference operators on weighted graphs for image processing on surfaces and point clouds," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3896–3909, 2014.

[15] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel, "Image denoising methods. A new nonlocal principle," *SIAM Review*, vol. 52, no. 1, pp. 113–147, 2010.

[16] Mark Pauly, Richard Keiser, and Markus H. Gross, "Multi-scale feature extraction on point-sampled surfaces," *Computer Graphics Forum*, vol. 22, no. 3, pp. 281–290, 2003.

[17] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, and Werner Stuetzle, "Surface reconstruction from unorganized points," *ACM SIGGRAPH Computer Graphics*, vol. 26, no. 2, pp. 71–78, 1992.

[18] A. Nouri, C. Charrier, and O. Lézoray, "Multi-scale saliency of 3d colored meshes," in *International Conference on Image Processing (IEEE)*, 2015, vol. to appear.

[19] Xiaobai Chen, Abulhair Saparov, Bill Pang, and Thomas Funkhouser, "Schelling points on 3d surface meshes," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 29 :1–29 :12, 2012.

[20] Daniela Giorgi, Silvia Biasotti, and Laura Paraboschi, "Shrec : Shape retrieval contest : Watertight models track," http ://watertight.ge.imati.cnr.it/.

Schéma de compression d'images intégrales basé sur l'extraction de vues

A. Dricot [1,2] J. Jung [1] M. Cagnazzo [2] B. Pesquet [2] F. Dufaux [2]

[1] Orange Labs [2] Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI

antoine.dricot;joelb.jung;{@orange.com}
cagnazzo;beatrice.pesquet;frederic.dufaux;{@telecom-paristech.fr}

Résumé

L'imagerie intégrale est une technologie qui permet de capturer une scène sous la forme d'une représentation dite ligth-field. Cette représentation offre plusieurs points de vue de la scène et permet d'éliminer beaucoup d'inconvénients connus en stéréoscopie et en auto-stéréoscopie par exemple. Les images intégrales ont cependant une résolution élevée et une structure en micro-images difficiles à encoder. Cet article propose un schéma de compression d'images intégrales basé sur l'extraction de vues. Les gains moyens en BD-rate par rapport à la référence HEVC sont de 15.7% (jusqu'à 31.3%). Les paramètres de ce schéma peuvent prendre un nombre étendu de valeurs. Une recherche exhaustive parmi ces valeurs permet d'abord de donner les résultats de la meilleure configuration. Puis un critère d'optimisation débit-distorsion est proposé pour éviter la recherche exhaustive, permettant de réduire le temps d'encodage tout en préservant les gains. Finalement, l'étude plus fine de l'impact des différents paramètres permet de réduire encore davantage le temps de codage.

Mots clefs

Imagerie Intégrale, Imagerie Plénoptique, Holoscopie, Compression d'Image, Codage Vidéo, Extraction de vues

1 Introduction

Le développement technologique lié à la vidéo 3D tend à créer des expériences de visualisation de plus en plus réalistes et immersives. Les technologies vidéo 3D actuellement disponibles sur le marché présentent cependant plusieurs limitations. L'utilisation de lunettes en stéréoscopie crée un manque de confort. De plus, le conflit entre les distances d'accommodation et de convergence n'est pas naturel pour le système visuel humain et peut causer des gènes. Malgré un nombre plus élevé de vues, les systèmes autostéréoscopiques sont limités par des éléments de perception non-naturels, comme par exemple le manque d'une parallaxe de mouvement fluide (i.e. une visualisation continue quand l'utilisateur se déplace devant l'écran), qui est un élément clé dans la perception de la profondeur [1].
L'imagerie intégrale est une technologie basée sur la photographie plénoptique [2]. Cette technique permet d'obtenir une représentation du *light-field* d'une scène [3], qui permet d'éliminer la plupart des inconvénients existants dans les technologies 3D actuelles (e.g. le conflit convergence-accommodation). L'acquisition en imagerie intégrale est basée sur l'utilisation d'un dispositif lenticulaire, constitué d'un ensemble de micro-lentilles, placé devant le périphérique de capture (i.e. la caméra ou son capteur). Chaque micro-lentille produit une micro-image (MI), et chaque MI contient l'information provenant de plusieurs angles de vue. L'image intégrale résultante correspond à un ensemble de MIs, comme illustré en Figure 1.
Des caméras plénoptiques sont déjà disponibles sur le marché, avec notamment *Lytro* [4] et *Raytrix* [5]. De plus, plusieurs institutions ont déjà montré de l'intérêt pour les technologies liées au *light-field* en travaillant sur des systèmes d'affichage [6] pour une visualisation sans lunette, visant à offrir une expérience réaliste et immersive. La téléprésence est un cas d'application cible intéressant, ainsi que la réalité virtuelle avec la navigation dans une scène en relief [7].
Pour gérer efficacement la résolution élevée des images intégrales et les caractéristiques spécifiques de leur structure, de nouvelles technologies de codage sont nécessaires. On propose dans cet article un schéma de compression efficace qui exploite les techniques d'extraction de vues afin de créer une image intégrale résiduelle qui est encodée. La performance de ce schéma dépend de plusieurs paramètres pouvant prendre un nombre étendu de valeurs. On propose donc un ensemble de méthodes itératives pour sélectionner la configuration la plus efficace en jouant sur le compromis débit-distorsion et complexité.
La suite de cet article est organisée comme suit. En Section 2, les méthodes d'extractions de vues existantes sont décrites, et l'état de l'art des méthodes de codage d'images intégrales est présenté. Le schéma de compression proposé est décrit puis les résultats expérimentaux sont montrés en Section 3. Les conclusions sont tirées dans la Section 4.

2 État de l'art

2.1 Extraction de vues

Plusieurs méthodes pour extraire des vues d'une image intégrale sont décrites dans [9]. La méthode basique extrait un *patch* (une zone carrée de pixels) de chaque MI,

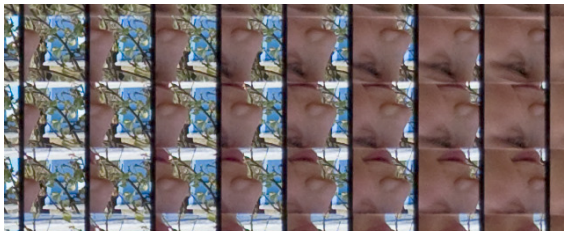


Figure 1 – *Micro-Images (MIs) - Laura [8]*

comme illustré en Figure 2 *(gauche)*. Cette méthode est basée sur les caractéristiques de la *focused plenoptic camera* [10] pour laquelle chaque MI contient de l'information spatiale et de l'information angulaire. Une méthode plus basique consiste à utiliser un patch de taille 1×1, i.e. un pixel par MI. La taille du patch définit la profondeur (distance à la caméra) dans la scène du plan sur lequel la mise au point sera faite dans la vue extraite : plus le patch est large, moins cette distance est élevée. L'angle de vue dépend de la position relative du patch dans la MI. Une méthode plus avancée permet d'atténuer les effets de blocs en lissant les transitions entre les patchs adjacents. Les pixels entourant chaque patch sont moyennés avec une pondération (les pixels les plus éloignés du centre ont un poids plus faible). Une méthode d'estimation de disparité (basée sur *block-matching*) est proposée dans [10] pour obtenir la profondeur relative des objets dans chaque MI. Une valeur de disparité par MI est estimée, correspondant à la taille du patch à utiliser. Dans les vues extraites par cette méthode dite *disparity-assisted patch blending extraction* (DAPBe [9]), tous les objets sont nets car les tailles de patch sont adaptées à la profondeur.

2.2 Codage d'images intégrales

Les images intégrales doivent avoir une grande résolution pour obtenir un nombre élevé de vues différentes avec une résolution suffisante. De plus, la structure en micro-images (MIs) donne un aspect de grille qui rend l'image difficile à encoder (voir Fig. 1). Une première approche consiste à appliquer la Transformée en Cosinus Discrète (*Discrete Cosine Transform* - DCT) sur les MIs, suivie d'une quantification et d'un codage sans perte [11]. Les corrélations

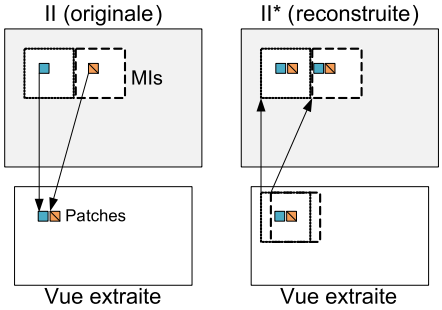


Figure 2 – *Extraction (gauche), Reconstruction (droite)*

inter-MIs peuvent également être exploitées en utilisant une *3D-DCT* sur les MIs superposées en un volume [12]. Dans [13], une transformée en ondelettes (*Discrete Wavelet Transform* - DWT) est appliquée aux MIs, suivie d'une DCT sur les blocs de coefficients transformés (transformée hybride à 4 dimensions). Les approches basées sur les transformées correspondent bien à la structure en MIs mais donnent des gains en compression limités en comparaison aux encodeurs standards actuels (H.264/AVC [14] et HEVC [15]). Dans [16] et [17], les images intégrales sont décomposées en vues, qui sont encodées avec l'encodeur MVC [18]. Cette approche est efficace sur les images générées par ordinateur (i.e. dont les MIs sont parfaitement alignées sur les pixels) mais reste limitée pour les contenus naturels. Le mode *Self-similarity* [19] est une autre approche basée sur le même principe que le mode *Intra Block Copy* [20], qui exploite les corrélations spatiales non-locales entre les MIs. Un algorithme de *block-matching* est utilisé, comme pour le mode inter de H.264/AVC et HEVC, mais au sein de l'image courante (zone causale déjà encodée). Cet outil apporte de bons gains pour les images fixes mais reste limité pour les séquences quand la prédiction temporelle est activée.

Un schéma de codage scalable est proposé dans [21]. Ce schéma offre une fonctionnalité intéressante de scalabilité côté écran (i.e. un flux adapté aux systèmes d'affichage 2D, multi-vues, et holoscopiques). La couche 0 correspond à la vue centrale, la couche 1 correspond à un ensemble de vues additionnelles, et la couche 2 est l'image intégrale. Cette scalabilité a donc un coût, car des vues supplémentaires sont encodées. Un schéma de prédiction inter-couches est proposé pour réduire le débit dédié à la couche 2, dans lequel une image intégrale est partiellement reconstruite à partir des vues (couche 1) et ajoutée aux images de référence.

Dans la Section 3, on propose un schéma de codage original pour les images intégrales. Bien qu'il soit basé sur l'extraction de vue, et permette une certaine forme de scalabilité, il diffère des méthodes scalables existantes : son but principal est l'efficacité de codage. Il utilise avantageusement le processus d'extraction pour reconstruire un prédicteur fiable et créer une image intégrale résiduelle qui est encodée.

3 Schéma de codage proposé

Le schéma de compression proposé (Fig. 3) est décrit dans cette section. Dans ce schéma, une image résiduelle $\mathrm{II_R}$ est encodée avec HEVC (flux *résiduel*). $\mathrm{II_R}$ est la différence entre l'image originale II et une image reconstruite II^*. II^* est reconstruite à partir de vues extraites de l'image originale II. Le nombre de vues n'est pas limité. Les valeurs de disparité utilisées pour l'extraction et la reconstruction sont encodées sans compression, et les vues extraites sont encodées avec 3D-HEVC (flux *vues*). De par leur résolution réduite, les vues représentent un nombre de bits réduit à encoder, en comparaison à II. De plus, elles ont un as-

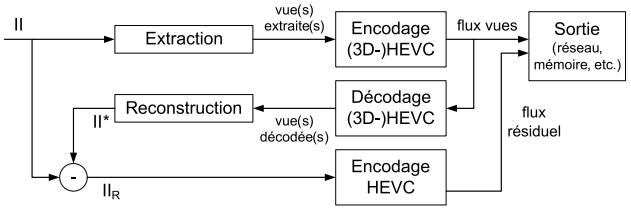


Figure 3 – *Schéma de codage proposé - encodeur*

pect d'images naturelles qui est moins couteux à encoder, en opposition à l'aspect donné par les MIs. Pour obtenir des vues avec un tel aspect, des méthodes d'extractions avancées sont utilisées (voir Sec. 2.1), basée sur l'extraction de patch et avec une moyenne pondérée sur les bords, ce qui empêche une reconstruction parfaite pour II^*. L'information manquante, correspondant à la différence entre II et II^*, est récupérée dans II_R. Par définition, pour une image reconstruite II^* proche de l'originale II, la différence donne des valeurs proches de zéro (un exemple de l'aspect de l'image résiduelle est illustré en Fig. 4). II_R a donc un aspect *lisse* avec peu de variations, qui est plus facile à encoder avec HEVC que II.

Comme l'illustre la Figure 2 *(droite)*, lors de la reconstruction de II^* à partir de vues, certains pixels provenant d'angles de vue différents sont remplacés par des pixels adjacent provenant de la même vue. Cependant la transformation d'un objet en changeant d'angle de vue n'est pas limitée à une simple translation (disparité) mais implique également un mouvement angulaire. Des erreurs sont donc introduites. Un filtrage passe-bas (e.g. moyenneur) est appliqué aux vues décodées avant la reconstruction pour atténuer ces erreurs. Les hautes fréquences dans la vue (qui correspondent par exemple au bruit et aux détails) sont filtrées tout en préservant la forme des objets.

Côté décodeur, les vues sont décodées et utilisées pour reconstruire II^*, et II_R est décodée puis ajoutée (somme) à II^* pour obtenir l'image de sortie.

Il existe un compromis entre le débit et la qualité des vues, et le débit de II_R. II^* doit être la plus proche possible de II afin de minimiser le coût de II_R, sans augmenter trop le coût des vues. Plusieurs combinaisons sont possibles pour les paramètres suivants : le QP utilisé pour encoder les vues ($\mathrm{QP_V}$), le QP utilisé pour encoder l'image résiduelle ($\mathrm{QP_R}$), et la taille (en pixels) du filtre passe-bas appliqué

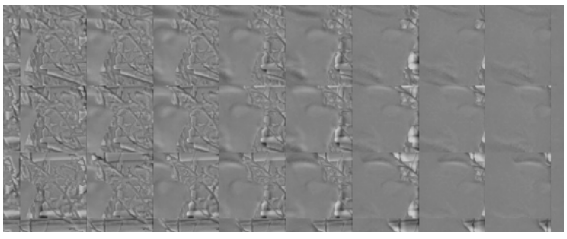
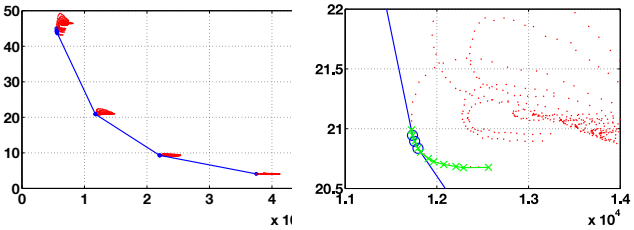


Figure 4 – *Micro-Images (MIs) in a residual image - Laura [8]*



(a) Toutes les configurations (b) Focus sur $\mathrm{QP_R}$ 20

Figure 5 – *Points débit-distorsion pour toutes les configurations (Fountain) - horizontal : débit (kbits/image), vertical : EQM*

aux vues décodées (B). Dans la suite, on propose différents compromis entre débit-distorsion et complexité, en explorant plusieurs méthodes pour paramétrer ces valeurs.

3.1 Recherche exhaustive de la configuration optimale

Nos expérimentations incluent sept images fixes [8], listées dans le Tableau 1. Ces images sont rognées pour retirer les MIs incomplètes, et débarrassées des pixels de la grille correspondant aux limites entre les micro-lentilles [22]. La méthode dite *disparity-assisted patch blending extraction* (DAPBe [9]) est utilisée pour extraire une vue unique. Les encodages de la vue et de l'image résiduelle sont effectués avec HEVC (HM14.0) dans la configuration *Intra main* [23], et les valeurs de disparité sont codées sur 4 bits par MI. Les résultats de compression sont donnés avec la métrique Bjøntegaard Delta (BD) rate [24]. L'ancrage de référence est II encodée avec HEVC sur l'intervalle de QP {25,30,35,40}, et les valeurs négatives représentent un gain sur la référence.

Comme en pratique, une très large partie du débit est dédiée à II_R, la valeur de $\mathrm{QP_R}$ est déterminée en fonction du débit (ou de la qualité) cible, et $\mathrm{QP_V}$ et B sont considérés comme des paramètres à optimiser pour un $\mathrm{QP_R}$ donné. Pour chaque $\mathrm{QP_R}$ dans l'intervalle {10,15,20,25}, des combinaisons de valeurs pour les paramètres $\mathrm{QP_V}$ et B (respectivement dans les intervalles {10,11,...,50} et {1,2,...,11}) sont testées par itérations, fournissant 1804 ($4 \times 41 \times 11$) points débit-distorsion (*RD-points*, illustrés par les points rouges dans la Figure 5). Pour chaque image, le Tableau 1 montre la configuration qui donne les meilleurs résultats en BD-rate.

Un gain moyen de 15.7% (jusqu'à 31.3% pour *Fredo*) est rapporté en utilisant les combinaisons optimales de paramètres. La valeur de $\mathrm{QP_V}$ augmente en rapport à $\mathrm{QP_R}$, offrant un compromis entre le débit de la vue et de II_R. Approximativement 97% du débit total est dédié à II_R en moyenne, principalement à cause de sa large résolution par rapport à la vue (e.g. pour *Fountain* 6512×4880 contre 960×720), qui représente les 3% restants (les valeurs de disparités utilisées pour l'extraction et la reconstruction représentent seulement 0.3%).

Les valeurs optimales pour $\mathrm{QP_V}$ et B sont ici sélection-

Image	BD-rate (%)	Pour chaque $\mathrm{QP_R}$ dans {10,15,20,25}	
		$\mathrm{QP_V}$	B
Fountain	-17.0	19 21 23 29	3 3 3 3
Fredo	-31.3	18 21 25 32	3 3 3 3
Jeff	-5.9	25 30 30 32	9 9 9 7
Laura	-11.2	22 25 27 31	4 4 4 4
Seagull	-13.7	20 21 25 29	3 3 3 3
Sergio	-23.6	19 19 24 32	4 2 2 2
Zenhgyun1	-7.5	25 26 30 32	9 9 9 7
Moyenne	-15.7		

Tableau 1 – *Gains BD-Rate avec les configurations optimales de* $\mathrm{QP_V}$ *et* B *pour chaque* $\mathrm{QP_R}$. *Les valeurs négatives représentent un gain sur la référence*

Image	BD-rate (%)	Temps (%)	Pour chaque $\mathrm{QP_R}$ dans {10,15,20,25}	
			$\mathrm{QP_V}$	B
Fountain	-17.0	48284	19 21 23 27	3 3 3 3
Fredo	-31.1	47067	18 21 25 28	3 3 3 3
Jeff	-5.9	48729	25 30 30 32	9 9 9 7
Laura	-11.2	49065	22 25 27 30	4 4 4 4
Seagull	-13.7	48836	19 21 25 29	3 3 3 3
Sergio	-23.5	48036	20 21 24 28	4 2 2 2
Zenhgyun1	-7.5	48554	25 26 31 30	9 9 9 7
Moyenne	-15.7	48367		

Tableau 2 – *BD-Rate, variations du temps de codage et configurations associées pour la méthode* $\mathrm{All_{RDO}}$

nées exhaustivement parmi les 1804 points, et dépendent de l'image testée. En Section 3.2, les résultats fournis par cette étude préliminaire sont utilisés pour déterminer un processus d'optimisation débit-distorsion (*Rate Distortion Optimisation - RDO*) qui sélectionne automatiquement les meilleures valeurs pour un $\mathrm{QP_R}$ donné.

3.2 Processus local d'optimisation débit-distorsion

La Figure 5 illustre les valeurs RD obtenues pendant la recherche exhaustive de la meilleure combinaison pour $\mathrm{QP_V}$, $\mathrm{QP_R}$ et B. On définit l'enveloppe convexe globale (GCH, en bleu) comme l'enveloppe convexe de tous les points, et l'enveloppe convexe locale (LCH, en vert) comme l'enveloppe convexe d'un ensemble de points ayant la même valeur de $\mathrm{QP_R}$. Les points optimaux pour un $\mathrm{QP_R}$ donné sont situés à l'intersection S entre LCH et GCH. D'après nos données expérimentales, on peut observer que pour chaque $\mathrm{QP_R}$, il existe au moins un point qui appartient à GCH (i.e. cette intersection n'est pas vide). Utiliser uniquement LCH peut donner des configurations sous-optimales, comme le montre sur la Figure 5 les points marqués d'une croix verte et qui sont situés sur la partie droite de GCH. GCH ne peut cependant être connue qu'en encodant l'image avec plusieurs $\mathrm{QP_R}$, ce qui multiplie le nombre de combinaisons testées.

L'objectif dans cette section est de pouvoir sélectionner la configuration, pour un $\mathrm{QP_R}$ donné, qui fournit des valeurs de débit et distorsion (respectivement R et D) qui minimisent le coût $D + \lambda R$, où λ est la pente de LCH dans S (donc de GCH). Dans la Figure 5, cela revient à trouver parmi les points marqués d'une croix, ceux qui sont également marqués d'un cercle.

La pente de GCH entre deux valeurs consécutives de $\mathrm{QP_R}$ évolue exponentiellement par rapport à $\mathrm{QP_R}$. Une estimation de $\lambda = f(\mathrm{QP_R})$ est donc possible. La fonction définie par $\lambda = 2^{a\mathrm{QP_R}+\mathrm{b}}$ (avec $a = 0.34$ et $b = -15.8$) a une excellente correspondance avec les données.

Une première méthode $\mathrm{All_{RDO}}$ est proposée, où les combinaisons de $\mathrm{QP_V}$ et B sont successivement testées pour un $\mathrm{QP_R}$ donné. La combinaison qui donne les valeurs de débit et distorsion minimisant le coût $D + \lambda R$ (avec λ déterminé comme ci-dessus) est sélectionnée. Les conditions de test sont telles que décrites en Section 3.1. Le Tableau 2 donne le BD-rate et la variation de temps de codage (pour chaque image et en moyenne) en référence à l'ancrage HEVC.

Les combinaisons sélectionnées par $\mathrm{All_{RDO}}$ sont très proches des configurations optimales déterminées dans la Section 3.1 (même valeurs pour B et seulement de légères différences pour quelques valeurs de $\mathrm{QP_V}$), et les gains de 15.7% sont préservés, ce qui montre la robustesse de l'estimation de $\lambda = f(\mathrm{QP_R})$. Le temps total d'encodage pour toutes les itérations est très large, avec une multiplication du temps de référence par 484 en moyenne. On note que les intervalles de valeurs testées pour $\mathrm{QP_V}$ et B ne sont pas utilisés entièrement et peuvent donc être réduits pour diminuer le nombre d'itérations. Deux variantes sont proposées dans la suite afin de réduire encore ce nombre.

Avec $\mathrm{B_{RDO}}$ et $\mathrm{B_{MSE}}$, les itérations sur B sont faites uniquement pour une valeur de $\mathrm{QP_V}$ (e.g. pour $\mathrm{QP_V} = 10$ dans nos expérimentations), et la meilleure valeur de B est conservée pour les autres itérations sur $\mathrm{QP_V}$. La meilleure valeur pour B est celle qui minimise le coût $D + \lambda R$ dans $\mathrm{B_{RDO}}$ (même processus d'optimisation que pour $\mathrm{All_{RDO}}$), et celle qui minimise l'Erreur Quadratique Moyenne (EQM) entre II^* et II dans $\mathrm{B_{MSE}}$. Le Tableau 3 montre les résultats BD-rate moyens et les variations en temps de codage associées pour ces méthodes.

Pour $\mathrm{B_{RDO}}$, le temps d'encodage total est réduit significativement (de 484 fois le temps de référence à 55 fois) car le nombre d'itérations est fortement réduit. Les gains de 15.7% sont préservés car B ne varie pas significativement par rapport à $\mathrm{QP_V}$. Les résultats de $\mathrm{B_{MSE}}$ montrent que le temps d'encodage peut encore être réduit (à 44 fois la référence) en sélectionnant B sans encoder l'image résiduelle à chaque itération, avec un gain presque aussi bon (15.3% en moyenne, e.g. avec une diminution de 1.7% pour *Seagull*, et 0.8% pour *Sergio*). On note que le nombre d'itérations sur $\mathrm{QP_V}$ peut être réduit encore en évitant de chercher sur tout l'intervalle {10,11,...,50}. Par exemple, on peut observer que le coût $D + \lambda R$ a un minimum local par rapport à

	$\mathrm{All_{RDO}}$	$\mathrm{B_{RDO}}$	$\mathrm{B_{MSE}}$	$\mathrm{QPV_{fixed}}$	$\mathrm{All_{fixed}}$
BD-rate	-15.7	-15.7	-15.3	-15.5	-8.5
Temps	48367	5526	4443	136	120

Tableau 3 – *Gains BD-Rate (%) et variations du temps de codage associée(%), en moyenne*

$\mathrm{QP_V}$ pour B et $\mathrm{QP_R}$ donnés. Les itérations peuvent donc être stoppées quand le coût commence à augmenter.

3.3 Détermination empirique de la configuration

Deux variantes supplémentaires sont définies, $\mathrm{QPV_{fixed}}$ et $\mathrm{All_{fixed}}$, dans lesquelles la valeur de $\mathrm{QP_V}$ est fixée empiriquement pour chaque $\mathrm{QP_R}$. Avec $\mathrm{QPV_{fixed}}$, B est sélectionnée par rapport à l'EQM entre II^* et II (comme pour $\mathrm{B_{MSE}}$ dans la Sec. 3.2), tandis que pour $\mathrm{All_{fixed}}$, la valeur de B est également fixée.

Les résultats de $\mathrm{QPV_{fixed}}$ dans le Tableau 3 montrent qu'assigner un $\mathrm{QP_V}$ à un $\mathrm{QP_R}$ réduit largement le temps d'encodage (seulement 1.4 fois la référence) et donne quand même un gain moyen proche de l'optimal avec 15.5%. Bien que le nombre d'images de test disponibles soit limité, les paramètres diffèrent peu d'une image à l'autre. Cette robustesse suggère des gains similaires sur d'autres images intégrales. Le temps de codage est de seulement 1.2 fois la référence pour $\mathrm{All_{fixed}}$. Par contre, le gain moyen descend à 8.5%, avec des pertes pour *Jeff* et *Zenhgyun1*. La valeur optimale de B dépend fortement du contenu de l'image et les itérations sur ce paramètre peuvent améliorer de manière significative la performance de codage, avec seulement une légère augmentation du temps de codage.

Des résultats préliminaires montrent des performances similaires entre le schéma proposé et le mode *Intra Block Copy* [20] (même principe que *Self-Similarity* [19]), et que combiner les deux méthodes améliore encore l'efficacité.

Le Tableau 4 montre les variations du temps d'encodage et de décodage du schéma proposé par rapport à la référence, ainsi que le pourcentage du temps total dédié à chaque tâche, pour l'image *Fountain* avec la méthode $\mathrm{QPV_{fixed}}$. Le schéma proposé a un temps d'encodage total de 1.3 fois le temps d'encodage de II avec HEVC (référence). 79% du temps est dédié à l'encodage de $\mathrm{II_R}$. L'extraction de la vue prend 7% du temps total, principalement à cause du temps dédié à l'estimation de disparité. Les onze itérations des étapes de filtrage, reconstruction, et différence représentent 12%. Le temps de décodage est de 2.4 fois le temps de référence, avec 46% dédié au décodage de $\mathrm{II_R}$. Il ne dépend pas du nombre d'itérations côté encodeur. La reconstruction (31%) et la somme de $\mathrm{II_R}$ et II^* (22%) représentent un pourcentage plus élevé au décodeur car le processus de décodage avec HEVC est bien plus rapide que l'encodage. Cette augmentation est plus élevé en bas débit, où le temps de décodage d'HEVC est encore réduit davantage, tandis que le temps de reconstruction et de la somme ne varie pas.

Temps (%)	Variation ancrage	Extr.	Rec.	HEVC		Autres
				Vue	$\mathrm{II_R}$	
Encodage	130	7	8	2	79	4
Décodage	240	.	31	1	46	22

Tableau 4 – *Fountain - Variation du temps de codage par rapport à l'ancrage, avec* $\mathrm{QPV_{fixed}}$*, et pourcentage du temps total pour chaque tâche, incluant : extraction, reconstruction, codage/décodage de la vue et de l'image résiduelle, et filtrage, différence et somme (autres).*

4 Conclusions

Un schéma efficace de compression d'images intégrales est proposé dans cet article. Une image intégrale résiduelle et une vue extraite y sont encodées. L'image résiduelle est issue de la différence entre l'image originale et une image reconstruite à partir de la vue. Un gain en BD-rate de 15.7% en moyenne, et jusqu'à 31.3%, est obtenu par rapport à la référence HEVC. La configuration du QP utilisé pour encoder la vue, et de la taille du filtre passe-bas appliqué à cette vue, ont un large impact sur l'efficacité du codage. Une méthode robuste d'optimisation débit-distorsion est d'abord modélisée pour sélectionner la meilleure configuration en préservant les gains optimaux. On limite ensuite le nombre d'itérations pour réduire le temps d'encodage, tout en conservant les gains. On montre finalement qu'il est possible d'assigner une valeur unique de QP pour la vue à une valeur de QP donnée pour l'image résiduelle, avec une perte d'efficacité minime, et que la taille du filtre passe-bas peut être sélectionnée avec un nombre réduit d'itérations. De cette étude résulte un codec réaliste du point de vue du compromis entre performances de codage et complexité. Les perspectives pour les futurs travaux incluent des tests du schéma avec plusieurs vues extraites, différents filtres pour la (les) vue(s), et l'utilisation de méthodes d'extractions plus avancées.

Références

[1] F. Dufaux, B. Pesquet-Popescu, et M. Cagnazzo. *Emerging technologies for 3D video : content creation, coding, transmission and rendering*. Wiley Eds, 2013.

[2] G. Lippmann. Epreuves reversibles donnant la sensation du relief. *J. Phys. Theor. Appl.*, 7(1) :821–825, 1908.

[3] A. Lumsdaine et T. Georgiev. Full resolution lightfield rendering. *Indiana University and Adobe Systems, Tech. Rep*, 2008.

[4] https ://www.lytro.com/.

[5] https ://raytrix.de/.

[6] M. P. Tehrani, T. Senoh, M. Okui, K. Yamamoto, N. Inoue, et T. Fujii. [m31261][FTV AHG] Multiple

aspects. Dans *ISO/IEC JTC1/SC29/WG11*, October 2013.

[7] M. P. Tehrani, T. Senoh, M. Okui, K. Yamamoto, N. Inoue, et T. Fujii. [m31103][FTV AHG] Introduction of super multiview video systems for requirement discussion. Dans *ISO/IEC JTC1/SC29/WG11*, October 2013.

[8] http ://www.tgeorgiev.net/.

[9] J. F. O. Lino. 2D image rendering for 3D holoscopic content using disparity-assisted patch blending. *Thesis to obtain the Master of Science Degree*, October 2013.

[10] T. Georgiev et A. Lumsdaine. Focused plenoptic camera and rendering. *Journal of Electronic Imaging*, 19(2) :021106, 2010.

[11] M. Forman, A. Aggoun, et M. McCormick. Compression of integral 3D TV pictures. Dans *Fifth International Conference on Image Processing and its Applications*, pages 584–588, Edinburgh, UK, July 1995. IET.

[12] M. C. Forman, A. Aggoun, et M. McCormick. A novel coding scheme for full parallax 3D-TV pictures. Dans *ICASSP Proceedings*, volume 4, pages 2945–2947, Nagoya, Japan, August 1997. IEEE.

[13] E. Elharar, A. Stern, O. Hadar, et B. Javidi. A hybrid compression method for integral images using discrete wavelet transform and discrete cosine transform. *Journal of display technology*, 3(3) :321–325, September 2007.

[14] D. Marpe, T. Wiegand, et G. J. Sullivan. The H. 264/MPEG4 advanced video coding standard and its applications. *Communications Magazine*, 44(8) :134–143, 2006.

[15] G. J. Sullivan, J. Ohm, W.-J. Han, et T. Wiegand. Overview of the high efficiency video coding (HEVC) standard. *TCSVT*, 22(12) :1649–1668, 2012.

[16] J. Dick, H. Almeida, L. D. Soares, et P. Nunes. 3D holoscopic video coding using MVC. Dans *EUROCON*, pages 1–4, Lisbon, Portugal, April 2011. IEEE.

[17] S. Shi, P. Gioia, et G. Madec. Efficient compression method for integral images using multi-view video coding. Dans *18th ICIP*, pages 137–140, Brussels, Belgium, September 2011. IEEE.

[18] J.-R. Ohm. Overview of 3D video coding standardization. Dans *3DSA*, Osaka, Japan, June 2013.

[19] C. Conti, P. Nunes, et L. D. Soares. New HEVC prediction modes for 3D holoscopic video coding. Dans *19th ICIP*, pages 1325–1328, Orlando, FL, September 2012. IEEE.

[20] D.-K. Kwon et M. Budagavi. Fast intra block copy (intrabc) search for hevc screen content coding. Dans *ISCAS*, pages 9–12. IEEE, 2014.

[21] C. Conti, P. Nunes, et L. D. Soares. Inter-layer prediction scheme for scalable 3-d holoscopic video coding. *Signal Processing Letters*, 20(8) :819–822, 2013.

[22] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, et C. Intwala. Spatio-angular resolution tradeoffs in integral photography. *Rendering Techniques*, 2006 :263–272, 2006.

[23] F. Bossen. Test conditions and software reference configurations. *JCT-VC L1100*, January 2013.

[24] G. Bjøntegaard. Calculation of average PSNR differences between RD-curves. Dans *VCEG Meeting*, Austin, USA, April 2001.

Segmentation de maillages 3D de pièces manufacturées numérisées : Application à la rétro-conception

S. Gauthier[1,2] R. Bénière[2] W. Puech[1] G. Subsol[1]

[1] Equipe ICAR, LIRMM, Université de Montpellier / CNRS, 860 rue de St Priest, 34 095 Montpellier Cedex 5, France
[2] C4W, 219 rue le Titien 34 000 Montpellier, France
E-mail : silvere.gauthier@lirmm.fr.

Résumé

De nos jours, il est de plus en plus fréquent et facile de numériser en 3D la surface des objets réels. Les maillages issus de ces objets numérisés sont souvent imprécis et bruités. De plus, nombre d'entre eux sont de haute résolution, donc très lourds. Dans une chaine de rétro-conception, bien souvent une étape de segmentation est nécessaire. Dans cet article, nous proposons une nouvelle approche de segmentation pour la rétro-conception utilisant les spécificités des maillages numérisés.

Mots clefs

Maillage 3D, Segmentation, Rétro-conception, Numérisation surfacique.

1 Introduction

La rétro-conception regroupe tous les procédés étudiant un objet afin d'en déterminer le fonctionnement interne ou la méthode de fabrication. Cela permet, par exemple, de pouvoir reconstituer numériquement un objet 3D composé d'un ensemble de primitives géométriques telles que des plans, des sphères ou des cylindres. De nombreuses méthodes de rétro-conception ont été proposées ces dernières années (voir par exemple [1], ou plus récemment [2]).
La plupart d'entre elles sont basées sur les paramètres de la courbure du maillage. Or, le calcul de ces paramètres sur un maillage numérisé discret peut être faussé par le bruit ou l'imprécision sur les coordonnées des sommets.

Ainsi, sur le maillage de la figure 1, nous avons calculé un paramètre de courbure, le shape index, par la méthode décrite dans [2] avec différentes valeurs de *k-voisinage*. Le *k-voisinage* d'un point P contient tous les sommets accessibles depuis P en parcourant au plus k arêtes.
Les points des parties convexes du maillage ont un shape-index entre 0 et 1 et sont représentés en bleu, les points des parties concaves entre -1 et 0 apparaissent en rouge et pour les points des parties localement planes, le shape-index n'est pas défini mais ils sont colorés en vert.
Avec un *k-voisinage* restreint (figure 1.a), bien que ce maillage soit issu d'un objet numérisé contenant des plans, il n'y a aucun sommet détecté comme plan. Tandis qu'avec un *k-voisinage* étendu (figures 1.b), plusieurs points sont détectés plans mais uniquement au centre des larges parties planes. En augmentant encore le *k-voisinage* (figure 1.c), les plans étroits deviennent difficilement détectables.

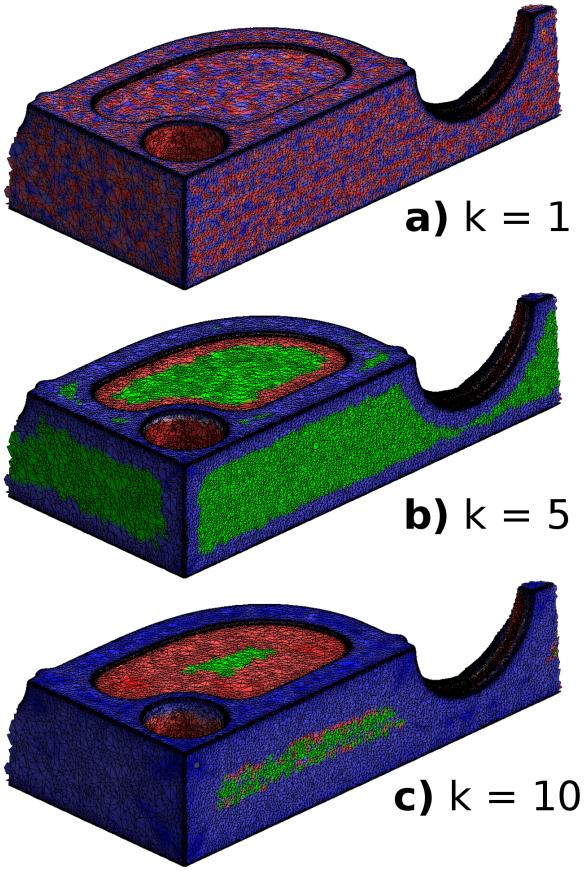


Figure 1 – *Représentation des valeurs du shape-index d'un maillage 3D en fonction du k-voisinage (vert : plan, bleu : convexe, rouge : concave).*

Nous voyons donc que l'utilisation directe du calcul des paramètres de courbure pour extraire des zones homogènes [3] ou des primitives [2] à partir de maillages 3D bruités n'est pas satisfaisante. Pour résoudre ce problème, nous

pouvons appliquer une segmentation avant de calculer la courbure, afin de limiter la portée du voisinage d'un sommet à d'autres sommets appartenant à la même zone homogène. Il devient alors possible d'étendre le *k-voisinage* afin de pallier au bruit, tout en évitant de déborder sur une zone voisine.
Pour valider cette approche, nous avons réalisé une segmentation manuelle de l'objet de la figure 1, en définissant les contours de sous-maillages "uniformes" (dans cet exemple, il s'agit de définir les arêtes). Les paramètres de courbure sont ensuite recalculés séparément pour chaque sous-maillage, et nous obtenons le résultat présenté dans la figure 2. Le calcul de courbure sur les sommets n'est plus perturbé par les zones voisines, et nous obtenons ainsi des paramètres de courbure plus cohérents.

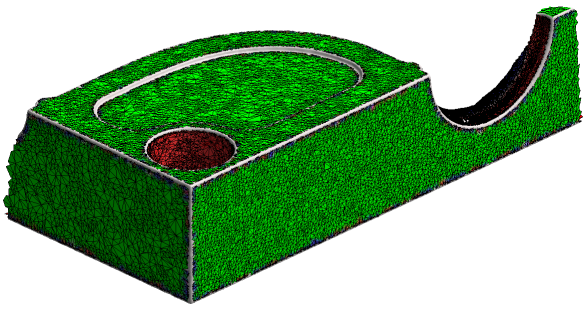


Figure 2 – *Représentation des valeurs du shape-index avec un 5-voisinage après segmentation (vert : plan, rouge : concave).*

Cette étape de segmentation va permettre d'améliorer la qualité du processus global de reconstruction et aussi de diminuer le temps de traitement.

La segmentation de maillages est un problème très courant en modélisation 3D (voir par exemple les revues de l'état de l'art [4] et [5]). Les méthodes peuvent être classées en trois catégories, à savoir les méthodes de croissance de régions [3][6], les méthodes utilisant des lignes caractéristiques [7], et enfin les méthodes de segmentation par patchs surfaciques [8].

Pour autant, les méthodes classiques ne permettent pas d'isoler les régions définissant les arêtes de l'objet, où le calcul de la courbure reste très instable car les valeurs sont souvent très élevées. De plus, détecter ces régions arêtes permettrait de les traiter différemment dans le processus de rétro-conception puisqu'elle vont définir de fait les relations de voisinage entre les primitives géométriques. Dans la suite, nous proposons donc une nouvelle approche de segmentation de maillages numérisés permettant de distinguer les arêtes du reste du maillage.

2 Présentation de la méthode

Cette segmentation se décompose en quatre parties, à savoir une détection des arêtes vives, une segmentation en régions basée sur les arêtes détectées, une fusion des petites régions et une fusion des régions non pertinentes. Un schéma récapitulatif est illustré figure 3, où les parties noires sont celles qui ne sont pas prises en compte.

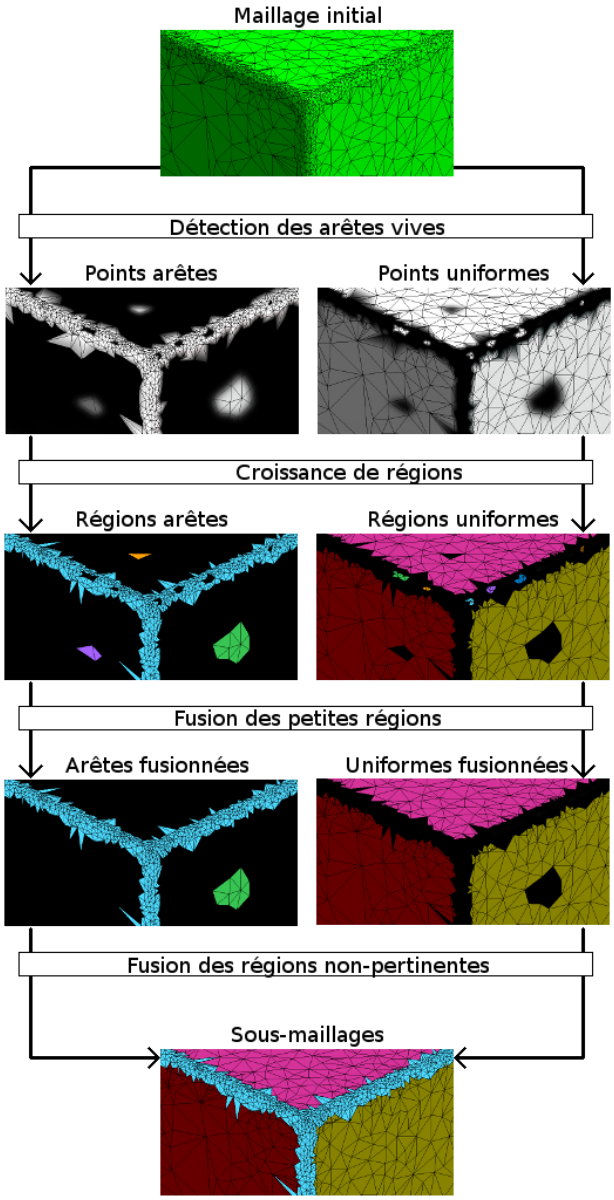


Figure 3 – *Décomposition de la méthode proposée.*

2.1 Détection des arêtes vives

L'enjeu de cette étape est de détecter les arêtes vives de l'objet. Pour cela, des valeurs de courbure moyenne (moyenne des courbures principales $k1$ et $k2$) sont calculées par la méthode de [2] sur les sommets, avec un voisinage de 2.

Les sommets sont ensuite catégorisés selon leur courbure, en tant que point arête ou uniforme, tel que :

$$|Uniforme| < Seuil < |Arete|. \quad (1)$$

Nous proposons d'utiliser un seuil basé sur les dimensions de l'objet :

$$Seuil = \sqrt[3]{\frac{NP}{VBE}}, \quad (2)$$

avec NP le nombre de points du maillage et VBE le volume de la boite englobante de l'objet (exprimé en m^3). Ces informations étant liées à la notion de courbure, elles doivent alors être prises en compte dans le seuillage (le seuil est donc exprimé en m^{-1}, comme l'est la courbure).

Cette formule a été obtenue de manière empirique à partir de nombreux maillages de pièces manufacturées numérisées. Ainsi, nous obtenons une détection efficace des arêtes, comme illustré figure 4, où les points arêtes sont en noir.

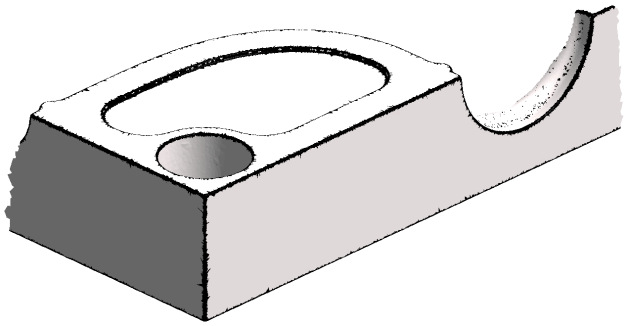


Figure 4 – *Détection des points arêtes vives (illustrés en noir).*

Nous avons constaté qu'un *k-voisinage* de 1 est sensible au bruit. Nous avons donc choisi un *k-voisinage* de 2 car au delà, les résultats perdent en précision (voir figure 1).
De la même manière, nous avons choisi de seuiller la courbure moyenne qui fournit de meilleurs résultats que l'utilisation des courbures principales.

2.2 Croissance de régions

Nous avons obtenu, grâce à l'étape précédente, un ensemble de points arêtes ou uniformes. Nous proposons ici de regrouper ces points afin de former, cette fois-ci, des régions arêtes ou uniformes. La segmentation va s'effectuer par une croissance de régions dont la connexité est définie par un voisinage "Triangle-Triangle".
La méthode initialise d'abord tous les triangles du maillage à une valeur signifiant qu'ils ne sont affectés à aucune région pour l'instant. Ensuite, un appel à la fonction de croissance, détaillée dans l'algorithme 1, est effectué sur chaque triangle de région indéfinie et ses voisins, jusqu'à ce que tous les triangles aient été traités.

Algorithme 1 Fonction de croissance

ENTRÉES: Le triangle de départ T
SORTIES: La région "arête" ou "uniforme" contenant T

```
Region ← List()
Region.PushBack(T)
i ← 1
tant que i < Region.Size() faire
  Ti ← Region[i]
  Marquer Ti comme traité
  pour tout triangle Tv voisin de Ti faire
    si Tv non-traité ET ((Ti et Tv arêtes) OU (Ti et Tv uniformes)) alors
      Region.PushBack(Tv)
    fin si
  fin pour
  i ← i + 1
fin tantque
renvoyer Region
```

Après cette croissance de régions, nous obtenons un ensemble de sous-maillages de type uniforme ou arête, comme illustré figure 5, où chaque couleur correspond à un sous-maillage différent. Nous remarquons alors que le bruit des maillages numérisés a pour effet de créer de nombreuses petites zones, ayant très peu de triangles, comme illustré sur le zoom de la figure 5. Nous proposons alors une étape permettant de fusionner ces régions avec leur plus grande région voisine.

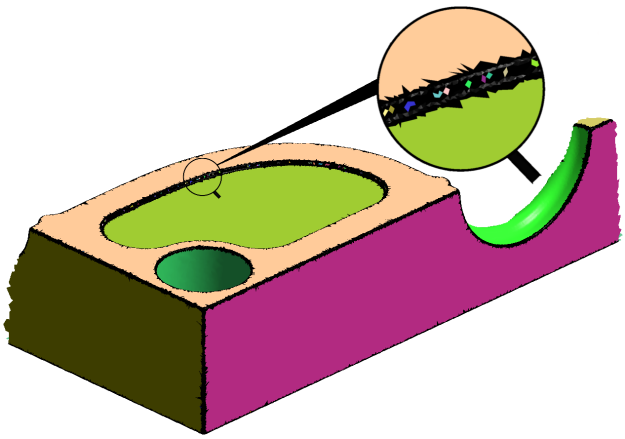


Figure 5 – *Segmentation par croissance de régions après la détection des arêtes vives.*

2.3 Fusion des petites régions

La fusion des régions utilise les prérequis de la segmentation, à savoir que le maillage ne peut comporter que deux types de région : "arête" ou "uniforme". Ainsi, nous savons qu'une région ne peut pas avoir une région de même type parmi ses voisines, conséquence de la croissance de régions. Néanmoins, il peut arriver que certaines régions

aient plusieurs voisines différentes, toutes du type opposé. Chaque petite région sera ainsi fusionnée avec la plus grande de ses régions voisines. Nous réitérons cette étape tant qu'il reste des régions à fusionner.
Pour une meilleure stabilité, nous fusionnons ces régions dans l'ordre croissant de leur aire afin de minimiser l'impact de l'ordre des régions.

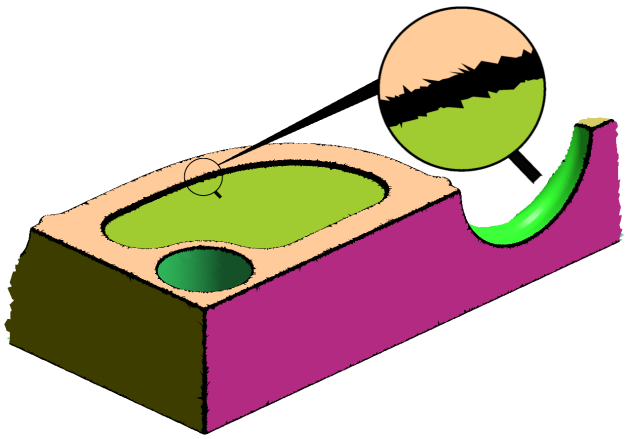


Figure 6 – *Fusion des petites régions à partir de l'exemple illustré figure 5.*

Après cette fusion de régions, nous obtenons un ensemble de sous-maillages, comme illustré figure 6, où chaque couleur correspond à un sous-maillage différent. Contrairement à la figure 5, nous remarquons que le maillage de la figure 6 ne contient plus de petite région.

2.4 Fusion des régions non pertinentes

Après la fusion des petites régions, il peut arriver que deux sous-maillages du même type soient voisins, ou qu'une région arête n'ait qu'un seul voisin alors que par définition, elle doit séparer deux régions uniformes. Nous devons donc fusionner ces régions non pertinentes.

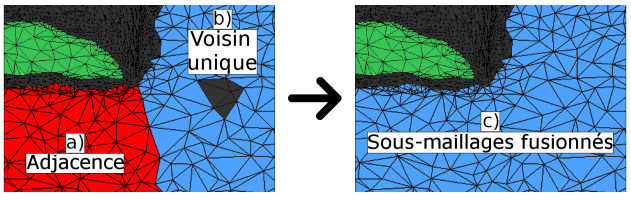


Figure 7 – *Fusion des régions non pertinentes.*

Dans le cas d'adjacence entre deux sous-maillages uniformes, illustré figure 7.a, nous fusionnons les régions en une seule. Dans le cas d'une région arête ayant un unique sous-maillage uniforme voisin, comme illustré figure 7.b, la région arête isolée est fusionnée avec le sous-maillage uniforme. Le résultat final est illustré figure 7.c.

3 Résultats expérimentaux

Dans cette section, des résultats obtenus avec notre méthode de segmentation sont étudiés. Les résultats présentés sont ceux obtenus sur trois maillages : *Shoe*, *Block* et *Mold*, illustrés respectivement figures 8.a, 8.b et 8.c.

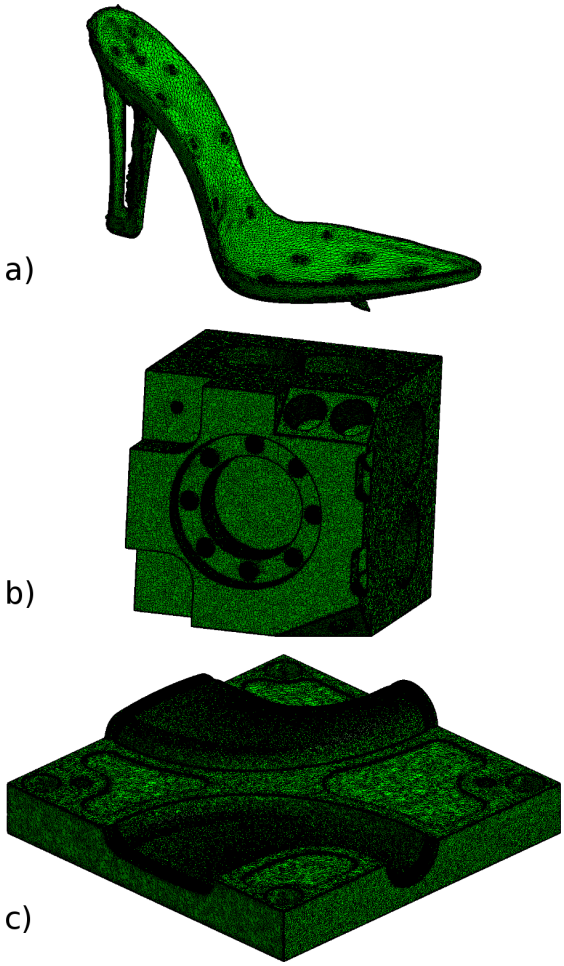


Figure 8 – *Exemple de maillages numérisés : a) Shoe - 74 536 triangles, b) Block - 1 125 832 triangles, c) Mold - 851 194 triangles.*

Ces trois maillages, issus directement de numérisation 3D, sont très différents par leur taille, leur forme et leur répartition des points. Ils sont respectivement composés de 74 536 triangles (37 252 sommets), 1 125 832 triangles (562 914 sommets) et 851 194 triangles (425 589 sommets).

Sachant que cette méthode est intégrée dans une application industrielle, il est à noter qu'aucun paramètre n'est demandé à l'utilisateur. Chaque paramètre utilisé lors de la segmentation est soit fixé, soit calculé, ce qui permet une grande simplification de son utilisation. Le but est ici de présenter une méthode pouvant être appréhendée par des utilisateurs non-experts.

3.1 Résultats de la segmentation

La première étape de notre segmentation consiste à détecter les arêtes vives du maillage. Les résultats de cette détection sur les trois maillages tests sont illustrés figure 9.

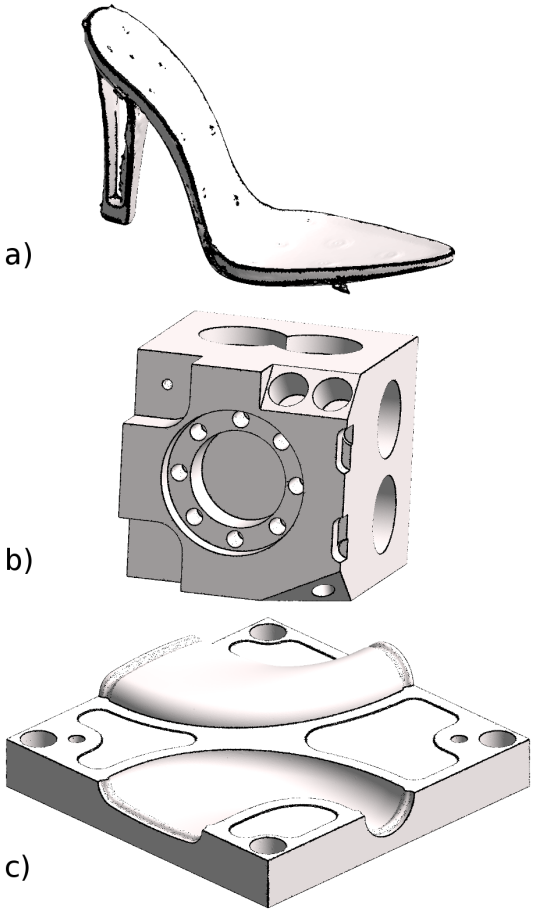


Figure 9 – *Points arêtes détectés (illustrés en noir).*

Ensuite, toutes les étapes de la segmentation ont été réalisées. Les résultats de cette segmentation sont illustrés figure 10 et détaillés dans le tableau 1. Chaque sous-maillage obtenu est colorié d'une couleur différente, et la partie noire restante du maillage correspond aux régions arêtes.

Maillage	a)	b)	c)
Régions avant fusion	207	224	592
Régions après fusion	12	71	32

Tableau 1 – *Résultats de la segmentation avec fusion*

Nous avons appliqué notre méthode sur 20 maillages très différents par la forme, la taille et la répartition des points, incluant ces trois exemples. Globalement, la segmentation obtenue correspond à une segmentation qu'un expert aurait pu faire manuellement.

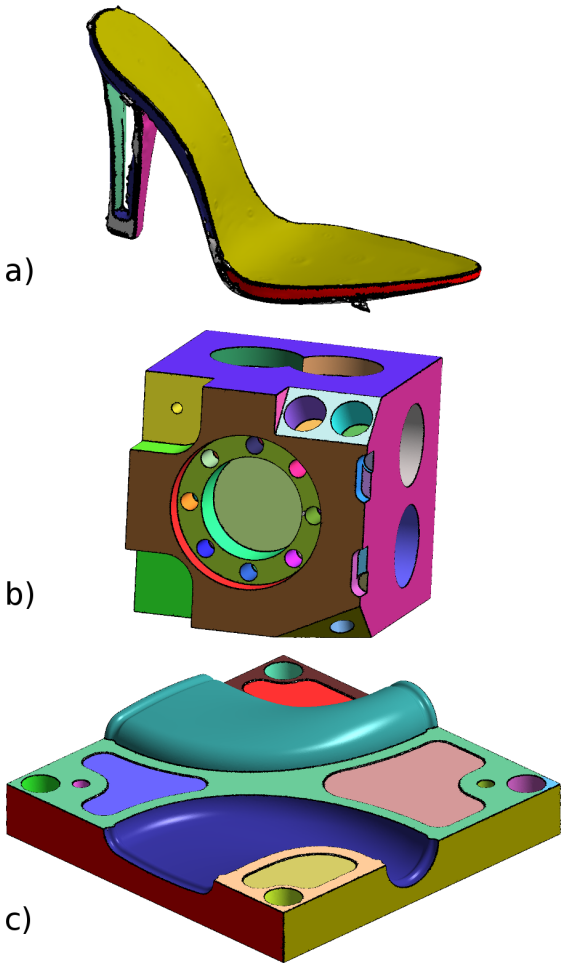


Figure 10 – *Segmentation des maillages tests.*

3.2 Application à la rétro-conception

Notre segmentation a été intégrée dans la solution de rétro-conception proposée par la société C4W.
Les résultats de l'extraction de primitives sur le maillage *Mold*, sans et avec la méthode de segmentation, sont illustrés figures 11.a et 11.b.

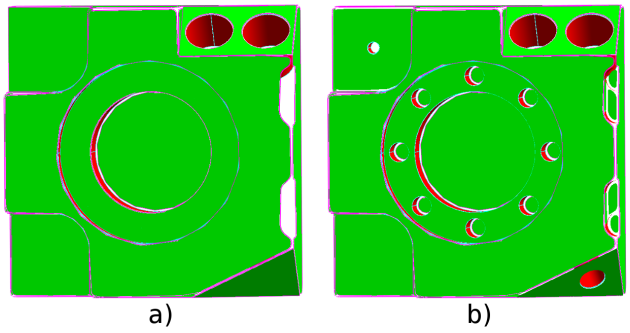


Figure 11 – *Extraction de primitives, a) sans et b) avec segmentation préalable.*

Nous avons alors pu vérifier expérimentalement notre hypothèse de départ, à savoir qu'une étape de segmentation permettait d'améliorer significativement l'extraction de primitives.
En effet, comme présenté dans la section 1, les courbures sont calculées en utilisant un voisinage des points, ce qui peut engendrer des effets de bord. La segmentation permet d'éviter ces effets de bord en limitant la portée du voisinage. Ainsi, les courbures utilisées pour l'extraction sont mieux calculées, d'où la différence notable des résultats.
De plus, nous avons observé un gain de temps non négligeable, se situant entre 40% et 75% sur la totalité de la solution, puisque la recherche des primitives se fait sur des régions plus cohérentes.
Des tests de performances ont été effectués, avec un processeur Intel® Core™ i5-4570 CPU @ 3.20GHz 3.20GHz.

Maillage	Triangles	Temps
a) Shoe	74 536	1s
b) Block	1 125 832	13s
c) Mold	851 194	17s

Tableau 2 – *Temps de segmentation*

Les résultats indiqués dans le tableau 2 montrent que la segmentation proposée s'exécute en quelques secondes. La méthode est donc adaptée à l'utilisation industrielle.

4 Conclusion et perspectives

Dans cet article, nous avons présenté une nouvelle méthode de segmentation des maillages numérisés. Aucun paramètre n'est demandé à l'utilisateur, ce qui est un avantage majeur dans le domaine industriel.
La méthode proposée se décompose en quatre parties, une détection des arêtes vives, une croissance de régions, ainsi que deux étapes de fusion de régions. Nous avons montré au travers de plusieurs résultats que notre méthode retournait des sous-maillages cohérents avec les maillages d'origine. De plus, après l'avoir intégrée dans le processus de rétro-conception de C4W, la segmentation s'est montrée très intéressante pour améliorer les résultats, que ce soit en terme de qualité ou de rapidité.

Un des plus gros avantages de notre méthode est de permettre l'extraction des arêtes du maillage. Ainsi, les sous-maillages sont plus homogènes, les temps de calculs sont plus courts et les calculs utilisés par la suite sont plus précis. De plus, la détection et l'extraction de ces arêtes apportent une information supplémentaire et nécessaire pour la reconstruction, à savoir la notion d'adjacence entre les sous-maillages et donc entre les primitives.

Notre méthode peut être améliorée, notamment au niveau du calcul des seuils. En effet, comme le propose [9], nous pouvons nous appuyer sur l'analyse de la distribution des valeurs de courbure sur un histogramme, illustré figure 12. Nous pouvons ainsi définir une meilleure valeur de seuil et donc améliorer nos résultats.

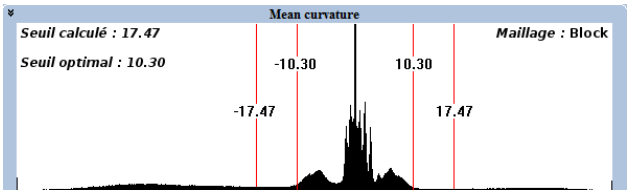


Figure 12 – *Distribution de la courbure moyenne de Block.*

Nous constatons que notre seuil est inférieur au seuil " optimal ", sélectionné manuellement par des experts. Une amélioration de notre segmentation utilisant cette distribution est en cours de développement.

Références

[1] Pál Benkő, Ralph R. Martin, et Tamás Várady. Algorithms for reverse engineering boundary representation models. *Computer-Aided Design*, 33(11) :839 – 851, 2001.

[2] R. Bénière, G. Subsol, G. Gesquière, F. Le Breton, et W. Puech. A comprehensive process of reverse engineering from 3d meshes to cad models. *Computer-Aided Design*, 45(11) :1382 – 1393, 2013.

[3] G. Lavoué, F. Dupont, et A. Baskurt. A new CAD mesh segmentation method, based on curvature tensor analysis. *Computer-Aided Design*, 37 :975–987, 2004.

[4] Ariel Shamir. A survey on mesh segmentation techniques. *Computer Graphics Forum*, 27(6) :1539–1556, 2008.

[5] S. Petitjean. A Survey of Methods for Recovering Quadrics in Triangle Meshes. *ACM Computing Surveys*, 2(34) :1–61, july 2002.

[6] L. Di Angelo et P. Di Stefano. Geometric segmentation of 3D scanned surfaces. *Computer-Aided Design*, 62 :44–56, 2015.

[7] J. Digne, J.-M. Morel, N. Audfray, et C. Mehdi-Souzani. The level set tree on meshes. Dans *Proceedings of the Fifth International Symposium on. 3D Data Processing, Visualization and Transmission*, Paris, France, May 2010.

[8] S. Delest, R. Boné, et H. Cardot. Etat de l'art de la segmentation de maillage 3D par patchs surfaciques. *GTMG '07 : Groupe de Travail en Modélisation Géométrique, Valenciennes : France*, mars 2007.

[9] Jack Szu-Shen Chen et Hsi-Yung Feng. Automatic prismatic feature segmentation of scanning-derived meshes utilising mean curvature histograms. *Virtual and Physical Prototyping*, 9(1) :45–61, 2014.

Etude de la dynamique du visage en situation d'interaction naturelle

Benjamin ALLAERT[1] José MENNESSON[1] Ioan Marius BILASCO[1] Chabane DJERABA[1]

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

{benjamin.allaert}@ed.univ-lille1.fr
{jose.mennesson, marius.bilasco, chabane.djeraba}@univ-lille1.fr

Résumé

L'importance de la dynamique faciale pour la reconnaissance d'expressions spontanées a été prouvée dans l'identification subtile des déformations physiques du visage. Les approches courantes de reconnaissance d'expressions faciales sont performantes sur des datasets où l'environnement est contrôlé et les expressions sont actées. Cependant, ces datasets ne reflètent pas les conditions rencontrées dans une situation d'interaction naturelle où la personne est libre de ses mouvements. Dans ces contextes, la présence de variations de pose et de larges déplacements rend l'analyse difficile. Cet article présente une synthèse originale des solutions de normalisation et d'extraction de mouvements du visage en présence d'expressions spontanées et propose d'étudier l'impact de ces méthodes de normalisation sur la dynamique faciale.

Mots clefs

Expressions faciales, Etude du mouvement, Normalisation

1 Introduction

La reconnaissance d'émotions à partir de flux vidéo se base généralement sur l'analyse des expressions faciales reposant sur le système *Facial Action Coding System (FACS)* de Paul Ekman [1]. Les six expressions faciales universelles (i.e. Joie, Peur, Tristesse, Surprise, Colère, Dégoût) *(6+N)* sont définies par une combinaison linéaire d'unités d'actions caractérisant les mouvements sur le visage et par un terme aléatoire capturant l'incertitude.

La majorité des systèmes proposés mesure leur performance sur des collections d'images où des acteurs reproduisent des expressions faciales face à une caméra. Au vu des dernières études sur la reconnaissance des expressions faciales [2, 3, 4], les méthodes obtiennent de très bonnes performances sur les datatsets où les expressions sont actées. Cependant ces *datasets* ne font pas l'unanimité car en présence d'expressions spontannées, des problèmes de variation de pose et de large déplacement apparaissent. Des *datasets* permettant d'analyser des expressions faciales spontannées mettent en évidence des problèmes de robustesse des précédentes approches et une baisse des performances sur la reconnaissance des expressions spontanées.

Les expressions spontanées sont plus difficiles à caractériser que les expressions actées car les déformations physiques du visage sont moins grandes entre deux images successives. Dans ce contexte, les systèmes s'appuyant sur l'analyse du mouvement obtiennent de meilleures performances [5, 6] car ils permettent de détecter de subtils changements entre deux images, tout en prenant en compte les contraintes physiques du visage.

Plusieurs approches passent par une étape de prétraitement avant d'extraire le mouvement du visage afin de s'abstraire du mouvement global dans la scène. La première catégorie d'approche conserve uniquement les visages frontaux afin de revenir dans un contexte maîtrisé comme dans les *datasets* contraints [7]. Cette méthode est principalement utilisée pour vérifier les performances des systèmes. La seconde catégorie consiste à analyser le visage dans sa position courante, avec une hypothèse forte de symétrie de l'expression faciale. Enfin, la troisième catégorie utilise des algorithmes de recalage pour amener le visage dans une configuration frontale [8, 9].

Les études de synthèse [10, 11] permettent d'avoir une vision globale sur les approches de reconnaissance d'expressions faciales en passant en revue les différentes étapes de traitement, de l'identification à la classification. Ils s'accordent majoritairement sur le fait que l'extraction du mouvement dans une situation d'interaction naturelle reste un problème ouvert dû aux grandes difficultés rencontrées dans un contexte d'interaction naturelle.

Deux défis majeurs pour la reconnaissance d'expression en présence d'interactions naturelles sont étudiés dans cet article : les variations de pose, et la présence de mouvement global de la personne dans le cadre de la scène. Dans la section 2, nous discutons de l'impact des variations de pose et de larges déplacements sur la reconnaissance d'expressions faciales. Pour chacune de ces problématiques, nous discutons des solutions existantes dans la littérature et des déformations du visage induites par les approches de normalisation. Les *datasets* les plus utilisés dans la littérature pour analyser les expressions faciales sont presentés en section 3. Dans la section 4, les approches récentes sont comparées sur les différents *datasets* et le bruit généré par les approches de normalisation du visage pour l'extraction du mouvement facial est étudié. Dans la section 5, les

limites des approches et des *datasets* existants sont discutées, et des pistes pour limiter le déphasage entre les situations d'interactions contrôlées et les situations d'interactions naturelles.

2 Larges Déplacements (LD) et Variations de Poses (VP)

En présence de larges déplacements, une étape d'alignement est souvent nécessaire afin d'enlever le mouvement global du visage et de conserver uniquement le mouvement local. Une solution simple consiste à appliquer une transformation entre les deux images pour aligner les visages [12, 13]. Ces solutions ont tendance à déformer le visage afin de le rendre invariant aux translations, rotations dans le plan et aux changements d'échelle. Souvent utilisés pour identifier les expressions faciales, les visages normalisés peuvent produire des changements morphologiques néfastes [14]. Les travaux de Black *et al.* [15] ont montré qu'une normalisation locale sur plusieurs régions est mieux adaptée pour s'abstraire du mouvement global tout en réduisant les déformations induites par la transformation.

En présence de changements de pose de la tête, les visages ne sont pas directement exploitables sur la plupart des systèmes actuels. Un premier traitement consiste à normaliser toutes les images de visage. L'alignement du visage consiste à modifier la position du visage dans l'espace 3D afin de l'amener dans une configuration 2D idéale où l'apparence du visage est préservée tout en minimisant les artefacts et la perte d'information. La perte d'information est dûe aux rotations hors plan du visage, comme illustré dans la figure 1, où la partie gauche du visage disparaît progressivement au fur et à mesure que la tête tourne. La complexité de la tâche d'alignement dépend de la correction à appliquer.

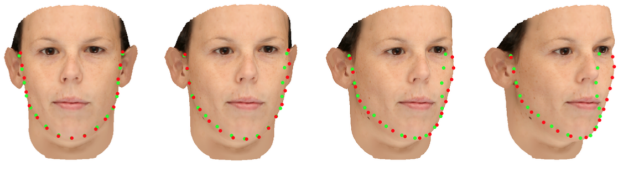


Figure 1 – *Erreur de correspondance des points faciaux du visage 2D (rouge) et 3D (vert) sous plusieurs angles [16].*

Des systèmes d'alignement sont proposés dans la littérature, cependant, la majorité de ces méthodes ne sont pas entièrement satisfaisantes. La majorité des systèmes aligne le visage en appliquant une transformation affine par rapport au centre des yeux et du nez car ces points restent stables indépendamment des expressions faciales [8, 13]. Cette solution souffre des mauvaises détections qui perturbent l'alignement. Des extensions de ces travaux utilisent un nombre plus important de points d'intérêt afin de garantir une meilleure stabilité en cas de mauvaise détection. Certaines approches [3] détectent des points d'intérêt uniquement sur la région centrale du visage. D'autres approches [17] prennent aussi en compte les contours du visage afin d'obtenir des informations complémentaires sur la morphologie du visage. Cependant, en présence de certaines expressions faciales, les contours du visage subissent des déformations impactant la précision de l'alignement.

L'inconvénient majeur des systèmes précédents réside dans les informations disparues et cela met en cause la robustesse des systèmes d'alignement. En effet, un système d'acquisition d'images ne fournit que la projection des scènes observées sur un plan en deux dimensions. L'image ne permet donc d'exploiter que le mouvement résultant de la projection sur le plan 2D, d'où la perte d'information.

Des travaux récents proposent d'utiliser l'information 3D pour aligner le visage en minimisant les pertes. Zhu *et al.* [18] proposent une approche robuste d'alignement du visage en appliquant la texture 2D sur un modèle 3D. La transformation géométrique est appliquée sur le modèle 3D afin d'aligner correctement le visage dans l'espace 2D. Le modèle 3D garantit la conservation de la morphologie du visage permettant d'identifier sa forme globale et ses contours. Comme le montre la Figure 2, l'utilisation du modèle 3D permet de trouver un alignement convenable du visage mais le visage n'est pas entièrement reconstitué. En effet, les parties cachées ne sont pas texturées car il est impossible d'obtenir ces informations avec une seule camera. Pour combler la perte d'information, des méthodes de remplissage de texture [19] sont utilisées afin de reconstruire le visage. Le mouvement est ensuite extrait sur ce visage aligné afin de reconnaître les expressions faciales. Plus ces régions sont importantes et texturées, plus la reconstruction est difficile et la perte d'expression faciale est importante. Ainsi, la prise en compte des zones reconstruites dans le calcul de la dynamique faciale est problématique.

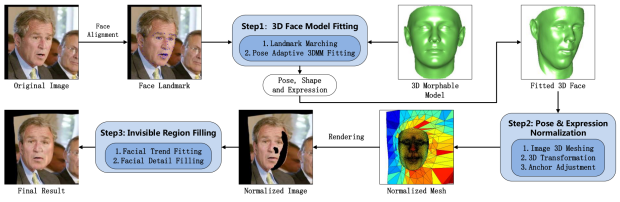


Figure 2 – *Alignement du visage en plaquant la texture sur un modèle 3D, extrait des travaux de Zhu et al. [18].*

Il est important de souligner que les approches d'alignement du visage sont employées principalement dans les systèmes de reconnaissance faciale où la présence de certains artefacts ou la perte de certaines informations est peu importante par rapport aux systèmes de reconnaissance d'expressions faciales. Or, lorsque l'on s'intéresse au mouvement, la présence des déformations liées aux méthodes d'alignement provoque l'apparition de faux mouvements

qui ne peuvent être supprimés sans impacter les mouvements pertinents [14]. Il faut considérer les limitations de l'alignement lorsque l'on extrait le mouvement, sur un visage, dans une situation d'interaction naturelle.

Dans la section suivante, nous analysons comment les critères définis précédement sont reflétés dans les différents corpus utilisés pour l'analyse des expressions faciales.

3 Les datasets

La majorité des systèmes de reconnaissance d'expressions mesurent leur performance dans des situations actées [20, 21], où le visage reste statique, ce qui permet d'analyser la dynamique faciale dans d'excellentes conditions. Dans ces *datasets*, l'expression est exagérée, jouée par des acteurs afin d'obtenir des déformations anormalement amplifiées par rapport au naturel. A l'inverse, il existe des *datasets* où le contexte d'interaction est naturel et les expressions faciales sont spontanées [22, 22, 23].

La caractérisation des émotions permet d'identifier la manière dont les états affectifs sont annotés. Dans la littérature, certaines émotions de base sont universellement reconnues (i.e. Joie, Peur, Tristesse, Surprise, Colère, Dégoût) *(6+N)*. Une alternative à cette représentation est l'utilisation de dimensions émotionnelles : "agréable ou non agréable" (Valence), "réveil ou soumission" (Arousal). La représentation dimensionnelle a un certain nombre d'avantages par rapport à la représentation catégorique. En effet, elle permet d'analyser un large panel d'émotions avec différentes nuances d'intensité qui semble plus adapté à caractériser une situation d'interaction naturelle.

Pour décrire les changements du visage, le système d'action faciale (FACS) proposé par Ekman [1] est largement utilisé. Certains *datasets* [24, 25, 26] contiennent une annotation continue sur l'enclenchement des muscles faciaux. Ces annotations sont utilisées pour caractériser les expressions faciales et permettent de mesurer l'extraction de mouvements sur le visage.

Certains *datasets* mettent en évidence plusieurs problèmes présentés dans la section 2. Il existe des *datasets* actés [20, 21, 27, 24] qui incluent partiellement des problématiques illustrées dans la section 2. Des *datasets* spontannés [28, 29, 22, 23, 26, 25] répondent à ces conditions en proposant d'étudier les problèmes de variations de pose (VP) et de larges déplacements (LD).

Les *datasets* couramment utilisés dans la littérature sont recensés dans le tableau 1. A ce jour, peu de *datasets* permettent d'analyser les émotions dans une situation d'interaction naturelle où les expressions sont spontanées. Certains *datasets* actés [20, 21] contiennent des extensions permettant d'étudier des expressions spontanées, cependant la majorité des travaux exploitent uniquement les données actés sur ces datasets. En présence d"interactions spontanées, il est difficile d'obtenir une annotation précise sur ces *datasets* car la durée des séquences est souvent longue et que les conditions de capture rendent l'annotation manuelle difficile.

Dataset	Type	Caract.	Longueur Seq.	AUs	VP	LD
CK+ [20]	Acté	6+N	10-60 img/seq	-	+	-
MMI [21]	Acté	6+N	40-520 img/seq	-	+	-
JAFFE [27]	Acté	6+N	2-3 img/seq	-	-	-
GEMEP [24]	Acté	6+N	100 img/seq	✓	+	+
SEMAINE [22]	Spont.	A/V	2K-25K img/seq	partielle	++	++
DISFA [25]	Spont.	-	4844 img/seq	✓	-	-

Tableau 1 – *Recensement des datasets couramment utilisés pour la reconnaissance des expressions faciales.*

Dans la section suivante, nous discutons de la manière dont les méthodes de l'état de l'art répondent aux défis posés par ces *datasets*.

4 Reconnaissance des émotions

Dans la littérature, nous distinguons deux catégories d'approche pour analyser les déformations du visage : les approches basées sur l'étude des caractéristiques du visage et les approches basées sur l'étude du mouvement. L'extraction du mouvement consiste à analyser les déformations apparentes du visage, causées par le mouvement relatif entre la caméra et la scène. Ambadar *et al.* [5] montrent l'importance de l'analyse de la dynamique faciale pour reconnaître les expressions car cela permet d'identifier plus subtilement les déformations physiques du visage.

Les récentes approches de la littérature qui extraient les expressions faciales pour analyser les états affectifs sont référencées dans le tableau 2. Ces approches ont été appliquées sur les différents *datasets* du tableau 1. Au vu des performances sur les *datasets* CK+, MMI, JAFFE, on constate que l'extraction des expressions faciales, où l'environnement est contrôlé et les expressions sont actées, est maîtrisée. En revanche, ce n'est pas le cas sur les *datasets* SEMAINE et DIFSA.

Des méthodes de normalisation sont appliquées dans la majorité des approches quel que soit le *dataset* utilisé. En effet, la normalisation est importante dans une situation d'interaction naturelle afin de s'assurer de la correspondance des régions. Cependant l'impact de la normalisation n'aura pas les mêmes conséquences sur les visages contraints car les problèmes de variation de pose et de larges déplacements sont très faibles. C'est pourquoi, une normalisation sur deux/trois points, comme le centre des yeux et le nez, est souvent employé dans un contexte contraint [8, 9, 4, 12, 30]. En revanche, dans une situation d'interaction naturelle, ces méthodes ne sont plus adaptées et nécessitent des techniques plus complexes [31, 32, 33].

La reconnaissance d'émotions via l'analyse des expressions faciales peut être réalisée en mode interactif ou en définissant une séquence type consistant en un ordre de mouvements faciaux. Dans un mode interactif, chaque image est analysée puis associée à une classe puis un taux de reconnaissance (ar) est calculé en sortie. La décomposition en séquence consiste à étudier l'activation et la durée des expressions faciales pour conserver l'information temporelle. La durée de ces séquences permet de distin-

guer des nuances d'intensité dans les expressions et varie en fonction des personnes. Les performances de ces approches sont calculées avec des méthodes de corrélation croisée (cc).

L'extraction de caractéristiques visuelles est utilisée par la majorité des approches. La prise en compte de l'information temporelle dans plusieurs travaux [8, 2, 34] apporte une hausse des performances sur la reconnaissance des émotions. Ceci met en évidence l'importance de l'évolution du contexte pour reconnaître un état émotionnel. Les approches basées sur les caractéristiques visuelles et/ou sur le mouvement sont appliquées localement ou globalement afin d'analyser les expressions faciales. Sanchez *et al.* [35] comparent deux approches de mouvement (dense, 15 points d'intérêt) au cours d'une même expérience sur le *dataset* MMI [21]. Ils constatent une amélioration de 3% avec l'approche dense. Les approches d'extraction des mouvements denses [12, 36, 4, 37] montrent de bonnes performances sur l'analyse des epressions actées. Comme souligné dans la section 2, la normalisation est un processus difficilement maîtrisé dans une situation d'interaction naturelle, ainsi, il est difficile de récupérer tous les avantages des approches denses dans ces conditions. Dans la suite ce cette section nous nous intéressons aux approches denses et aux solutions pour adapter ces approches aux situations d'interactions naturelles.

Reference	Methode	Norm.	Type	Performances
Zhao *et al.* [8]	LBP-TOP	2pts	6+N	CK+ ar :95.2%
Sikka *et al.* [9]	Dense BoW	2pts	6+N	CK+ ar :95.9%
Happy *et al.* [30]	Salient Patches	2pts	6+N	CK+ ar :94.14%
Su *et al.* [37]	Optical flow	Edges	6+N	CK+ ar :93.89%
Koelstra *et al.* [4]	FFDs	2pts	AUs	MMI cr :94.3%
Jiang *et al.* [2]	LPQ-TOP	4pts	AUs	MMI cr :94.7%
Jiang *et al.* [38]	LPQ	N/A	AUs	GEMEP cr :66%
Yang *et al.* [33]	LBP,LPQ	3D	6+N	GEMEP ar :84%
Liao *et al.* [12]	Optical Flow	2pts	6+N	JAFFE ar :92.5%
Zhang *et al.* [36]	patch based Gabor	N/A	6+N	JAFFE ar :92.93%
Cruz *et al.* [31]	LPQ	3D	A/V	SEMAINE ar :55%
Nicolle *et al.* [32]	App&Geom Feat.	CLM	A/V	SEMAINE cc :0.46
Sandbach *et al.* [13]	LBP	3pts	AUs	DISFA cc :0.342

Tableau 2 – *Les récentes approches de la littérature. (cc : person's cross correlation, ar : average recognition rate, cr : classification rate)*

Contrairement aux approches basées caractéristiques, les approches denses ne nécessitent pas d'étape de segmentation d'image et permettent ainsi d'exploiter toute l'information contenue dans l'image en restant au plus près de la réalité. Parmi ces approches, les méthodes de flux optiques ont été largement utilisées et ont montrées leur efficacité pour analyser les déformations physiques du visage [15, 39]. Cependant, ces approches sont très sensibles à l'environnement et il est difficile de dissocier le mouvement réel du mouvement apparant dans les images. En effet, un système d'acquisition d'images ne fournit que la projection des scènes observées sur un plan en deux dimensions. L'image ne permet donc d'exploiter que le mouvement résultant de la projection sur le plan 2D du mouvement réel 3D, couplé avec le mouvement de la caméra. Liao *et al.* [12] utilisent les flux optiques pour analyser les régions les plus distinctives afin de reconnaitre les expressions faciales. Chaque modèle de mouvement observé est associé à l'une des 6 expressions universelles

L'analyse du mouvement à partir de séquences d'images décrivant la dynamique faciale dans une situation d'interaction naturelle est particulièrement délicat. La difficulté réside dans la spécificité des mouvements associés au visage. Ces mouvements se caractérisent entre autres par des fortes variations de pose et de larges déplacements. Dans ce contexte, les techniques standards issues de la littérature, qui s'appuient sur des caractéristiques stables de la fonction de luminance et sur l'hypothèse d'un mouvement rigide, s'avèrent en général mal adaptées.

La méthode de Farneback [40] permet de calculer le flux optique dense rapidement en utilisant des polynômes, où le mouvement local est diffusé sur l'ensemble de ces voisins pour calculer le mouvement global. Cependant, en présence de discontinuité de mouvement, l'erreur produite localement est propagée sur l'ensemble des voisins avec une intensité moindre, ce qui réduit la qualité du mouvement global. C'est pourquoi, il est important de prendre en compte les variations de pose et de large déplacement lorsqu'on utilise les méthodes de flux optique sur un visage.

La conception de méthodes alternatives dédiées aux mouvements d'objet non rigide constitue un vaste champ d'investigation relativement peu abordé. Les régions occultées de l'image courante sont définies par un ensemble de pixels qui disparaissent dans l'image suivante en présence de large déplacement et de variation de pose comme illustré sur la figure 1. Ces pixels n'ont pas de correspondance dans l'image suivante et donc le mouvement associé n'est pas observable dans ces régions. Pour résoudre ces problèmes, des méthodes sont proposées pour détecter les contours assujetis aux risques d'occultations en s'appuyant sur les flux optiques [41]. La majorité des méthodes calcule le flux optique sur l'ensemble du visage et supprime l'information de mouvement aux alentours de ces contours pour réduire le bruit provoqué par ces discontinuités. Cependant, les fortes variations de pose impliquent une perte importante des informations du visage et de ce fait, réduisent la précision des algorithmes de reconnaissance d'expressions faciales. D'autres méthodes utilisent ces contours pour calculer un flux optique localement afin de réduire ces discontuinités de mouvement. Pour cela, ils utilisent des méthodes de remplissage de mouvement en se basant sur les flux optiques du voisinage et sur les contraintes physiques [42].

La prise en compte de toutes ces contraintes augmente considérablement le temps de calcul. Toutefois, ces solutions permettent d'adapter les méthodes de flux optiques pour analyser subtilement les expressions faciales dans une situations d'interaction naturelle.

5 Discussion

Les récentes approches de reconnaissance d'expressions faciales sont validées sur des *datasets* où l'interaction est naturelle car les expressions observées sont spontanées et reflètent mieux la réalité. Dans ce contexte, l'analyse des expressions est soumise à certaines contraintes liées à la liberté de mouvement du sujet (notamment de sa tête) et au système d'acquisition. De fait que la plupart des approches requièrent des images de visage quasi-frontale, une étape de sélection des visages à traiter ou une étape de normalisation du visage sont nécessaires pour garantir, en entrée du processus d'analyse, uniquement des visages frontaux. Toutefois, il est important d'adapter les méthodes de normalisation en fonction du contexte et des caractéristiques des approches sous-jacentes. La normalization permet d'améliorer les performances des systèmes caractérisant statiquement la texture du visage. En revanche, elle induit des pertes d'information ou des modifications morphologiques du visage qui provoquent des incohérences sur les données extraites lorsque la dynamique de la texture est caractérisée, par exemple, par l'intérmediaire des flux optiques.

La prise en compte de l'information temporelle, de l'étude du mouvement et des méthodes denses nous semblent mieux adaptées pour répondre aux problématiques soulevées par les nouveaux *datasets* qui s'approchent de plus un plus d'un environnement de captation naturel où les sujets s'expriment de manière non-actée et non-posée. Cette observation est également confortée pas le renouveau des approches caractérisant la dynamique du visage par rapport aux nombreuses approches traitant uniquement de l'analyse statiques du visage. Il en est de même pour les méthodes de classifications où l'analyse continue est mieux adaptée pour caractériser un état affectif qu'une analyse discrète.

Références

[1] P. Ekman et E.L. Rosenberg. *What the face reveals : Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[2] B. Jiang, M. Valstar, B. Martinez, et M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *Cybernetics*, 44(2) :161–174, 2014.

[3] M. Valstar et M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *Systems, Man, and Cybernetics, Part B : Cybernetics*, 42(1) :28–43, 2012.

[4] S. Koelstra, M. Pantic, et I.Y. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *PAMI*, 32(11) :1940–1954, 2010.

[5] Z. Ambadar, J.W. Schooler, et J.F. Cohn. Deciphering the enigmatic face the importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, 16(5) :403–410, 2005.

[6] E.G. Krumhuber, A. Kappas, et A.SR. Manstead. Effects of dynamic aspects of facial expressions : a review. *Emotion Review*, 5(1) :41–46, 2013.

[7] T. Ahonen, A. Hadid, et M. Pietikainen. Face description with local binary patterns : Application to face recognition. *PAMI*, 28(12) :2037–2041, 2006.

[8] G. Zhao et M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *PAMI*, 29(6) :915–928, 2007.

[9] K. Sikka, T. Wu, J. Susskind, et M. Bartlett. Exploring bag of words architectures in the facial expression domain. Dans *ECCV*, pages 250–259. Springer, 2012.

[10] E. Sariyanidi, H. Gunes, et A. Cavallaro. Automatic analysis of facial affect : A survey of registration, representation and recognition. 2014.

[11] Z. Zeng, M. Pantic, G. Roisman, T.S. Huang, et al. A survey of affect recognition methods : Audio, visual, and spontaneous expressions. *PAMI*, 31(1) :39–58, 2009.

[12] C. Liao, H. Chuang, C. Duan, et S. Lai. Learning spatial weighting via quadratic programming for facial expression analysis. Dans *CVPRW*, pages 86–93. IEEE, 2010.

[13] G. Sandbach, S. Zafeiriou, et M. Pantic. Markov random field structures for facial action unit intensity estimation. Dans *ICCVW, 2013*, pages 738–745. IEEE, 2013.

[14] S.W. Chew, P. Lucey, S. Lucey, J. Saragih, J.F. Cohn, I. Matthews, et S. Sridharan. In the pursuit of effective affective computing : The relationship between features and registration. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, 42(4) :1006–1016, 2012.

[15] M.J. Black et Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. Dans *Computer Vision*, pages 374–381. IEEE, 1995.

[16] C. Qu12, E. Monari, T. Schuchert, et J. Beyerer21. Adaptive contour fitting for pose-invariant 3d face shape reconstruction. 2015.

[17] A. Dhall, A. Asthana, R. Goecke, et T. Gedeon. Emotion recognition using phog and lpq features. Dans *FG*, pages 878–883. IEEE, 2011.

[18] X. Zhu, Z. Lei, J. Yan, D. Yi, et S.Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. Dans *CVPR*, pages 787–796, 2015.

[19] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, et S. Shan. Morphable displacement field based image matching for face recognition across pose. Dans *ECCV*, pages 102–115. Springer, 2012.

[20] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, et I. Matthews. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. Dans *CVPRW*, pages 94–101. IEEE, 2010.

[21] M. Pantic, M. Valstar, R. Rademaker, et L. Maat. Web-based database for facial expression analysis. Dans *ICME*, pages 5–pp. IEEE, 2005.

[22] G. McKeown, M. Valstar, R. Cowie, M. Pantic, et M. Schröder. The semaine database : Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing*, 3(1) :5–17, 2012.

[23] F. Ringeval, A. Sonderegger, J. Sauer, et D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. Dans *FG*, pages 1–8. IEEE, 2013.

[24] T. Bänziger, M. Mortillaro, et K.R. Scherer. Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5) :1161, 2012.

[25] M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I.R. Fasel, et J.R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6) :22–35, 2006.

[26] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, et J.M. Girard. Bp4d-spontaneous : a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10) :692–706, 2014.

[27] M. Lyons, S. Akamatsu, M. Kamachi, et J. Gyoba. Coding facial expressions with gabor wavelets. Dans *FG*, pages 200–205. IEEE, 1998.

[28] L. Yin, X. Chen, Y. Sun, T. Worm, et M. Reale. A high-resolution 3d dynamic facial expression database. Dans *FG*, pages 1–6, 2008.

[29] I. Sneddon, M. McRorie, G. McKeown, et J. Hanratty. The belfast induced natural emotion database. *Affective Computing*, pages 32–41, 2012.

[30] SL. Happy et A. Routray. Automatic facial expression recognition using features of salient facial patches. *Affective Computing, IEEE Transactions on*, 6(1) :1–12, 2015.

[31] A. Cruz, B. Bhanu, et S. Yang. A psychologically-inspired match-score fusion model for video-based facial expression recognition. Dans *ACII*, pages 341–350. Springer, 2011.

[32] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, et M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. Dans *ICMI*, pages 501–508. ACM, 2012.

[33] S. Yang et B. Bhanu. Facial expression recognition using emotion avatar image. Dans *FG*, pages 866–871. IEEE, 2011.

[34] C. Weng et S. Lai. Online facial expression recognition based on combining texture and geometric information. Dans *ICIP, 2014*, pages 5976–5980. IEEE, 2014.

[35] A. Sánchez, J.V. Ruiz, A. Moreno, A.S. Montemayor, J. Hernández, et J.J. Pantrigo. Differential optical flow applied to automatic facial expression recognition. *Neurocomputing*, 74(8) :1272–1282, 2011.

[36] L. Zhang et D. Tjondronegoro. Facial expression recognition using facial movement features. *Affective Computing*, 2(4) :219–229, 2011.

[37] M. Su, Y. Hsieh, et D. Huang. A simple approach to facial expression recognition. *Proceeding WSEAS 2007*, 2007.

[38] B. Jiang, B. Martinez, M. Valstar, et M. Pantic. Decision level fusion of domain specific regions for facial action recognition. Dans *ICPR*, pages 1776–1781. IEEE, 2014.

[39] J.F. Cohn, A.J. Zlochower, J.J. Lien, et T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. Dans *FG*, pages 396–401. IEEE, 1998.

[40] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. Dans *Image Analysis*, pages 363–370. Springer, 2003.

[41] P. Sundberg, T. Brox, M. Maire, P. Arbeláez, et J. Malik. Occlusion boundary detection and figure/ground assignment from optical flow. Dans *CVPR*, pages 2233–2240. IEEE, 2011.

[42] J. Revaud, P. Weinzaepfel, Z. Harchaoui, et C. Schmid. EpicFlow : Edge-Preserving Interpolation of Correspondences for Optical Flow. Dans *Computer Vision and Pattern Recognition*, 2015.

Dans l'approximation de calcul pour la réduction de la complexité du codage et décodage des données multimédia pour des applications mobiles

Maher Jridi et Ayman Alfalou
ISEN-Brest/Lab@ISEN, Équipe Vision
ISEN-Brest, 20 Rue Cuirassé Bretagne, CS 42807 Brest Cedex 2
maher.jridi@isen-bretagne.fr, ayman.alfalou@isen-bretagne.fr

Résumé

Avec l'avènement du big-data et des réseaux de capteurs d'images, le besoin en compression de données devient de plus en plus crucial. D'un autre côté, les systèmes portables à base de batteries ont des contraintes de consommation de plus en plus sévères. Par conséquent, les algorithmes de compression doivent être à la fois efficaces (en termes de compromis rapport de compression / qualité d'image), mais aussi faible-consommation pour permettre leur embarquement dans des systèmes portables. Dans ce papier nous présentons une approche à base d'approximation de calcul pour la réduction de complexité des algorithmes de compression. Plus concrètement, nous présentons ici une nouvelle méthode d'approximation de la DCT (discrete cosine transform) de taille 4 points et nous l'utilisons pour le calcul de la DCT et IDCT (DCT inverse) pour toute les tailles en puissance de 2. Des architectures reconfigurables et évolutives de la DCT sont présentées. Les résultats de simulation et de synthèse sur FPGA montrent le fort potentiel de la méthode proposée.

Mots clefs

Compression d'images, approximation de calcul, DCT.

1 Introduction

Les demandes en terminaux mobiles connectés et objets connectés (IoT) sont en fortes augmentations. En 2019, certaines études prévoient 1,5 terminal mobile par personne [1]. Les opérations de codage, décodage, téléchargement et affichage du contenu multimédia sont les plus utilisés dans ces systèmes. Par conséquent, une réalisation faible-consommation de ces fonctionnalités est recommandée. La puissance totale consommée par ces fonctions se compose de deux parties. La puissance consommée dans le traitement (codage, décodage, etc.) et la puissance consommée dans la transmission qui dépend de quantité d'information transmise. Les nouveaux standards de compression tel que HEVC [2] permettent de réduire les quantités d'information transmise pour une même qualité d'image. Malheureusement, ceci se fait au détriment de la complexité de calcul. Il devient alors nécessaire de réduire la complexité des traitements pour avoir une diminution de la consommation de puissance.
D'un point de vue algorithmique, la réduction de la complexité de traitement dans les codeurs vidéo peut être réalisée par des approximations de calcul. Les opérations de transformation et de filtrage peuvent tolérer certaines erreurs et donc peuvent être approximées. Le profilage de l'encodeur vidéo [3] montre que sous certaines conditions, le temps passé dans les fonctions de transformation peut atteindre 25% du temps d'encodage global. Ainsi, l'approximation de la DCT et IDCT peut être intéressant pour la réduction de la facture énergétique.

L'objectif principal des approximations de la DCT est de supprimer les multiplications qui consomment du temps et des ressources matérielles. Le besoin d'approximation est plus important pour les DCT de grandes tailles puisque la complexité de la transformation augmente quadratiquement avec la taille. Dans ce travail, nous avons choisi d'appliquer une approximation sur la DCT utilisée dans le standard HEVC. Il s'agit de DCT entière de taille 4-, 8-, 16- et 32-point. Nous avons réalisé l'approximation de la DCT de taille 4-point et nous l'avons utilisé pour approximer les DCT de taille 8-, 16- et 32-point. Ce processus d'utilisation de DCT de petites tailles pour obtenir des DCT de grandes tailles n'est pas possible pour le calcul exact de la DCT, contrairement aux autres transformations comme la DFT. Le processus de décomposition est rendu possible grâce à l'approximation de calcul. Ceci permet d'offrir plusieurs options de reconfigurabilité et d'évolutivité de l'architecture de DCT. De plus, des architectures séries (contrainte de surface) ou parallèles (contrainte temps-réel) peuvent être envisagées.

Dans la section 2 nous donnons les détails nécessaires pour l'approximation de la DCT. Les architectures reconfigurables et évolutives sont données dans la section 3. Les résultats de synthèse et de simulation sont analysés dans les sections 4 et 5 avant la conclusion.

2 Approximation de la DCT

Pour un vecteur d'entrée *X*, le vecteur des coefficients de sortie de la DCT *Y* est donné par $Y = C_N X$, où C_N est le noyau de calcul de la DCT de taille N-point. Il est montré dans [4] qu'il est possible de décomposer ce noyau afin de

reconstruire ses éléments à base de matrices carrées de petites tailles. Dans ce travail, nous faisons le choix de ne pas présenter les détails de cet algorithme pour éviter tout encombrement. La nouveauté de l'étude présentée dans ce papier est que nous reconstruisons toutes les matrices de la DCT à base de la matrice 4-point prise du standard HEVC et présentée par (1) :

$$C_4 = \begin{pmatrix} 64 & 64 & 64 & 64 \\ 83 & 36 & -36 & -83 \\ 64 & -64 & -64 & 64 \\ 36 & -83 & 83 & -36 \end{pmatrix} \quad (1)$$

Le choix cette matrice est justifié par la grande efficacité de la DCT basée sur (1) ce qui explique son adoption par le dernier standard de codage vidéo. Nous proposons l'approximation de la matrice C_4 par :

$$\tilde{C}_4(i,j) = 32 \times \Re\left(\frac{C_4(i,j) + 32 \times \mathrm{sgn}\left(C_4(i,j)\right)}{64}\right) \quad (2)$$

Où $\Re$ et sgn sont les opérateurs d'arrondi et de signe, respectivement. Ainsi, la matrice $\tilde{C}_4$ est donnée par (3) :

$$\tilde{C}_4 = \begin{pmatrix} 64 & 64 & 64 & 64 \\ 64 & 32 & -32 & -64 \\ 64 & -64 & -64 & 64 \\ 32 & -64 & 64 & -32 \end{pmatrix} \quad (3)$$

3 Architectures de DCT proposées

3.1 Approximation de calcul de la DCT

La matrice dans (3) présente l'avantage d'être orthogonalisable et fait une approximation assez efficace de la matrice exacte. De plus, ses éléments sont des entiers en puissance de 2 ce qui permet de réaliser les multiplications par des opérations de décalage. L'architecture de la DCT 4-point est montrée dans la Figure 1. Nous pouvons remarquer que le calcul de la DCT nécessite 8 opérateurs d'additions ainsi que 2 opérateurs de décalages simples. Notons que les décalages au niveau de l'étage de sortie ne sont pas prise en compte puisqu'ils peuvent être inclus dans le processus de quantification.

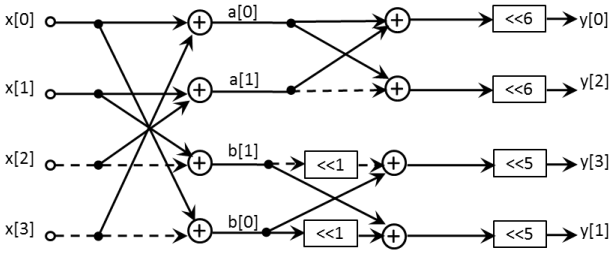


Figure 1 – *Structure de calcul de la DCT 4-point*

Enfin, conformément à l'algorithme de décomposition présenté dans [4], toutes les matrices de DCT de taille sous forme d'une puissance de 2 peuvent être approximées par une recomposition matricielle à base de la matrice donnée dans (3). Par exemple, le calcul de DCT 8-point, montré dans la Figure 2, utilise une unité d'addition à l'entrée, une unité de permutation en sortie qui ne présente pas un coût matériel et deux unités de calcul approximé de la DCT 4-point. Notons aussi que par analogie à la structure de la Figure 2, toutes les tailles de DCT supérieures à 8 présentent cette même structure.

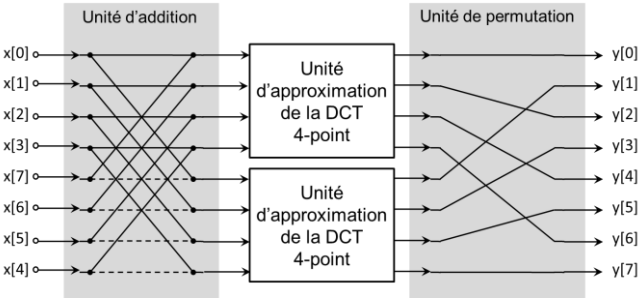


Figure 2 – *Structure de calcul de la DCT 8-point*

3.2 Approximation de calcul de la DCT inverse

Un autre avantage de l'approximation de calcul de la DCT présentée dans ce travail est que le calcul de l'IDCT (DCT inverse) se fait de façon similaire au calcul de la DCT directe. En effet, pour un vecteur d'entrée *U*, le vecteur *V* de sortie de l'IDCT est donné par $V = C_N^T U$. La réalisation de l'IDCT 4-point est présentée dans la Figure 3. Nous pouvons aisément voir des similarités entre les structures présentées dans la Figure 1 et la Figure 3. Il devient alors envisageable de trouver des architectures reconfigurables qui permettent d'implanter les DCT et IDCT sur le même circuit. En effet, pour des applications mobiles, où les circuits doivent supporter à la fois des fonctionnalités comme la capture vidéo et le play-back, la réalisation de la transformation et la transformation inverse sur le même circuit prend tout son intérêt.

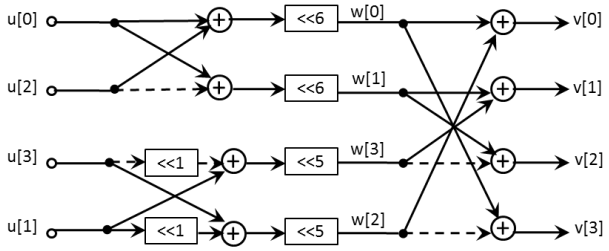


Figure 3 – *Structure de calcul de l'IDCT 4-point*

Pour réaliser cela, il faut voir que la principale différence entre les deux structures des Figure 1 et 3, est que l'unité d'addition est située à l'entrée pour la DCT et à la sortie pour l'IDCT. Une architecture unifiée qui réalise le calcul de la DCT 32-point et de l'IDCT 32-point est alors proposée et présentée dans la Figure 4.

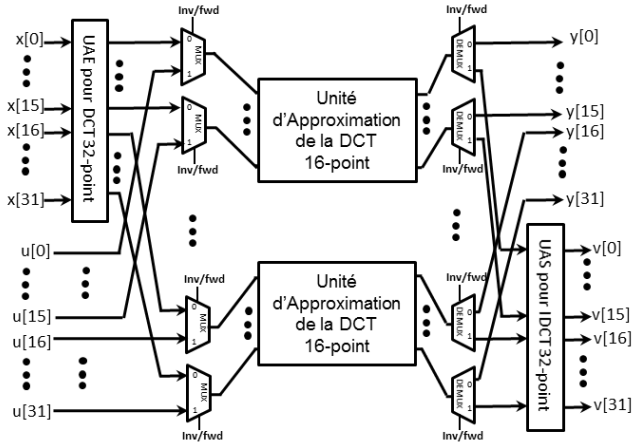


Figure 4 – *Architecture unifiée pour le calcul de la DCT et de l'IDCT 32-point*

Les unités UAE et UAS mentionnent respectivement les unités d'addition à l'entrée et à la sortie. Le choix de calcul de la DCT ou de l'IDCT se fait au moyen de multiplexeurs. Une entrée de sélection Inv/Fwd est située aux niveaux des multiplexeurs pour indiquer le choix de la transformation (directe ou inverse).

4 Résultats d'implantation sur FPGA

L'approximation de la DCT proposée ainsi que celles présentées dans [5-8] ne requièrent pas de multiplications contrairement au calcul exact de la DCT qui consomme 22, 86 et 342 multiplications respectivement pour les DCT de tailles 8-, 16- et 32-point. De plus, les nombres d'additions requises par les algorithmes d'approximation existants [5-8] et par celui présenté dans ce travail sont très comparables, mais aussi largement inférieurs à ceux requis par le calcul exact de la DCT. En effet, l'algorithme proposé consomme 8, 24, 64 et 160 additions alors que l'algorithme de référence (calcul exact) requiert l'utilisation de 8, 28, 100 et 372 additions respectivement pour des DCT de taille 4-, 8-, 16- et 32-point.

Des architectures pipelines de l'algorithme de référence, proposé et existants ont été décrites en utilisant le langage VHDL et synthétisées sur des FPGA de type Spartan 6 LX45T de Xilinx. Les résultats de synthèse pour les algorithmes existants [5-8] et celui présenté dans ce travail sont quasi-identiques. Le tableau 1 présente une comparaison par rapport à l'algorithme de référence. Nous pouvons voir facilement l'intérêt des approximations sur toutes les contraintes d'implantation. En effet, l'algorithme proposé consomme moins de LUT (look-up table), permet de réaliser la transformation en moins de temps (T période d'horloge) et assure le calcul de la DCT avec une consommation de puissance très inférieure à celle de l'algorithme de référence. Ceci permet d'avoir un faible produit surface-temps (ADP = area delay product) et une faible EoC (Energie par coefficient de sortie) ce qui a pour conséquence d'augmenter la durée de vie des batteries pour des applications mobiles.

Design	**N**	**LUT**	**T (ns)**	**P (mW)**	**ADP**	**EoC (pJ)**
Ref	4	304	4.33	3.02	1316	9.80
	8	1392	4.98	13.26	6932	41.27
	16	5002	5.65	55.68	28261	137.63
	32	18772	6.21	209.13	115552	365.25
Proposé	4	76	2.28	1.41	173	1.60
	8	240	2.31	2.93	554	2.53
	16	672	2.34	6.29	1572	3.67
	32	1760	2.37	16.71	4171	6.18

Tableau 1 –*Résultats de synthèse sur FPGA pour l'algorithme proposé et l'algorithme de référence (calcul exact)*

De plus, l'architecture unifiée pour le calcul de la DCT et IDCT 32-point permet d'aller au-delà de ces améliorations de performances. Le Tableau 2 résume les résultats de synthèse en termes d'occupation de surface (nombre de LUT) pour l'algorithme de référence et l'algorithme proposé. Les architectures 1 et 2 du tableau 2 font référence respectivement à la réalisation séparée (un circuit pour la DCT et un autre pour la IDCT) et à la réalisation unifiée.

Design	**Directe**	**Inverse**	**Directe + Inverse**
Ref	18772	16488	35260
Architecture 1	1760	2022	3782
Architecture 2	-	-	2524

Tableau 2 –*Résultats de synthèse sur FPGA pour l'algorithme proposé (architecture séparée et unifiée) et l'algorithme de référence (calcul exact)*

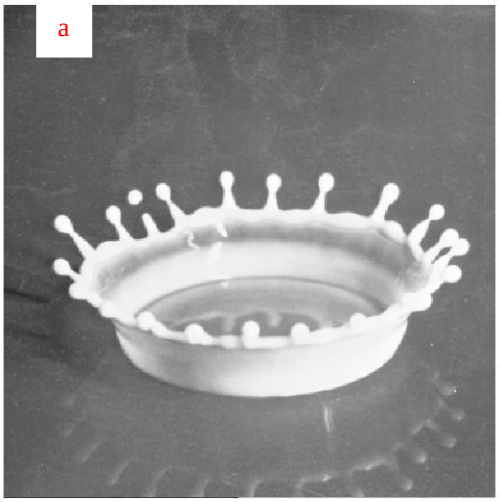
A partir des résultats présentés dans ce tableau, nous trouvons que l'architecture unifiée permet d'avoir un gain de plus de 33% par rapport à l'architecture 1 et utilise 14 fois moins de ressources matérielles que l'architecture de l'algorithme de référence.

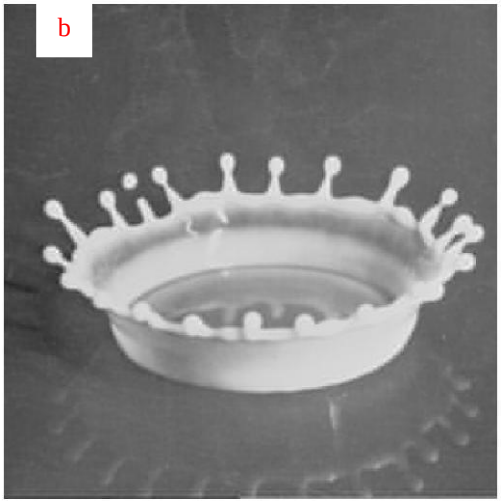
De plus, nous avons mis en place des architectures reconfigurables qui permettent de calculer 8 vecteurs de DCT 4-point, 4 vecteurs de DCT 8-point, 2 vecteurs de DCT 16-point ou 1 vecteur de DCT 32-point en partageant les ressources matérielles. Un gain en surface autour de 54% a été obtenu. Enfin, il est important d'indiquer que les fréquences maximales de fonctionnement sont élevées ce qui permet d'obtenir des débits (throughput) considérables autour de 13,21 GS/s, rendant ainsi possible le traitement de séquences vidéos 8K au format UHDTV.

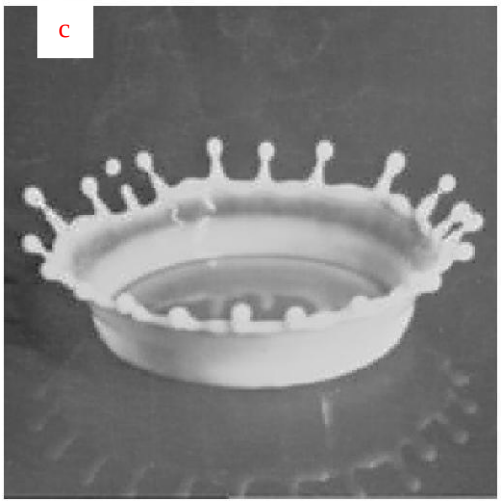
5 Résultats de simulation

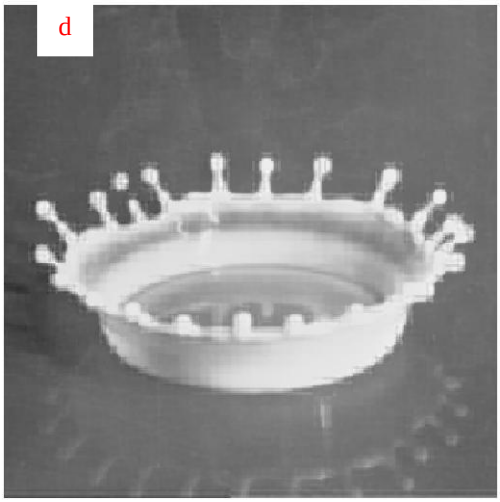
Nous discutons ici les performances de la méthode d'approximation proposée en termes de qualité d'image reconstruite après compression. Nous comparons les performances par rapport à l'algorithme de référence et aux algorithmes d'approximation existants. Nous avons utilisé le schéma de compression indiqué dans [8] pour comparer les différentes méthodes. Pour chaque bloc de 2D-DCT de taille *NxN* nous considérons que *r* coefficients pris du bloc *NxN* dans l'ordre zigzag. Tous les autres coefficients sont fixés à zeros. Plus particulièrement, nous ne retenons que *r* coefficients de sortie tels que $r_{\min} \le r \le r_{\max}$. Par exemple, pour un bloc de taille 8x8, nous considérons $r_{min}=1$ et $r_{max}=32$, ce qui correspond à un rapport de compression entre 50% et 98.43%. De plus, pour garder les mêmes dynamiques de rapport de compression pour les différentes tailles de DCT, nous multiplions r_{min} et r_{max} par 4 à chaque fois que la taille *N* double. De ce fait, le nombre des coefficients retenus varie dans l'intervalle [4,128] et [16,512], respectivement pour des blocs de tailles 16x16 et 32x32.

Les résultats de simulations sont présentés dans Figure 5. Nous avons considéré des images en niveau de gris sur 8-bits prise de la base d'image dans [9]. Les transformations existantes, théorique et proposée de tailles 8x8 ont été appliquées à cette image. Il est évident que la reconstruction avec la transformation exacte donne les meilleurs résultats au niveau de la qualité visuelle. Mais nous pouvons observer aussi que l'approximation proposée et celles existantes donnent aussi des qualités d'image acceptables. Il s'agit de mesure de qualité subjective avec laquelle nous pouvons confirmer que la méthode proposée et celle dans [5] présentent une qualité d'image similaire alors que la méthode dans [8] présente une moins bonne qualité notamment aux niveaux des contours.









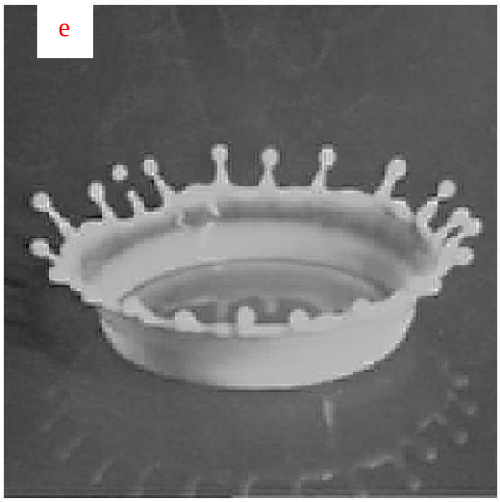


Figure 5 – Reconstruction d'une image après compression et décompression. Image d'origine (a). Image reconstruite avec transformation exacte (b), proposée (c), approximation [5] (d), approximation [8] (e).

D'un point de vue plus objectif, nous avons considéré le PSNR comme paramètre de mesure de qualité. Nous avons varié le nombre de coefficients retenus *r* et nous avons calculé la moyenne de PSNR obtenu pour 30 images prises de la base d'images dans [9], pour les différentes transformations. Les résultats de ces simulations sont présentés dans la Figure 6. Pour la DCT proposée de taille 8-points, les performances en termes de PSNR sont identiques à celles obtenues par la méthode BAS [7].

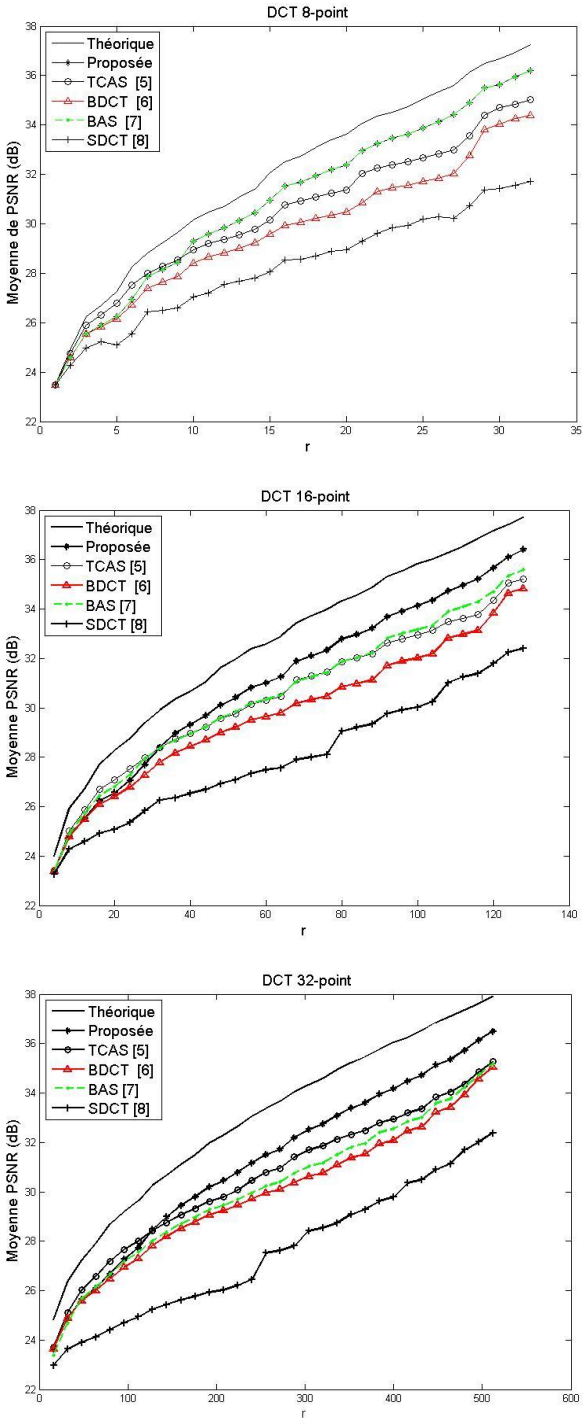


Figure 6 – Moyenne PSNR pour 30 images

De manière plus globale, les moyennes de PSNR indiquent que la transformation proposée présente un réel avantage en termes de qualité d'image en comparaison aux résultats obtenus par les différentes transformations.

6 Conclusion

Nous avons montré à travers ce travail que l'approximation de calcul des transformations peut présenter un réel intérêt pour la réduction de la complexité de calcul des algorithmes de codage et de décodage des données multimédias tout en gardant une certaine qualité d'images. Éliminer les multiplications et réduire le nombre d'additions permet ainsi de réduire les ressources matérielles requises, réduire le chemin critique et par conséquent augmenter la fréquence maximale de fonctionnement. Ceci affecte aussi la facture énergétique des applications de traitement d'images qui constitue la pierre angulaire des contraintes de conception des systèmes embarqués pour les prochaines années.

Références

[1] Cisco-Networks, "Visual Networking Index: Global Mobile Data Traffic Forecast Update 20142019," 3 Feb. 2015.

[2] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," IEEE Trans. Circuits and Syst. Video Technol, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] F. Bossen, B. Bross, K. Suhring, and D. Flynn, "HEVC complexity and implementation analysis," IEEE Trans. Circuits and Syst. Video Technol, vol. 22, no. 12, pp. 1685–1696, 2012.

[4] W.-H. Chen, C. Smith, and S. Fralick, "A fast computational algorithm for the discrete cosine transform," Communications, IEEE Trans. on, vol. 25, no. 9, pp. 1004–1009, Sep 1977.

[5] M. Jridi, A. Alfalou, and P. K. Meher, "A generalized algorithm and reconfigurable architecture for efficient and scalable orthogonal approximation of DCT," Circuits and Systems I: Regular Papers, IEEE Transactions on, vol. PP, no. 99, pp. 1–9, 2014.

[6] S. Bouguezel, M. Ahmad, and M. Swamy, "Binary discrete cosine and Hartley transforms," IEEE Trans. Circuits and Syst. I, vol. 60, no. 4, pp. 989–1002, 2013.

[7] ——, "A novel transform for image compression," in Circuits and Systems (MWSCAS), 2010 53rd IEEE International Midwest Symposium on, 2010, pp. 509–512.

[8] T. Haweel, "A new square wave transform based on the DCT," Signal. Process., vol. 82, 2001.

[9] S. University of Southern California and I. P. Institute, USC-SIPI Image Database. http://sipi.usc.edu/database/.

Réseaux de neurones convolutionnels profonds pour la reconnaissance d'action dans les vidéos

Omar Seddati, Stéphane Dupont, Saïd Mahmoudi
Computer Science - TCTS Lab
{omar.seddati, stephane.dupont, said.mahmoudi}@umons.ac.be

Résumé

Cet article traite la reconnaissance d'actions humaines dans des vidéos. Les vidéos contiennent deux types d'information, le premier est statique relatif à l'apparence et le deuxième est dynamique lié aux mouvements. Dans ce travail, nous nous sommes intéressés principalement à l'information dynamique. Nous utilisons des réseaux de neurones convolutionnels profonds (ConvNets) qui prennent des flux optiques denses en entrée. Nos contributions résident dans l'étude et l'amélioration (menée sur le benchmark HMDB51) des composants d'un système de reconnaissance d'actions dans des vidéos. Nous avons commencé par le choix de l'algorithme de flux optique (deux variantes), la représentation des cartes de flux optique (trois approches), la fusion des décisions (cinq méthodes), la longueur de la séquence en entrée (trois variantes). Et pour finir, nous présentons un ConvNet avec une nouvelle architecture (avec moins de paramètres et des vecteurs de caractéristiques plus courts) qui permet d'obtenir sur les bases HMDB51 et UCF101 les taux de classification parmi les plus élevés de la littérature[1].

Mots clefs

Reconnaissance d'actions, ConvNets, flux optiques, HMDB51, UCF101.

1 Introduction

La reconnaissance des actions humaines dans les vidéos est devenue un domaine de recherche très actif au cours des deux dernières décennies. Cependant, trouver des caractéristiques pertinentes et des systèmes de classification performants pour ce genre d'application représente toujours un réel défi. Cela est dû au grand nombre de configurations spatiales et temporelles possibles dans ce type de contenu (cela inclut les mouvements de la caméra), ainsi que la disponibilité limitée de ressources (bases de données vidéo) pour les entraînements, et la difficulté de l'annotation des vidéos.

Jusqu'à présent, plusieurs approches proposées dans le domaine du traitement d'images classiques ont été réutilisées pour les vidéos. Quelques-unes de ces méthodes ont été étendues pour les adapter à une utilisation avec les vidéos. Par exemple, 3D-SIFT [4], SURF étendu (ESURF) [5], HOG3D [6]. Vig et al. [12] ont utilisé des algorithmes de saillance inspirés par des mécanismes attentionnels humains pour trouver les régions informatives. En général, l'utilisation de sac-de-caractéristiques avec des caractéristiques espaces-temps locales [8] permet d'atteindre de hautes performances pour la classification d'actions (sans utilisation de suivi ou de segmentation).

En outre, les flux optiques ont été largement considérés comme un élément d'information supplémentaire. L'idée est d'utiliser les flux optiques pour faire un échantillonnage dense et de suivre les points caractéristiques dans chaque trame. Ces approches ont montré d'excellentes performances sur une variété de jeux de données. Subramanian et al. [7] ont proposé un système basé sur les flux optiques 3D (combinaison de flux optique 2D et flux de profondeur) et d'un système d'inférence neuro-floue métacognitive (McFIS). Nous pouvons également noter que de nombreuses approches ont été utilisées afin d'examiner le mouvement de la caméra. Uemura et al. [10] utilisent les caractéristiques de l'image et des techniques de segmentation pour distinguer les éléments mobiles et statiques dans la scène, quel que soit le mouvement de la caméra. Wu et al. [11] ont proposé une approche utilisant des méthodes d'optimisation pour séparer les composants de mouvements induits par les différents objets de ceux induits par la caméra. Jiang et al. [13] applique un partitionnement de données aux trajectoires denses, afin de donner une représentation robuste (aux mouvements de la caméra), à travers l'identification des points de référence de mouvements globaux et locaux. Wang et Schmid [14] supposent que deux trames consécutives sont liées par une homographie, et utilisent les trajectoires denses pour améliorer la reconnaissance d'actions en estimant le mouvement de la caméra. D'autres ont utilisé des approches basées sur l'apprentissage profond. Karpathy et al. [3] ont utilisé des ConvNets [1] multirésolution appliqués directement aux pixels. Ils ont également étudié de multiples techniques de fusion afin d'améliorer l'extraction des caractéristiques spatio-temporelle et la classification.

[1]Ce travail a été financé en partie par le projet IMOTION de Chist-Era avec la contribution du Fonds belge de la Recherche Scientifique (FNRS), contrat numéro R.50.02.14.F.

En général, l'utilisation des ConvNets permet de faire de la classification sans avoir recours à des systèmes d'extraction de caractéristiques classiques. Cependant pour le cas des vidéos, aujourd'hui nous pouvons trouver dans la littérature de plus en plus de travaux qui utilisent les flux optiques comme pré-calcul. Par exemple, Simonyan et al. [2] ont proposé une architecture qui combine deux ConvNets (un pour l'information spatiale appliqué à des pixels, et l'autre pour l'information temporelle appliqué à des cartes denses de flux optique). Il s'agit d'une adaptation de la configuration utilisée dans [18] (classification d'images) pour faire de la reconnaissance d'actions. Dans le travail que nous proposons dans cet article, nous nous basons sur l'approche de Simonyan et al. [2], et nous étudions davantage les différentes parties du système proposé. Nos contributions dans ce travail se présentent comme suit :

- Nous avons d'abord comparé deux approches d'extraction de flux optique.
- Nous proposons l'utilisation de différentes représentations pour les vecteurs de déplacement des cartes denses de flux optique.
- Nous comparons plusieurs techniques de fusion de décision. Dans l'architecture proposée, l'intégration des informations provenant des différentes trames d'une vidéo est réalisée à la fin.
- Afin d'extraire des caractéristiques liées aux mouvements, nous présentons une séquence de cartes en entrée des ConvNets proposés. Dans cet article nous avons testé trois longueurs de séquences différentes.
- Enfin, nous présentons aussi une nouvelle architecture performante, avec moins de paramètres, et qui produit des vecteurs de caractéristiques plus courts (ce qui les rend intéressants pour faire de l'indexation vidéo par exemple).

Des tests ont été menés avec la première division de la base de données HMDB51 [16]. En utilisant notre ConvNet temporel nous atteignons une précision de 50.33%, comparé au 46,6% obtenu dans [2] (lorsque les mêmes données d'entraînement sont utilisées). En même temps, une fusion des décisions du ConvNet temporel avec notre ConvNet spatial permet d'atteindre une précision de 59.08%.

2 Approche

Notre système est basé sur l'utilisation des ConvNets, qui représentent l'état de l'art dans le domaine de la reconnaissance d'objets dans des images. Les ConvNets appliqués aux images 2D sont adaptés pour capturer des configurations spatiales. Pour capturer des informations temporelles (qui sont liées à des changements qui se produisent entre les images d'une vidéo) nous utilisons une séquence d'images en entrée. Le premier inconvénient de cette approche est l'augmentation significative du nombre de dimensions et la complexité du problème résultante. Afin de rendre le problème moins complexe, nous utilisons des cartes de flux optique denses en entrée au lieu des pixels. Chaque carte représente le déplacement de chaque

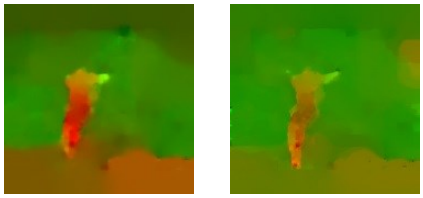


Figure 1: *Cartes de flux optique dense pour la même paire d'images (gauche: Brox ; droite : TV-L1).*

pixel entre deux images successives. Avec cette méthode, l'information temporelle locale est projetée dans un espace similaire à l'espace des pixels. Cela permet de réaliser des systèmes performants pour le traitement d'informations temporelles en utilisant des ConvNets.

Les algorithmes de calcul flux optique :

Après avoir analysé et inspecté visuellement le résultat de l'application de différents algorithmes de calcul de flux optiques. Nous avons remarqué que les résultats sont fort différents. Certains produisent beaucoup plus de bruit que d'autres (plusieurs paramètres ont été testés). Cela nous a conduits à garder deux méthodes parmi celles testées : TV-L1 [17] et la méthode Brox proposée dans [15] et utilisé par Simonyan et al. dans [2] (on peut voir un exemple de résultat des deux approches pour la même entrée dans la figure 1).

La représentation des flux optiques :

Nous avons ensuite sélectionné et testé différents formats de données pour les vecteurs de flux optiques qui représentent les déplacements des pixels. Nous avons commencé par la représentation classique, dans laquelle deux valeurs sont utilisées, une pour la composante horizontale, et l'autre pour la composante verticale. Ensuite, nous avons testé une représentation polaire (r, thêta) du vecteur. Nous avons également testé l'utilisation

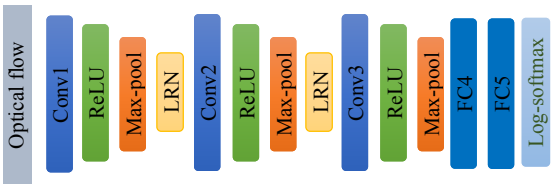


Figure 2: *L'architecture de base des ConvNets utilisés pour les tests.*

du rayon sans angle, afin de vérifier que le ConvNet capture des caractéristiques plus intéressantes que la simple forme des objets en mouvement dans la scène.

Architecture :

Afin de choisir la meilleure configuration des données (la méthode des flux optiques + la représentation des flux optiques), nous avons d'abord utilisé une version basse résolution de HMDB-51 et une sélection de petites architectures pour nos ConvNet. La Figure 2 montre l'architecture de base de ces derniers, à savoir trois couches de convolution et deux couches FC (fully connected). Cela permet de réduire le temps de calcul lors de la comparaison

des différentes approches. Ensuite nous avons sélectionné et testé deux configurations de données avec un ConvNet qui a la même architecture que celle utilisée dans [2]. Finalement nous avons entraîné notre nouvelle architecture dont la configuration est reprise dans le tableau 5. Tous les résultats sont reportés dans la section 4.

Fusion des décisions :

La taille du ConvNet augmente rapidement avec la longueur des séquences de cartes de flux optique denses utilisées en entrée. Cette contrainte nous impose l'utilisation de courtes sous-séquences des trames vidéo (ou plutôt de cartes de flux optique). Ce qui augmente le risque de sélectionner sur une partie de la vidéo qui ne possède pas suffisamment d'informations pour identifier l'action. Pour résoudre ce problème, nous utilisons une approche de fenêtre glissante lors de la phase du test, et nous estimons la probabilité pour chaque action en utilisant plusieurs sous-séquences de la même vidéo. Dans [2], les auteurs calculent la probabilité moyenne pour chaque action à partir des probabilités provenant des différentes sous-séquences. La classe gagnante est celle qui possède la moyenne la plus élevée. Dans notre travail, nous étudions plus en détail cette étape et nous comparons les résultats de quatre autres techniques de fusion de décisions. (1) l'approche Maximax où le gagnant est celui qui a la probabilité la plus élevée. (2) La méthode Maximin, dont le gagnant a la probabilité minime la plus élevée. (3) La méthode basée sur le vote majoritaire sélectionne la classe qui a la majorité des voix, sachant que nous avons un vote pour chaque échantillon. (4) Enfin, la méthode produit (produit des probabilités) donne le meilleur résultat et nous l'avons utilisée pour tous les tests intermédiaires.

3 Implémentation

Pour l'extraction des flux optiques, nous utilisons la boîte à outils OpenCV, offrant des implémentations GPU de quelques algorithmes comme Brox et TV-L1 [15, 17]. Ce calcul est effectué une seule fois (pour chaque algorithme et résolution), et le résultat est stocké sur le disque.

Lors de l'entraînement, les cartes de flux optique sont stockées dans la RAM pour garantir un accès rapide. Afin de réduire la quantité de mémoire nécessaire pour stocker ces cartes, nous calculons pour chaque carte le maximum (M) et le minimum (m) de toutes les valeurs (ces valeurs sont stockées dans un vecteur H). Chaque valeur v de la carte est alors remplacée par la version 8-bit quantifiée de $(255 \times (v - m) / (M-m))$. Les cartes sont ensuite compressées au format JPEG et stockées sur le disque. De plus, avant de commencer l'entraînement d'un ConvNet, nous chargeons le vecteur H et toutes les cartes JPEG (sans décompression) dans la mémoire vive. Pendant l'entraînement, une fois qu'une séquence a été aléatoirement sélectionnée pour être utilisée comme entrée, les données concernées sont décompressées et le vecteur H est utilisé pour faire la reconstruction. Cela permet de remettre à l'échelle les valeurs de chaque carte (les vecteurs de déplacement des flux optiques) à leurs valeurs d'origine. Nous avons également évalué l'importance du format utilisé pour représenter les vecteurs de déplacements des flux optiques. Nous avons testé les trois techniques décrites dans la section 2. Les résultats seront présentés dans la section suivante.

Notre système est basé sur l'utilisation de la boîte à outils Torch [20] qui offre un ensemble puissant et varié d'outils, notamment pour la mise en place et l'entraînement de ConvNets.

Dans cette section, nous allons décrire la méthode utilisée pour entraîner et évaluer une architecture similaire à celle utilisée dans [2]. La résolution des cartes de flux optiques stockés est de 256×256. Le ConvNet prend des séquences qui ont une résolution de 224×224. Cela permet de faire de l'augmentation de données en recadrant la séquence. Pour les autres expériences, quelques modifications minimes sont apportées lorsque la longueur de la séquence et/ou la résolution des cartes changent.

Entraînement : lors de l'entraînement, à chaque itération, nous sélectionnons de façon aléatoire 128 séquences (mini-batch = 128) qui seront utilisées comme entrée de notre ConvNet. Chaque séquence contient dix flux optiques successifs (20 cartes de taille 224 x 224) appartenant à la même vidéo. Nous entraînons le réseau à l'aide de l'algorithme du gradient stochastique (mini-lot). Le momentum est égal à 0,9 et le taux d'apprentissage vaut 10^{-2} à l'instant t = 0. Il est ensuite réduit au fur et à mesure jusqu'à 10^{-3}. L'entraînement est arrêté après 70 époques (1 époque = 102400 échantillons traités par le ConvNet).

Surapprentissage: Afin de réduire le surapprentissage, nous utilisons la technique dropout, ainsi que l'augmentation de données en recadrant et en retournant au hasard les cartes de la séquence d'entrée. Nous estimons aussi le mouvement de la caméra en calculant la moyenne à travers les cartes de la même composante (horizontale et verticale), puis on soustrait cette moyenne de la séquence correspondante.

Test: nous avons utilisé une approche similaire à celle proposée dans [2], qui consiste à prendre 25 séquences uniformément réparties sur la vidéo sélectionnée. Nous générons par la suite 10 séquences pour chaque séquence, en recadrant (quatre coins + le centre) et en tournant les cartes de la séquence. La décision finale est obtenue en utilisant la technique du produit au lieu de la moyenne.

4 Evaluation

Nous allons commencer par la présentation des résultats intermédiaires obtenus en utilisant de petits ConvNets, tous basés sur l'architecture montrée dans la figure 2. Les différences entre ces ConvNets sont minimes, mais nécessaires pour adapter l'architecture de base aux différents types d'entrées. Par exemple, nous devons ajuster la première couche si nous changeons la longueur des séquences d'entrée.

Jeu de données : l'évaluation est effectuée en utilisant la base de données HMDB51, qui représente un benchmark largement utilisé dans le domaine de la reconnaissance d'actions. Nous avons utilisé dans ce qui suit la première division (split 1). En effet, HMDB51 contient 51 catégories d'actions. Pour chaque catégorie nous avons 70 vidéos pour l'entraînement et 30 vidéos pour le test (180 images / vidéo en moyenne).

L'algorithme de flux optique et le format des vecteurs de déplacements : le tableau 1 montre les résultats obtenus pour deux méthodes de flux optiques, ainsi que les différents formats de représentation (section 2). Tout d'abord, nous pouvons constater que l'algorithme TV-L1 donne de meilleurs résultats. En second lieu, la sélection du format de représentation a un impact significatif sur les résultats. Enfin, l'utilisation du rayon seulement donne de meilleurs résultats qu'une représentation polaire. Très probablement, les cartes d'angle sont moins appropriées pour les ConvNets et perturbent les informations provenant de cartes de rayon.

	Précision (%)		
Format	(x, y)	(Θ,r)	(r)
TV-L1	39.67	28.19	32.94
Brox	35.55	24.36	27.12

Tableau 1 : *Précision en fonction de l'algorithme de flux optique et du format.*

La longueur de la séquence : Dans [2] des longueurs de séquence (L) différentes ont été testés. Ils ont remarqué que pour L> 10 l'amélioration est non significative. Nous avons soupçonné que la raison est liée à la taille du réseau (un nombre élevé de paramètres). Donc, nous avons répété l'expérience avec un plus petit ConvNet. Les résultats sont rapportés dans le tableau 2.

	Précision (%)		
Longueur	10	20	25
	39.67	40.39	42.48

Tableau 2 : *La précision pour différentes longueurs de séquences.*

Afin de pouvoir comparer nos résultats avec ceux obtenus dans [2], nous avons entraîné une architecture identique à celle utilisée dans [2] en utilisant les deux algorithmes de flux optique. Les résultats sont reportés dans le tableau 3. Comme on peut le constater, l'utilisation de l'algorithme TV-L1 permet d'améliorer les résultats.

Dropout	TV-L1	Brox
0.5	48.69%	43.27%
0.9	51.24%	46.13%

Tableau 3 : *Comparaison des résultats obtenus en utilisant TV-L1 et Brox avec l'architecture de [2].*

Fusion des décisions : Comme expliqué dans la section 2, nous avons testé cinq techniques de fusion de décisions. Le tableau 4 montre les résultats de notre meilleur ConvNet en utilisant ces cinq techniques de fusion.

Méthode	Maximin	Mean	Maximax	M.V.	Produit
Précision %	44.38	49.73	43.40	48.36	51.24

Tableau 4 : *Précision en utilisant différentes techniques de fusion.*

Le ConvNet temporel de [2] atteint une précision de 46,6% sur la base de données HMDB51 (sans utilisation de données supplémentaires). Notre analyse nous a permis d'apporter des modifications au système de base et d'améliorer les résultats par une grande marge. Comme dans la majorité des travaux basés sur les ConvNets, une variante de l'architecture AlexNet [21] est utilisée. Cette architecture a été proposée pour faire de la classification d'images. À la base, cette architecture a été entraînée sur ImageNet [19], un benchmark avec 1000 catégories et plus d'un million d'images. Le nombre important de catégories et d'exemples dans ce genre de benchmark peut justifier le besoin d'utiliser une architecture avec autant de paramètres. Tandis que dans le cas d'un jeu de données comme HMDB51 nous pouvons utiliser une architecture avec moins de paramètres.

Ind	Type	Nombre de filtres	Taille des filtres	Stride
1	Conv	32	7×7	2
2	ReLU	-	-	-
3	Maxpool	-	3×3	2
4	Normalisation	-	5	-
5	Conv	96	5×5	1
6	ReLU	-	-	-
7	Maxpool	-	3×3	2
8	Normalisation		5	-
9	Conv	96	3×3	1
10	ReLU	-	-	-
11	Maxpool	-	3×3	2
12	Conv	96	3×3	1
13	ReLU	-	-	-
14	Maxpool	-	3×3	2
15	Conv	96	3×3	1
16	ReLU	-	-	-
17	FC	1024	-	-
18	ReLU	-	-	-
19	Dropout (0.5)	-	-	-
20	FC	512	-	-
21	ReLU	-	-	-
22	Dropout (0.5)	-	-	-
23	FC	51	-	-
24	LogSoftmax	-	-	-

Tableau 5 : *La configuration du nouveau ConvNet temporel.*

Dans ce travail nous présentons une nouvelle architecture (tableau 5) qui permet d'obtenir des résultats très proches de notre meilleur ConvNet (50.33% au lieu de 51.24% pour

HMDB51 division 1), tout en réduisant considérablement le nombre de paramètres (entraînement plus rapide + nécessite moins de ressources). Nous avons aussi entraîné une architecture similaire sur le jeu de données UCF101 [22]. La seule modification apportée se situe au niveau de la dernière couche FC pour obtenir 101 sorties au lieu de 51. Sur ce jeu de données, nous avons atteint une précision de 72.69% (split 1). Ce résultat est inférieur au 81% obtenu dans [2] pour le même jeu de données. Cependant il faut noter qu'UCF101 a presque deux fois le nombre de catégories et de vidéos que HMDB51. Ce qui pourrait justifier le besoin de l'utilisation d'une architecture plus grande.

Finalement, nous avons aussi entraîné un ConvNet spatial (une variante de l'architecture VGG [9]) sur ImageNet. À la fin de l'entraînement, le taux d'erreur sur les données de validation est de 40.4% (top 1). Nous avons utilisé ce ConvNet pour extraire des vecteurs de caractéristiques avec 4096 valeurs à partir des images extraites de HMDB51 (25 images par seconde). Ces vecteurs ont été stockés et réutilisés par la suite pour entraîner un réseau de neurones avec trois couches FC (4096-4096-51). Pour le test nous avons utilisé la même approche que [2]. Notre réseau de neurones réalise un taux de classification de 43.2% sur HMDB51. Et une fusion des décisions en provenance du ConvNet spatial avec le ConvNet temporel (la nouvelle architecture) permet de faire passer la précision à 59.08%.

5 Conclusion

Dans ce travail, nous avons réalisé une analyse approfondie des différentes parties du système proposé dans [2]. Cette étude a été accompagnée de plusieurs propositions pour améliorer la reconnaissance d'actions, à commencer par une comparaison de deux méthodes de flux optiques et le format des différentes représentations. Nous avons également testé cinq méthodes de fusion de décision et montré expérimentalement que notre analyse était permet d'améliorer les résultats. Enfin, nous avons proposé une nouvelle architecture plus rapide et qui demande moins de ressources que celle utilisée dans [2], et cela tout en gardant des taux de classification parmi les plus élevés de la littérature.

Références

[1] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

[2] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (pp. 568-576).

[3] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4] Scovanner, P., Ali, S., & Shah, M. (2007, September). A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the 15th international conference on Multimedia (pp. 357-360). ACM.

[5] Willems, G., Tuytelaars, T., & Van Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In Computer Vision–ECCV 2008 (pp. 650-663). Springer Berlin Heidelberg.

[6] Klaser, A., & Marszalek, M. (2008). À spatio-temporal descriptor based on 3d-gradients.

[7] Subramanian, K., Radhakrishnan, V. B., & Sundaram, S. (2014, April). An optical flow feature and McFIS based approach for 3-dimensional human action recognition. In Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2014 IEEE Ninth International Conference on (pp. 1-6). IEEE.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In VS-PETS, 2005.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[10] Uemura, H., Ishikawa, S., & Mikolajczyk, K. (2008, September). Feature Tracking and Motion Compensation for Action Recognition. In BMVC (pp. 1-10).

[11] Wu, S., Oreifej, O., & Shah, M. (2011, November). Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 1419-1426). IEEE.

[12] Vig, E., Dorr, M., & Cox, D. (2012). Space-variant descriptor sampling for action recognition based on saliency and eye movements. In Computer Vision–ECCV 2012 (pp. 84-97). Springer Berlin Heidelberg.

[13] Jiang, Y. G., Dai, Q., Xue, X., Liu, W., & Ngo, C. W. (2012). Trajectory-based modeling of human actions with motion reference points. In Computer Vision–ECCV 2012 (pp. 425-438). Springer Berlin Heidelberg.

[14] Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 103(1), 60-79.

[15] Brox, T., Bruhn, A., Papenberg, N., & Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In Computer Vision-ECCV 2004 (pp. 25-36). Springer Berlin Heidelberg.

[16] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011, November). HMDB: a large video database for human motion recognition. In Computer

Vision (ICCV), 2011 IEEE International Conference on (pp. 2556-2563). IEEE.
[17] Sánchez Pérez, J., Meinhardt-Llopis, E., & Facciolo, G. (2012). TV-L1 optical flow estimation.
[18] Zeiler, M. D., & Fergus, R. (2013). Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901.
[19] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. ILSVRC-2012, 2012. URL http://www.image-net.org/challenges/LSVRC/2012/.
[20] Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). Torch7: A matlab-like environment for machine learning. In BigLearn, NIPS Workshop (No.EPFL-CONF-192376).
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.
[22] SOOMRO, Khurram, ZAMIR, Amir Roshan, et SHAH, Mubarak. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

Modèles prédictifs pour les gliomes diffus de bas grade sous chimiothérapie

M. Ben Abdallah[1] M. Blonski[1,2] S. Mézières[3] Y. Gaudeau[1,4] L. Taillandier[1,2] J.M. Moureaux [1]

[1] Université de Lorraine, Centre de Recherche en Automatique de Nancy (CRAN), CNRS UMR 7039, Faculté de Médecine - Bât D - BP 184, Vandoeuvre-lès-Nancy, 54505, France

[2] Service de Neuro-Oncologie, Centre Hospitalier Régional Universitaire de Nancy, Avenue du Maréchal Lattre de Tassigny,54035 Nancy, France

[3] Université de Lorraine, Institut de Mathématiques Elie Cartan, INRIA BIGS CNRS UMR 7502, BP 239, F-54506 Vandoeuvre-lès-Nancy Cedex, France

[4] Université de Strasbourg, 30 Rue du Maire André Traband, Haguenau, 67500, France

Résumé

Les gliomes diffus de bas grade sont des tumeurs cérébrales primitives rares des adultes. Ces tumeurs progressent de manière continue au cours du temps et se transforment, par la suite, en tumeurs de grade supérieur dont la malignité est associée à un handicap neurologique et à une issue fatale. La taille de la tumeur est l'un des facteurs pronostiques les plus importants. De ce fait, il est d'une grande importance d'évaluer le volume tumoral pendant le suivi des patients. On recommande, pour ce faire, l'utilisation de l'IRM comme modalité. En outre, si la chirurgie reste la première option thérapeutique pour les gliomes diffus de bas grade, la chimiothérapie est de plus en plus utilisée (avant ou après une chirurgie potentielle). Cependant, des questions cruciales et difficiles restent à résoudre : l'identification de sous-groupes de patients qui pourraient bénéficier de la chimiothérapie, la détermination du meilleur moment pour entamer une chimiothérapie, la définition de la durée de la chimiothérapie et l'évaluation du meilleur moment pour effectuer une chirurgie ou, le cas échéant, une radiothérapie. Dans ce travail, nous nous proposons d'aider les cliniciens dans la phase de prise de décision, en concevant de nouveaux modèles prédictifs dédiés à l'évolution du diamètre tumoral. Nous proposons deux modèles statistiques (linéaires et exponentiels) que nous avons testés sur une base de données de 16 patients dont la chimiothérapie a duré entre 14 et 32 mois, avec une durée moyenne de 22,8125 mois. Le choix du modèle le plus approprié a été réalisé avec le critère d'information d'Akaike corrigé. Les résultats sont très prometteurs, avec des coefficients de détermination, pour le modèle linéaire, variant entre 0,79 et 0,97 et une valeur moyenne de 0,90. Cela montre qu'il est possible d'alerter le clinicien sur un changement de la dynamique du diamètre tumoral.

Mots clefs

Modèle, Gliomes Diffus de Bas Grade, Chimiothérapie.

1 Introduction

Les Gliomes Diffus de Bas Grade (GDBG) sont des tumeurs cérébrales rares et infiltrantes. Trois phases caractérisent leur évolution :

- La première phase est asymptomatique avec une évolution linéaire et une vitesse de croissance tumorale constante (environ 3,5 mm par an) [1].
- La deuxième phase est associée à peu de symptômes neurologiques (des crises épileptiques le plus souvent), se caractérise par une évolution linéaire et une vitesse de croissance tumorale constante d'environ 4 mm par an [2].
- La troisième phase correspond à la transformation anaplasique (gliome de haut grade) avec une augmentation de la vitesse de croissance tumorale.

Le but des traitements est de retarder autant que possible la transformation anaplasique tout en préservant la qualité de vie [3]. La chirurgie fonctionnelle est la première option thérapeutique en raison de son impact évident sur la survie tout en maintenant ou en améliorant la qualité de vie [4]. La radiothérapie, qui a été le premier traitement complémentaire proposé, a le même impact, quel que soit le moment, de sa réalisation (tôt au moment du diagnostic ou plus tard après progression) [5]. En outre, de plus en plus souvent, la chimiothérapie est prescrite de manière plus précoce dans la gestion des patients, soit en traitement néoadjuvant (comme un potentiel premier traitement avant la chirurgie) soit en cas de progression après la chirurgie pour les patients non ré-opérables [6] [7]. Il est maintenant admis que la modalité d'IRM est recommandée dans le suivi des patients atteints de GDBG [6] et permet d'évaluer le volume tumoral. En effet, le contrôle du volume est

essentiel dans la décision thérapeutique. Dans ce contexte, deux techniques ont été proposées dans la littérature pour estimer le volume tumoral : la technique des trois diamètres [2] et la segmentation [1] qui est considérée comme le gold standard. Quelle que soit la technique adoptée, il faudrait que la stratégie de traitement soit personnalisée pour chaque patient [8]. Dans le cadre de la chimiothérapie, il est très important d'être en mesure d'évaluer l'impact et le bénéfice attendu du traitement afin de le contrôler. De ce fait, la modélisation de la réponse à la chimiothérapie aidera les cliniciens dans la prise de décision en identifiant les sous-groupes de patients qui pourraient bénéficier de la chimiothérapie, en déterminant le meilleur moment pour entamer ce traitement, en fixant la durée de la chimiothérapie et en évaluant le meilleur moment pour effectuer une intervention chirurgicale. En d'autres termes, les cliniciens pourront plus facilement proposer une solution adaptée à chaque patient. A notre connaissance, deux études récentes ont été menées pour répondre à cette problématique et ont offert des résultats intéressants [9] [10]. Les solutions proposées dans ces études sont basées sur une approche microscopique, à l'échelle cellulaire. En outre, la mesure du volume est obtenue par la méthode des trois diamètres. Ici, nous proposons une approche différente, à l'échelle macroscopique, qui est basée sur la conception de deux modèles prédictifs pour les patients de GDBG sous chimiothérapie Témolozomide. Une autre différence clé se trouve dans l'évaluation du volume tumoral qui est basée ici sur la segmentation des IRM. Nos modèles s'appuient sur les données et s'adaptent en temps réel à chaque nouvelle IRM. Cela rend notre approche pratique en routine clinique pour aider, simplement et rapidement, les médecins dans l'établissement de leur stratégie de traitement. Les deux modèles proposés nous permettent de prédire le diamètre tumoral comme une variable évoluant en fonction du temps et sont basés sur un jeu de données d'apprentissage issues de 5 IRM depuis le début de la chimiothérapie. Nous montrons que si un nouveau diamètre observé (correspondant à l'examen d'IRM en cours) est à l'intérieur de l'intervalle de prédiction donné par le modèle, une évolution normale est prévue. Au contraire, si la nouvelle observation est à l'extérieur de l'intervalle de prédiction, un changement dans la dynamique d'évolution du diamètre tumoral est prédit et un message d'alerte est affiché au clinicien afin de l'aider dans la prise de décision.

Le reste du papier est organisé comme suit. Dans le paragraphe 2, la méthodologie et les données utilisées sont présentées. Les modèles sont introduits dans le paragraphe 3. Les résultats sont donnés dans le paragraphe 4. Enfin, le paragraphe 5 prévoit une discussion et une conclusion.

2 Matériels et méthodes

Certains patients atteints de GDBG doivent subir une ou plusieurs cures de chimiothérapie. Après la confrontation avec la question de l'instant optimal d'initiation du traitement, les médecins doivent choisir l'instant d'arrêt de ce dernier (si, bien sûr, il n'y a pas de problème de tolérance). Il n'y a, actuellement, pas de consensus au sein de la communauté de neuro-oncologie sur la modalité de détermination de cet instant. Pour répondre à cette problématique, nous proposons des modèles statistiques de prédiction qui sont basés sur l'évolution du diamètre tumoral au cours du temps. Ce diamètre est obtenu à partir du volume tumoral, qui est numériquement reconstruit à partir de segmentations manuelles avec le logiciel OsiriX. Dans cette étude, nous ne considérerons que les patients en première ligne de chimiothérapie, qui n'auraient subi aucun traitement préalable, hormis la chirurgie. Nous allons également étudier un seul type de chimiothérapie, la Témozolomide (TMZ), qui est la molécule la plus largement utilisée pour les GDBG en raison de sa bonne tolérance par la plupart des patients et de son utilisation pratique (traitement oral). Spontanément, sans aucun traitement, le diamètre tumoral du GDBG, évolue linéairement au cours du temps [2]. Nous avons remarqué que, sous chimiothérapie, pour certains patients, le diamètre évolue linéairement, tandis que, pour d'autres patients, l'évolution est exponentielle. Mais pour une très petite minorité de cas, la courbe d'évolution du diamètre tumoral au cours du temps ne correspond à aucune fonction connue ou spécifique. Dans ce papier, nous proposons, selon la catégorie de la progression du patient, une prédiction de l'instant d'arrêt de la chimiothérapie sur la base d'un modèle soit linéaire soit exponentiel. Notre base de données initiale comprenait 21 patients qui ont suivi une chimiothérapie TMZ. Cinq patients ont été exclus parce qu'ils ne suivaient ni le modèle exponentiel ni le modèle linéaire. Pour ces cinq patients, et pour d'autres cas similaires, nous allons adopter l'approche classique du suivi de la dynamique de croissance tumorale. Dans le présent article, nous allons uniquement aborder les cas de patients suivant un modèle linéaire ou exponentiel, ce qui représente la majorité des cas de notre base de données (16/21 patients), et nous allons laisser la discussion des autres cas pour la suite de nos travaux. Nos modèles prédictifs ont donc été appliqués sur 16 patients dont le traitement a duré entre 14 et 32 mois, avec une durée moyenne de 22,8125 mois. Dans cette première étude, nous avons inclus des tumeurs indépendamment de leurs statuts pathologiques (astrocytomes, oligodendrogliomes, tumeurs mixtes) et moléculaires (y compris IDH 1/2 ou 1p19q).

3 Analyse statistique

Nos modèles de prédiction utilisent un ensemble de données d'apprentissage de diamètres tumoraux $(d_i)_{i=0...n}$, obtenus après segmentation d'examens IRM en période de chimiothérapie, et les temps d'acquisition correspondants $(t_i)_{i=0...n}$. Le nombre de données d'apprentissage disponibles n est faible, souvent autour de 6. Dans tous les cas, afin de se conformer à notre analyse statistique, n doit être au moins égal à 5. Les deux modèles sont testés sur l'ensemble des données de chaque patient et le coefficient de détermination R^2 est calculé pour le modèle linéaire

afin d'évaluer la qualité de sa prédiction. Dans cette étude, le critère d'information d'Akaike corrigé (AIC_c) [11] est appliqué pour distinguer les patients suivant une évolution linéaire de ceux suivant une évolution exponentielle. AIC_c est défini comme suit :

$$AIC_c = n \ln \frac{SS}{n} + 2K + \frac{2K(K+1)}{n-K-1} \quad \text{avec} \qquad (1)$$

SS : C'est la somme du carré de la distance verticale du point par rapport à la courbe.

K : C'est le nombre de paramètres du modèle de régression plus un. Il est égal à 3 pour le modèle linéaire et à 4 pour le modèle exponentiel.

Le meilleur modèle est celui qui a la valeur minimale de AIC_c.
En outre, les deux modèles affichent un message pour alerter le clinicien de toute augmentation ou de toute diminution significative du diamètre tumoral ou pour indiquer une évolution régulière qui pourrait être prise en considération pour mettre fin au traitement.

3.1 Modèle linéaire

Soit D la variable aléatoire représentant le diamètre tumoral suivant un modèle de régression linéaire simple :

$$D = b_0 + b_1 T + \epsilon \quad \text{où}$$

T : la variable aléatoire représentant le moment d'observation.
ϵ : l'écart inexpliqué entre le diamètre observé et le diamètre expliqué par le modèle de régression linéaire.
b_0 : la valeur initiale du diamètre tumoral.
b_1 : la vitesse de de croissance du diamètre tumoral.

On suppose que les ϵ_i (avec $i = 1...n$ le numéro de l'observation) suivent une loi normale $\mathcal{N}(0, \sigma)$, remplissent la condition d'homoscédasticité et sont mutuellement indépendants. La première étape du modèle linéaire est d'estimer les paramètres b_0 et b_1 à partir des données d'apprentissage $(T_i, D_i)_{i=0,...,n}$ en utilisant la méthode des moindres carrés. Nous vérifions que les hypothèses initiales sont remplies : la normalité des termes d'erreur statistique ϵ_i est vérifiée à l'aide du test de Shapiro-Wilk et leur indépendance avec le test de Durbin Watson. En ce qui concerne la condition d'homoscédasticité, elle est vérifiée avec un test de White. Une fois ces conditions établies, nous appliquons un test de Student sur b_1 afin d'évaluer l'importance de la régression. La qualité de prédiction du modèle linéaire est analysée avec le coefficient de détermination R^2.
Une fois le modèle validé, nous convenons de l'évolution linéaire de la tumeur avec les paramètres estimés $\hat{b}_0$, $\hat{b}_1$. Une nouvelle observation est, ensuite, prédite à l'instant T_{n+1} comme suit :

$$\hat{D}_{n+1} = \hat{b}_0 + \hat{b}_1 T_{n+1}$$

Un intervalle de prédiction est également défini pour un risque de première espèce α donné, fournissant ainsi la gamme de valeurs dans laquelle varie $\hat{D}_{n+1}$ pour un T_{n+1} donné. Si la nouvelle observation D_{n+1} est à l'intérieur de l'intervalle de prédiction, une évolution normale du diamètre tumoral est annoncée. Si elle quitte l'intervalle de prédiction, un changement dans la dynamique d'évolution du diamètre tumoral est observé. Cela peut être médicalement interprété soit positivement (dans le cas où le diamètre diminue en dessous de la borne inférieure de l'intervalle de prédiction, annonçant vraisemblablement une bonne réponse au traitement) soit négativement (croissance de la tumeur sous chimiothérapie au-dessus de la borne supérieure de l'intervalle de prédiction, le patient ne répond pas au traitement).

3.2 Modèle exponentiel

Soit $D = ae^{-bT} + c + \epsilon$ la variable aléatoire représentant le diamètre tumoral suivant un modèle exponentiel.
La première étape consiste à estimer les paramètres du modèle a, b, c à partir des données d'apprentissage $(T_i, D_i)_{i=0,...,n}$. Le terme c assure l'adéquation des données avec la réalité où le diamètre tumoral n'égale jamais zéro, indépendamment de l'efficacité de la chimiothérapie. Comme pour le modèle linéaire, on suppose que les termes d'erreur statistique ϵ_i suivent une loi normale $\mathcal{N}(0, \sigma)$, remplissent la condition d'homoscédasticité et sont mutuellement indépendants. La normalité des ϵ_i est vérifiée avec le test de Shapiro-Wilk, l'homoscédasticité est évaluée avec le graphe des résidus en fonction des valeurs fittées, et l'hypothèse d'indépendance est contrôlée à l'aide du graphe de chaque résidu en fonction du résidu précédent. Après avoir confirmé ces hypothèses, un test de Student est appliqué sur a et b afin d'évaluer l'importance de la régression.
Une fois le modèle validé, le diamètre d'une nouvelle observation est estimé ainsi que l'intervalle de prédiction pour un risque de première espèce α donné. Si la nouvelle observation est à l'intérieur de l'intervalle de prédiction, un message est affiché pour annoncer une évolution normale du diamètre. Dans le cas contraire, un message, similaire à ceux décrits pour le modèle linéaire, est affiché annonçant un changement dans la dynamique d'évolution du diamètre tumoral.

4 Résultats

Les résultats de cette étude ont été implémentés en R. Parmi les 16 patients de notre base de données, 13 sont classés comme linéaires et 3 sont classés comme exponentiels. Les coefficients de détermination R^2 pour les patients linéaires varient entre 0,77 et 0,97 avec une moyenne de 0,90. Le risque de première espèce α est fixé à 0,1. Dans cet article, nous présenterons, pour illustration, un exemple de chacun des deux modèles prédictifs.

4.1 Exemple pour le modèle linéaire

La chimiothérapie du patient 1 a duré 27 mois. Pour la phase d'apprentissage, nous disposions de 6 points et nous avons prédit le point précédant la fin de la chimiothérapie. L'estimation des paramètres b_0 et b_1 a donné les valeurs 5,268757 et -0,025601 respectivement. Nous avons ensuite testé l'hypothèse de normalité avec le test de Shapiro-Willk : $W = 0,9888 > Wcrit = 0,713$ pour un risque de première espèce égal à 0,01. Nous avons validé la condition d'homoscédasticité avec le test de White : $f(2,3) = 30,82 > White = 3,7468$ pour un risque de première espèce égal à 0,01. Nous avons également confirmé l'hypothèse de non corrélation avec le test de Durbin Watson : $DW = 1,9082$, valeur comprise entre 1,5 et 2,5 ce qui permet de valider l'hypothèse de non corrélation des résidus. Nous avons également vérifié l'importance de la régression avec un test de Student sur b_1. Nous avons ensuite calculé le coefficient de détermination R^2 qui est égal à 0,87. Nous avons également testé le modèle exponentiel. Les coefficients a, b et c sont estimés à -0,61778, 0,08615304 et 4,71256 respectivement. Nous avons évalué l'hypothèse de normalité avec le test de Shapiro-Wilk : $W = 0,9021 > Wcrit = 0,713$ pour un risque de première espèce égal à 0,01. Les conditions d'homoscédasticité et de non corrélation ont été vérifiées avec les graphes correspondants. Quant à la significativité de la régression, nous l'avons confirmée avec un test de Student sur a et b.
Afin de sélectionner le modèle approprié, nous avons calculé AIC_c pour les deux modèles. $AIC_{c_{Linéaire}} = -6,495425 < AIC_{c_{Exponentiel}} = -1,268149$. Par conséquent, nous avons choisi le modèle linéaire pour nos données. Nous avons ensuite prédit la dernière valeur du diamètre avant la fin de la chimiothérapie, ainsi que l'intervalle de prédiction pour un risque de première espèce égal à 0,1. La courbe des données d'apprentissage, de la nouvelle observation à l'instant suivant et de l'intervalle de prédiction dans la Figure 1 montre clairement que les données suivent la droite de régression et que la nouvelle observation est au-dessus de la borne supérieure de l'intervalle de prédiction. Le message d'alerte affiché est le suivant : "Augmentation significative du diamètre tumoral".

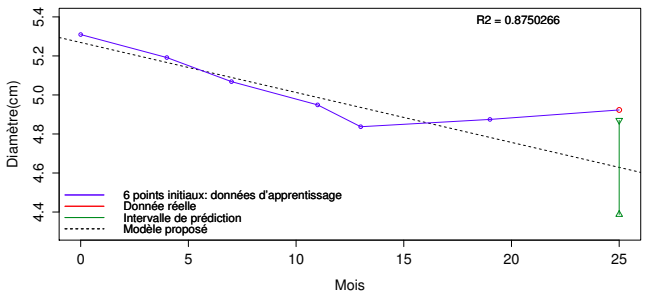


Figure 1 – *Les données d'apprentissage, la nouvelle observation et l'intervalle de prédiction avec le modèle linéaire pour le patient 1.*

4.2 Exemple pour le modèle exponentiel

La chimiothérapie du patient 2 a duré 26 mois. Pour la phase d'apprentissage, nous avions eu 6 points et nous avons prédit le point précédant la fin de la chimiothérapie. L'étape d'apprentissage a donné les valeurs -2,47087, 0,1252605 et 4,93097 pour l'estimation de a, b et c respectivement. L'hypothèse de normalité a été évaluée avec le test de Shapiro-Willk : $W = 0,8965 > Wcrit = 0,713$ pour un risque de première espèce égal à 0,01. Les conditions d'homoscédasticité et de non corrélation ont été vérifiées avec les graphes correspondants. Quant à la significativité de la régression, nous l'avons confirmée avec un test de Student sur a et b. Nous avons également testé le modèle linéaire, mais comme l'hypothèse de non corrélation n'a pas été confirmée ($DW = 1,0995 < 1,5$), ce modèle a été rejeté. A l'aide du modèle exponentiel sélectionné, nous avons finalement estimé le dernier point précédant la fin du traitement ainsi que l'intervalle de prédiction pour un risque de première espèce égal à 0,1. La Figure 2 montre la courbe des données d'apprentissage, des données prédites, de la nouvelle observation et de l'intervalle de prédiction. Les données réelles semblent coller parfaitement au modèle exponentiel et la nouvelle observation est à l'intérieur de l'intervalle de prédiction. Le message affiché est le suivant : "Le diamètre tumoral évolue normalement".

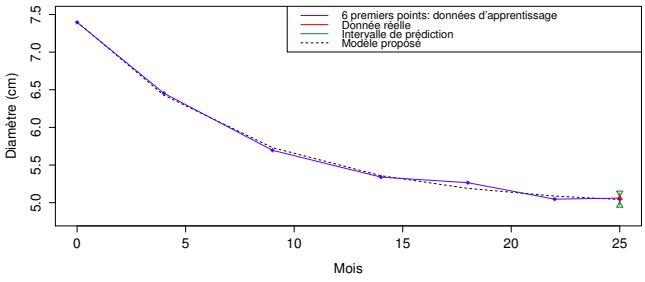


Figure 2 – *Les données d'apprentissage, la nouvelle observation et l'intervalle de prédiction avec le modèle exponentiel pour le patient 2.*

5 Discussions et conclusions

Dans cet article, nous avons présenté une nouvelle approche pour une aide à la décision des cliniciens traitant des patients atteints de GDBG et en cours de chimiothérapie. Cette approche est basée sur deux modèles statistiques de prédiction (linéaire et exponentiel) qui sont fonctions du diamètre tumoral, étant donné que la taille de la tumeur est l'un des facteurs pronostiques les plus importants. Notre étude, menée en partenariat avec le CHRU de Nancy, s'appuie sur les données de 16 patients ayant suivi un traitement de chimiothérapie TMZ. Ce traitement a duré entre 14 et 32 mois, avec une durée moyenne de 22,8125 mois. Dans cette base de données, 13 patients ont suivi le modèle linéaire tandis que 3 patients ont suivi le modèle exponentiel. Sur la base d'un apprentissage incluant au moins 5

IRM, le modèle est capable de prédire le diamètre tumoral suivant avec une précision importante. Grâce à l'intervalle de prédiction correspondant, il est possible de vérifier si la nouvelle observation correspond au diamètre prédit. Si le diamètre réel se situe dans l'intervalle de prédiction, un message annonçant une évolution normale est affiché. Dans le cas contraire, le clinicien est alerté par un message d'une augmentation significative du diamètre tumoral. Il est à noter que la validation des modèles devrait normalement s'appuyer sur un jeu de données différent du jeu de données d'apprentissage. Cependant, le nombre limité de cas ne permet pas de mener à bien cette procédure. Afin d'augmenter la taille des données d'apprentissage et de permettre à nos modèles de prédire plus tôt l'évolution des diamètres en période de chimiothérapie, un nouveau protocole a été mis en place au CHRU de Nancy. Les nouveaux patients qui acceptent de suivre ce protocole bénéficient d'examens IRM supplémentaires non injectés au début du traitement. De plus, nous sommes en cours d'élargissement de notre base de données de GDBG grâce à une collaboration avec le CHRU de Montpellier. Enfin, comme d'autres facteurs moléculaires (mutations IDH 1 ou 2, codélétion 1p19q, la méthylation du promoteur MGMT, le promoteur TERT ou les mutations ATRX, etc.) ou radiologiques (IRM de perfusion, spectroscopie RM, imagerie PET, etc.) pourraient influencer de manière significative la croissance de la tumeur, nous prévoyons d'inclure certains d'entre eux dans la conception de nos modèles. Notre objectif à long terme est de concevoir un outil d'aide à la décision qui, sur la base de différents facteurs (diamètre tumoral, paramètres radiologiques et moléculaires, etc.), livrera un message au clinicien sur l'état d'évolution du GDBG afin de permettre un prise en charge thérapeutique personnalisée des patients.

Références

[1] J. Pallud, D. Fontaine, H. Duffau, E. Mandonnet, N. Sanai, L. Taillandier, P. Peruzzi, R. Guillevin, L. Bauchet, V. Bernier, M. H. Baron, J. Guyotat, et L. Capelle. Natural history of incidental world health organization grade ii gliomas. *Annals of Neurology*, 68 (5) :727–733, Novembre 2010.

[2] E. Mandonnet, J.-Y. Delattre, M. L. Tanguy, K. R. Swanson, A. F. Carpentier, H. Duffau, P. Cornu, R. Van Effenterre, E. C. Jr Alvord, et L. Capelle. Continuous growth of mean tumor diameter in a subset of grade ii gliomas. *Annals of Neurology*, 53 :524–528, Avril 2003.

[3] H. Duffau. Preserving quality of life is not incompatible with increasing overall survival in diffuse low-grade glioma patients. *Acta Neurochirurgica*, 157 (2) :165–167, Février 2015.

[4] A. S. Jakola, K. S. Myrmel, R. Kloster, S. H. Torp, S. Lindal, G. Unsgård, et O. Solheim. Comparison of a strategy favoring early surgical resection vs a strategy favoring watchful waiting in low-grade gliomas. *JAMA*, 308 (18) :1881–1888, Novembre 2012.

[5] M. J. van den Bent, D. Afra, O. de Witte, M. ben Hassel, S. Schraub, K. Hoang-Xuan, P. O. Malmström, L. Collette, M. Piérart, R. Mirimanoff, A. B. Karim, et EORTC. Long-term efficacy of early versus delayed radiotherapy for low-grade astrocytoma and oligodendroglioma in adults : the eortc 22845 randomised trial. *The Lancet*, 366 (9490) :985–990, Septembre 2005.

[6] R. Soffietti, B. G. Baumert, A. von Deimling, H. Duffau, M. Frénay, W. Grisold, R. Grant, F. Graus, K. Hoang-Xuan, M. Klein, B. Melin, J. Rees, T. Siegal, A. Smits, R. Stupp, et W. Wick. Guidelines on management of low-grade gliomas : report of an efns-eano task force. *European Journal of Neurology*, 17 (9) :1124–1133, Septembre 2010.

[7] M. Blonski, J. Pallud, C. Gozé, E. Mandonnet, V. Rigau, L. Bauchet, M. Fabbro, P. Beauchesne, M. H. Baron, D. Fontaine, P. Peruzzi, A. Darlix, H. Duffau, et L. Taillandier. Neoadjuvant chemotherapy may optimize the extent of resection of world health organization grade ii gliomas : a case series of 17 patients. *Journal of Neuro-oncology*, 113 (2) :267–275, Juin 2013.

[8] H. Duffau et L. Taillandier. New concepts in the management of diffuse low-grade glioma : Proposal of a multistage and individualized therapeutic approach. *Neuro-Oncology*, Août 2014.

[9] B. Ribba, G. Kaloshi, M. Peyre, D. Ricard, V. Calvez, M. Tod, B. Cajavec-Bernard, A. Idbaih, D. Psimaras, L. Dainese, J. Pallud, S. Cartalat-Carel, J.-Y. Delattre, J. Honnorat, E. Grenier, et F. Ducray. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. *Clinical Cancer Research*, 18 (18) :5071–5080, Septembre 2012.

[10] P. Mazzocco, C. Barthélémy, G. Kaloshi, M. Lavielle, D. Ricard, A. Idbaih, D. Psimaras, M.-A. Renard, A. Alentorn, J. Honnorat, J.-Y. Delattre, F. Ducray, et B. Ribba. Prediction of response to temozolomide in low-grade glioma patients based on tumor size dynamics and genetic characteristics. *CPT : Pharmacometrics & Systems Pharmacology*, Octobre 2015.

[11] H. Motulsky et A. Christopoulos. *Fitting Models to Biological Data using Linear and Nonlinear Regression : A practical guide to curve fitting*. Oxford University Press, 2004.

Analyse de la symétrie en 4D, le cas de la paralysie faciale

Paul Audain Desrosiers, Boulbaba Ben Amor, Mohamed Daoudi
Telecom Lille, CRIStAL (UMR CNRS 9189)
Yasmine Bennis, Pierre Guerreschi
Service de chirurgie plastique reconstructrice et esthétique, Université de Lille 2
{desrosiers, boulbaba.benamor, mohamed.daoudi}@telecom-lille.fr
yasbennis0311@gmail.com, pierre.guerreschi@chru-lille.fr

Résumé

Dans ce papier, nous abordons le problème de la quantification du niveau d'asymétrie du visage dans le cas de la paralysie faciale. L'objectif final est de proposer des outils pour aider les cliniciens en chirurgie réparatrice à mieux évaluer qualitativement et quantitativement les résultats de leurs interventions. La méthode proposée prend un visage 3D et par une approche Riemannienne récemment développée, permet l'analyse de la déformation de celui-ci. Autrement dit, après le pré-traitement, chaque trame dans le flux de visages 3D est approximée par une collection ordonnée de courbes radiales. L'analyse élastique de la forme des courbes par rapport à leurs symétries dans le visage donne lieu à un champ scalaires de déformation, appelé DSF. Les DSF obtenus révèlent le niveau d'asymétrie bilatéral du visage. Pour illustrer la méthode, un ensemble de données ont été recueillies (de patients) avant et après l'injection de toxine botulique (BT) dans le visage. Les résultats obtenus montrent que l'approche proposée permet aux cliniciens d'évaluer l'asymétrie de la dynamique de formes faciales avant et après intervention.

Mots clefs

Séquence de trames 3D, géométrie Riemannienne, champs scalaires de déformations, symmétrie faciale, paralysie faciale.

1 Introduction

La paralysie faciale est une pathologie invalidante dont l'impact fonctionnel, esthétique et psychologique est majeur. On déplore 15000 nouveaux cas par an en France de paralysie faciale juste pour son origine virale. Au delà de 18 mois d'atteinte, les chances de récupération sont minimes et il faut envisager un traitement palliatif. De ce fait, le traitement médical qui est largement utilisé depuis 1989 dans la chirurgie plastique [1, 2] est l'injection des faible doses de toxine botulique (BT) dans le visage pour compenser la sur-activité de la partie saine de celui-ci. De plus, de multiples procédés chirurgicaux sont combinés pour obtenir une réanimation faciale globale la plus symétrique possible aussi bien au repos que lors du mouvement. Pour guider ses choix thérapeutiques, le chirurgien a besoin d'outils qui lui permettent de mesurer d'une manière objective la symétrie du visage et de ses mouvements avant et après traitement. A ce jour, l'évaluation clinique reste subjective et très dépendante du chirurgien et donc manque de reproductibilité. Les méthodes les plus courantes comprennent l'utilisation d'échelles cliniques validées (House brackmann, Sunnybrook and Terzis) [3] et l'analyse de photographies numériques et de captures vidéos. Des méthodes dites objectives consistant à relever des mesures, avec des règles et des compas sur les données images ont été suggérées. Cependant, ces mesures peuvent être faussées si la position de la personne n'est pas frontale face à la caméra, en outre elles ne permettent pas de mesurer la dynamique faciale ou les différences de formes.

Les récents progrès technologiques ont permis le développement de techniques d'acquisition tridimensionnelle (3-D) non-invasive de la surface faciale, telle que la numérisation par lumière structurée, pour mesurer la forme faciale et son évolution dans le temps (4D ou 3D+t). Ainsi, les erreurs liées à la subjectivité et à l'anthropométrie directe sont évitées et on obtient des informations denses et précises sur la dynamique faciale qui s'ajoute à celles obtenues par analyse statique. Nous voulons concentrer notre travail sur les techniques de réanimation dynamique du tiers inférieur du visage. Deux procédés chirurgicaux prédominent – la greffe nerveuse (anastomose hypoglosso-faciale) et les transferts musculaires (libres ou myoplastie d'allongement du temporal). C'est le chirurgien qui choisit, en fonction de son expérience et du patient, la technique la plus adaptée. Puisqu'il n'existe toujours pas de moyen objectif de comparer les résultats de ces deux techniques chirurgicales, la question de la meilleure stratégie thérapeutique est encore largement débattue.

L'objectif principal de ce travail est de proposer d'utiliser les avancées technologiques et méthodologiques en acquisition et traitement de visages tridimensionnelles dynamique (4D). Cela a pour but de mesurer quantitativement et de manière reproductible les séquelles de la paralysie faciale et l'efficacité des traitements chirurgicaux dans le cadre de la réanimation faciale inférieure. Une analyse de la dissymétrie faciale sera réalisée au cours du temps pour différentes expressions faciales prédéfinies.

2 Etat de l'art

Bien que les technologies d'acquisition 4-D aient fait des progrès importants, ces dernières années, les méthodes d'analyse de formes 3D non-rigides évoluant dans le temps sont encore à un état embryonnaire. La symétrie d'un visage dans un état dynamique est un élément clef pour mesurer l'efficacité des traitements chirurgicaux. Actuellement, la seule échelle d'une mesure subjective de la symétrie faciale à l'étage inférieur est celle de Terzis [3] qui classe les déformations faciales en 5 classes qui reflètent la qualité de la symétrie du visage. Or, la majorité des travaux existants essaient de mesurer la symétrie du visage 3D statique [4]. Peu de résultats existent à l'heure actuelle pour mesurer la symétrie dans des images 4D. Dans l'article de T. Al-Anezi [5], les auteurs utilisent des repères anatomiques (appelés landmarks). Il est reconnu que la détection de points correspondants à ces repères anatomiques est un problème difficile à résoudre. Nous proposons dans ce projet d'analyser toute la forme géométrique 3D du visage, dans sa globalité, sans avoir à détecter ces points. Une première tentative proposée récemment a permis d'obtenir des résultats encourageant concernant l'étude de la symétrie du visage dynamique (4D) [4].
Nous devrions donc apporter des réponses aux questions fondamentales suivantes,

1. Quelles représentations mathématiques de formes faciales statiques (3D) et de leurs dynamiques (4D) adaptées à l'analyse ?
2. Comment quantifier d'une manière précise les déformations le long d'une séquence de formes faciales 3D ?
3. Comment quantifier la symétrie (et l'asymétrie) bilatérale du visage et de ses dynamiques dans le temps ?

Plusieurs méthodes ont été proposées dans la littérature comme l'algorithme Non-rigid ICP [6], le Free Form Deformation (FFD) [7] et le Thin-plate Spline (TPS) [8]. La plupart de ces algorithmes essayent de trouver un alignement spatial optimal entre deux visages 3D. Cependant, leurs fonctions coût qui minimisent la distance entre les maillages 3D ne sont pas une métrique, elle n'est même pas symétrique. Autrement dit, l'enregistrement optimale d'une forme faciale 3D F^1 à une autre forme F^2 peuvent ne pas être le même que l'alignement de F^2 à F^1, ce qui rend difficile l'interprétation des résultats.
Notre approche est basée sur l'idée que l'espace des formes faciales 3D est une structure non-linéaire. Afin qu'on puisse analyser cet espace de formes des visages 3D, il nous faut des outils issus de la géométrie différentielle. Ces outils vont nous permettre non seulement de calculer les déformations optimales entre deux surfaces faciales en calculant les distances géodésiques entre elles, mais aussi de calculer des statistiques telles que la moyenne et la variance. La modélisation de l'évolution temporelle d'une surface faciale est un sujet difficile et nouveau. Nous avons proposé dans [9] une première modélisation des expressions faciales dynamiques (4D) qui approxime une forme faciale 3D par une collection ordonnée de courbes radiales. Puis une analyse élastique de la forme de ces courbes permet à la fois un alignement et une comparaison optimale deux courbes faciales. L'analyse élastique de formes faciales 3D permet entre autre d'obtenir un champs de déformation optimal appelées DSF qui permet d'aller d'une forme à une autre sur l'espace de formes. Dans [9], ces DSFs ont été utilisés pour quantifier les activités faciales dans le temps. Les déformations sont calculées sur une fenêtre de temps glissante puis classifiées par des approches d'apprentissage automatique type Modèles de Markov cachés ou forêts aléatoires. Des résultats très encourageants ont été obtenus sur la base BU-4DFE [10], base spécialiées dans la reconnaissane des expressions faciales. Ce même descripteur DSF a donné de bons résultats pour le calcul de la symétrie d'un visage 3D statique [11].

Dans cet article, nous utiliserons la même approche Riemannienne d'analyse de formes 3D pour quantifier l'asymétrie bilatérale d'un visage 3D en mouvement. Dans la section 3 nous présentons le protocole d'acquisition de séquences faciales 3D sur lesquelles nous avons fait nos tests, nous décrivons brièvement l'étape de prétraitement cruciale pour l'évaluation de l'asymétrie faciale. Dans la section 4, nous décrivons notre pipeline de quantification de l'asymétrie faciale statique et dynamique. Nous rapportons les résultats des tests pré- et post- injection BT dans la section 5. La section 6 donne quelques conclusions et perspectives sur le travail.

3 Jeu de données et Pré-traitement

La nouveauté de ce sujet nous a permis de travailler en étroite collaboration avec les cliniciens dans le domaine de la paralysie faciale. Dans notre protocole d'acquisition, le patient se présente devant le scanner, puis le clinicien lui demande d'effectuer certaines expressions faciales. Pour chacune des expressions, nous avons capturé une séquence de trames 3D.

3.1 Acquisitions du jeu de données 4D

Les acquisitions sont obtenues pour 5 patients atteints de paralysies faciales. Une première session s'est déroulée avant l'injection du BT, et une seconde session a été programmée 15 jours plus tard (donc après l'injection du BT). Pour une acquisition optimale, le patient a été invité à s'asseoir face à un scanner 3D à une distance de 80 cm approximativement, puis on demande aux patients d'effectuer une série d'expressions faciales qui inclut (1) sourire normal, (2) sourire forcé et (3) hausser les sourcils. De plus, nous avons scanné un sujet de contrôle (un sujet à visage sain) afin de pouvoir faire des comparaisons. La durée moyenne des acquisitions pour une séquence 3D est de 4 à 6 secondes avec un total de 60 à 105 trames 3D. Dans le processus d'acquisitions, nous avons utilisé le scanner 3D MTH

d'ARTEC [1]. Ce scanner permet de capturer des séquences de trame 3D avec 15 tps (15 trames par seconde). Chaque trame est un maillage composé d'environ de 5500 points avec une résolution 3D de 0.5 mm.

3.2 Pré-traitements des données

Les acquisitions 3D présentent souvent des trous dues à l'absorption de la lumière par les zones chevelues du visage, l'auto-occultation (des parties qui en cachent d'autres) ou simplement lors de l'ouverture de la bouche. Ces données sont également bruitées et contient des parties indésirable (vêtements, cheveux, etc.). Afin d'extraire la partie utile à l'analyse et corriger les imperfections citées ci-dessus, nous avons appliqué les étapes suivantes aux différents flux de visages 3D :

- Remplissage des trous dans les maillages par interpolation,
- Détections du bout du nez pour chaque trame 3D,
- Recadrage du visage, le bout du nez sert de référence pour définir une sphère avec un rayon constant qui sert à découper la partie informatique de visage (masque facial),
- Réduction du bruit et lissage uniforme par un filtrage Laplacien,
- Application d'un opérateur miroir sur le visage 3D par rapport au plan YZ afin de générer un maillage miroir (qui sera utilisé plus tard dans l'évaluation de la symétrie),
- Utilisation de l'algorithme ICP (Iterative Closest Point) entre le visage et son miroir (obtenu à l'étape précédente) afin de corriger les problèmes de pose.

Les filtres implémentés dans la bibliothèque VTK [2] nous ont permis de mettre en place ce pipeline, qui est appliqué séparément sur chaque trame dans la séquence 3D. De ce fait, on obtient le visage orignal et son miroir. L'étape suivante consiste à extraire des courbes radiales et de quantifier le niveau d'asymétrie du visage 3D en comparant chaque forme 3D de leur réflexion, c'est l'objet de la section suivante.

4 Asymétrie bilatérale

Notre objectif est de mesurer le niveau de l'asymétrie bilatérale d'un visage statique, et celui de son évolution dans le temps au cours d'une expression faciale. Rappelons que dans le cas de l'analyse 3D statique, la méthode la plus utilisée est la détection d'un plan de symétrie du visage (appelé aussi "plan miroir"), puis appliquer des algorithmes d'alignement comme Iterative Closest Point (ICP) pour quantifier la différence des deux moitiés du maillage, comme proposé dans [12]. Cependant la détection du plan de symétrie avec précision pour comparer les deux demis visages n'est pas aussi simple. Car, dans le cas de la paralysie faciale, le visage est déjà asymétrique ce qui rend très délicat l'application de la méthode proposée par dans [12]. Pour mesurer le niveau d'asymétrie bilatérale du visage, nous proposons d'abord d'aligner le visage 3D et son miroir en utilisant une approche de recalage rigide type ICP (Iterative Closest Point) ou le plus proche voisin itérée afin de corriger la pose d'un visage par rapport à l'autre (considéré comme référence). Puis, nous approximons les deux surfaces faciales par des collections de courbes radiales comme proposé dans [9]. C'est la métrique Riemannienne (élastique) définit sur les courbes radiales qui prendra en compte les déformations entre deux surfaces faciales. Cette approche achève un recalage non-rigide (plus fin que le premier) en mettant en correspondance les parties anatomiques des deux parties du visage en question et en même temps calcule une distance géodésique entre les deux formes qui reflète la longueur du plus court chemin qui séparent ces formes (visage 3D et son miroir). Un champs scalaire de déformation optimales est produit comme illustré dans la figure 1.

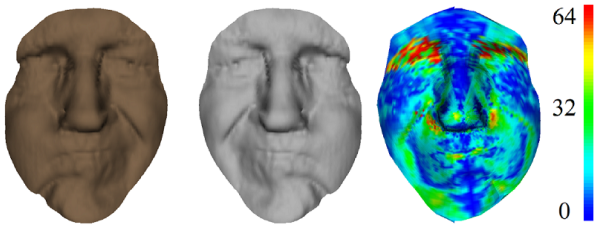


Figure 1 – *Un exemple de champs scalaires de déformations appelé DSF (à droite) calculés entre le visage (à gauche) et son miroir (au centre).*

C'est la comparaison du visage 3D (original) avec son visage miroir (obtenu par réflexion) qui permet de produire les champs scalaires de déformations, représentées ici par une cartographie couleur. Plus la couleur est chaude (rouge) plus l'asymétrie faciale est élevée, et plus s'est froide (vers le bleu) plus l'asymétrie est faible.

5 Expérimentations et discussions

Pour quantifier le niveau d'asymétrie du visage dans le cas d'un paralysie faciale, nous utilisons la méthode qui est décrite dans la section 4 ainsi que le jeu de données dans la section 3.1. Nous avons défini un protocole d'acquisition des séquences 3D, et les séquences sont traitées trame par trame. Pour chaque trame, on détecte le bout du nez, on calcule son miroir, puis on aligne la trame originale avec son miroir grâce à ICP. Et ensuite, le DSF est calculé pour chacune des trames. Dans nos expérimentations, un visage 3D est représenté par 100 courbes radiales, chacune contient 50 points. Par conséquence, les champs scalaires de déformations sont composés de 5000 valeurs chacun.

1. http://www.artec3d.com/fr/hardware/artec-mht/
2. http://www.vtk.org

Concrètement, quatre expressions faciales ont été réalisées : (1) neutre, (2) sourire simple, (3) sourire forcé, (4) hausser les sourcils. De plus, un sujet de contrôle (personne jugé sans aucune paralysie faciale) est utilisé afin de pouvoir comparer et interpréter les résultats . Nous tenons compte de l'état pré-opératoire (sans injection BT) et post-opératoire du patient (avec Injection BT).

1. État neutre – dans cette expérience, nous avons restreint notre étude à un simple visage neutre statique (au repos). Car, il est nécessaire de tenir compte du visage du patient à l'état neutre avant et après l'injection du BT. Pour une meilleure observation, une illustration de l'asymétrie bilatérale obtenue avec l'aide de DSF est donnée dans la figure 2. Dans cette figure (a) avant injection BT et (b) après injection BT, il est possible d'observer des changements mineurs entre l'état pré-opératoire et post-opératoire. En revanche, il existe une grande différence lors de la comparaison avec le visage du sujet de contrôle (c). Nous pourrons noter que même le visage du sujet de contrôle n'est pas parfaitement symétrique. Pour une analyse plus fine, il est important que le patient réalise une expression faciale par exemple un sourire simple.

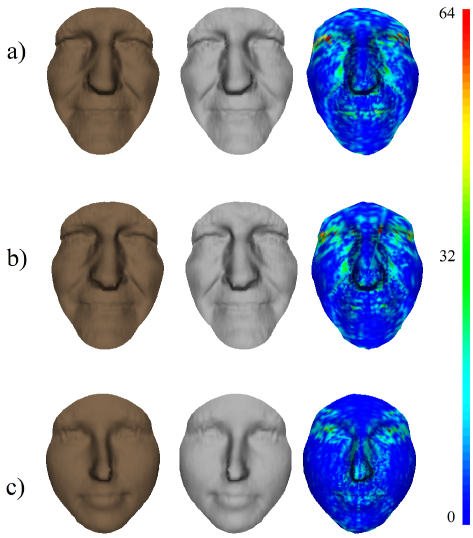


Figure 2 – *DSFs calculés sur un patient en (a) pré-injection puis (b) post-injection et sur (c) le sujet de contrôle.*

2. Sourire normal – dans cette expérience, le clinicien demande aux patients d'effectuer un sourire normal. Dans la figure 3 nous pouvons observer le degré d'asymétrie sur une trame apex (expression maximale) du sourire. Cependant, en comparant les résultats obtenus avec ceux de l'expérience précédente (Figure 2), le DSF permet d'observer une grande différence en terme d'asymétrie, entre le résultat pré-opératoire (a) et le post-opératoire (b). En effet, on peut constater une nette amélioration du niveau d'asymétrie du visage du patient après l'injection du BT. Nous rappelons que ces acquisitions en 4D (3D+t) en post-opératoire sont obtenues deux semaines après l'injection du BT. Le niveau d'asymétrie est évalué dans chaque image de la séquance 3D.

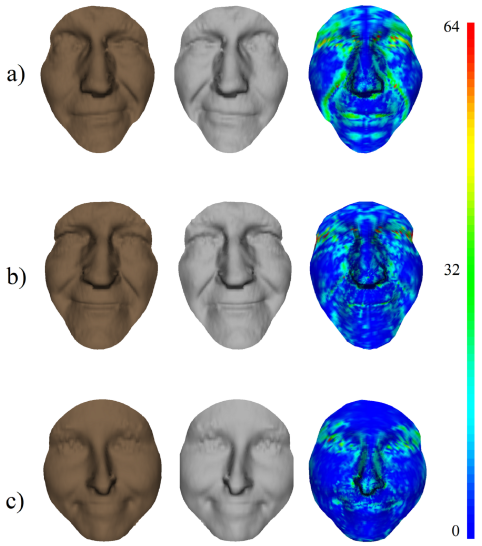


Figure 3 – *Sourire naturel, DSF (a) pré-opératoire, (b) post-opératoire, et (c) sujet de contrôle.*

La figure 4 permet d'observer la distance géodésique, écart global entre la forme générale et son miroir entre et son évolution dans le temps pour les trois expériences. De ce graphique, on voit une diminution du niveau d'asymétrie du visage entre les trames 28 et 60 (courbe bleu) comparée à la courbe rouge (avant injection), ce qui montre clairement l'effet du BT à freiner la sur-activité de la partie saine du visage et donc rendre le visage et son évolution dans le temps symétrique, bien sûr dans le cadre de cette expérience (sourire normal).

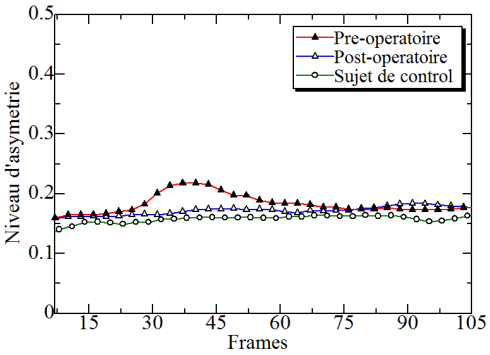


Figure 4 – *Niveau d'asymétrie global du visage dans une séquence 3D dans le cas d'un sourire normal.*

3. Sourire forcé – pour aller plus loin dans nos analyses, le clinicien demande aux patients de réaliser un sourire forcé

ce qui engendre plus de déformations faciales, puis nous avons utilisé la même technique décrite dans l'expérience précédente. En observant la figure 5, il est possible d'établir les mêmes conclusions que dans le cas précédent, car le niveau d'asymétrie du visage est relativement faible après l'injection du BT.

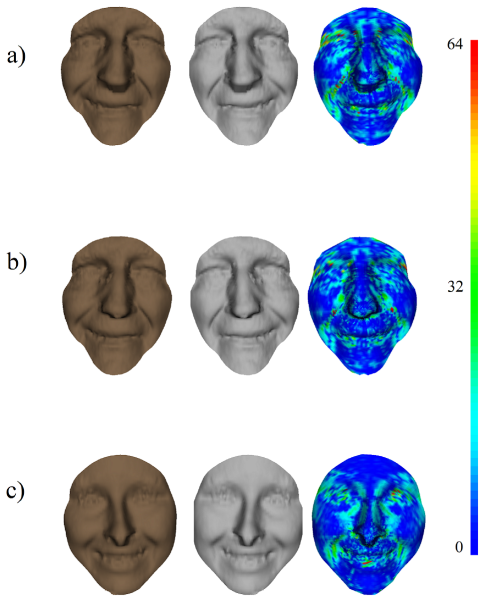


Figure 5 – *Sourir forcé, résultat de l'asymétrie du visage avant BT en (a), après BT (b), et en (c) le sujet de control.*

Dans la figure 6, et pendant l'intervalle de temps entre la trame 14 jusqu'à la fin, le niveau d'asymétrie (pre-opératoire, courbe rouge) est plus élevé que le niveau d'asymétrie (post-opératoire, courbe bleu), tandis que le niveau d'asymétrie du visage du sujet de contrôle (courbe verte) reste très bas et presque stable par rapport aux deux autres courbes.

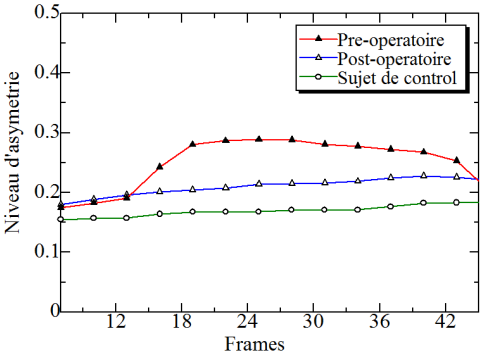


Figure 6 – *Evolution de la symétrie global du visage dans une séquence de trames 3D pour un sourir forcé .*

4. Hausser les sourcils – dans cette expérience, il est demandé aux patients de hausser les sourcils. En observant la figure 7, il est possible de dire que le niveau d'asymétrie en mode pre-opératoire et post-opératoire sont comparables. Ceci peut être expliqué par le fait que l'injection du BT n'est concentrée que sur les muscle zygomatiques et non sur les muscles qui contrôle le mouvement des sourcils.

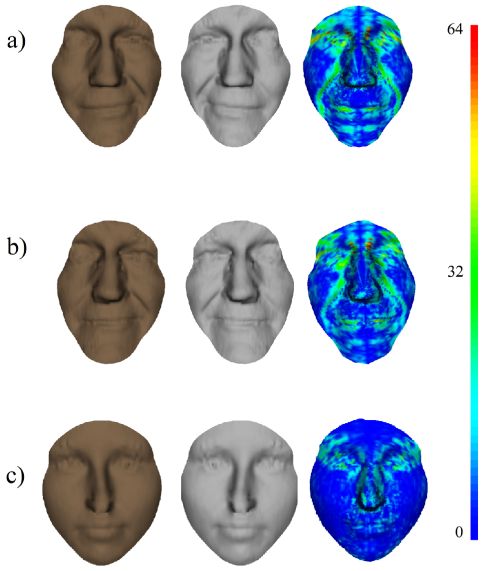


Figure 7 – *Hausser les sourcils, résultats de l'asymétrie avant et après injection du BT.*

De plus la figure 8 confirme les résultats, car dans les deux courbes (rouge et bleu) le niveau d'asymétrie est toujours élevé tandis que celui du sujet de contrôle est très bas et stable.

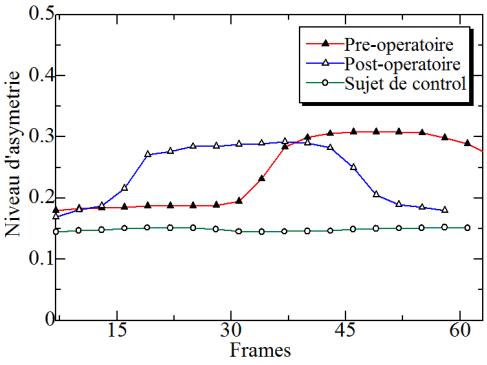


Figure 8 – *Evolution dans le temps du niveau de symétrie global dans le cas de hausser les sourcils.*

6 Conclusion

Dans ce papier, nous avons proposé une nouvelle technique permettant de quantifier le niveau de l'asymétrie faciale à partir des données 4D. L'ingrédient principal de cette

approche est le calcul de champs scalaires de déformations (DSF) [9]. L'approche permet de comparer un visage donné à sa réflexion et de réaliser un recalage point-à-point afin de mesurer avec précision l'asymétrie du visage et l'étendre dans le temps. De ce fait, une base de données de cinq patients a été recueillie dans des conditions cliniques. Nous avons démontré avec des séquences 3D dynamiques, l'utilité de la méthode proposée. En particulier, la comparaison de l'asymétrie faciale en utilisant les champs scalaires de déformations (DSF) avant et après l'injection du BT montre que l'approche proposée est prometteuse. Ce travail propose ainsi un outil d'évaluation quantitative pour les cliniciens ne nécessitant aucun repère anatomique (landmarks) et qui prend en compte l'aspect temporel de la dynamique faciale.

Références

[1] RP. Clark et CE Berris. Botulinum toxin : a treatment for facial asymmetry caused by facial nerve paralysis. *Plast Reconstr Surg*, 84(2) :353—5, 1989.

[2] R Filipo, I Spahiu, E Covelli, M Nicastri, et GA Bertoli. Botulinum toxin in the treatment of facial synkinesis and hyperkinesis. *The Laryngoscope*, 122(2) :266–70, 2012.

[3] Terzis JK et Noah ME. Analysis of 100 cases of free-muscle transplantation for facial paralysis. *Plast Reconstr Surg*, 99(7) :1905–21, 2013.

[4] F.M. Sukno, M.A. Rojas, J.L. Waddington, et P.F. Whelan. On the quantitative analysis of craniofacial asymmetry in 3d. Dans *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8, May 2015.

[5] T. Al-Anezi, B. Khambay, M.J. Peng, E. O'Leary, X. Ju, et A. Ayoub. A new method for automatic tracking of facial landmarks in 3d motion captured images (4d). *International Journal of Oral and Maxillofacial Surgery*, 42(1) :9 – 18, 2013.

[6] Shiyang Cheng, Ioannis Marras, Stefanos Zafeiriou, et Maja Pantic. Active nonrigid ICP algorithm. Dans *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*, pages 1–8, 2015.

[7] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, et Daniel Rueckert. Recognition of 3D facial expression dynamics. *Image and Vision Computing*, 30(10) :762–773, 2012.

[8] Tianhong Fang, Xi Zhao, Omar Ocegueda, Shishir K. Shah, et Ioannis A. Kakadiaris. 3d/4d facial expression analysis : An advanced annotated face model approach. *Image and Vision Computing*, 30(10) :738–749, 2012.

[9] Boulbaba Ben Amor, Hassen Drira, Stefano Berretti, Mohamed Daoudi, et Anuj Srivastava. 4-d facial expression recognition by learning geometric deformations. *IEEE T. Cybernetics*, 44(12) :2443–2457, 2014.

[10] Yi Sun et Lijun Yin. Facial expression recognition based on 3d dynamic range model sequences. Dans *Proceedings of the 10th European Conference on Computer Vision : Part II*, ECCV '08, pages 58–71, 2008.

[11] Baiqiang Xia, Boulbaba Ben Amor, Hassen Drira, Mohamed Daoudi, et Lahoucine Ballihi. Combining face averageness and symmetry for 3d-based gender classification. *Pattern Recognition*, 48(3) :746–758, 2015.

[12] Wei Quan, Bogdan J. Matuszewski, et Lik-Kwan Shark. Facial asymmetry analysis based on 3-d dynamic scans. Dans *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, SMC 2012, Seoul, Korea (South), October 14-17, 2012*, pages 2676–2681, 2012.

Évaluation de la qualité audiovisuelle des applications de visiophonie

Ines Saidi [1,2] Lu Zhang [2] Olivier Deforges [2] Vincent Barriac [1]

[1] Orange Labs-SVQ/MOV, 2 avenue Pierre Marzin, F-22307 Lannion Cedex
[2] IETR, CNRS UMR 6164, INSA Rennes, 20 avenue des Buttes de Coesmes, 35708 Rennes, France

Résumé

Les services de télécommunication évoluent en intégrant de nouvelles modalités au-delà du simple échange vocal bien connu de tous. Ainsi, voit-on de plus en plus de services mêlant le son, le texte et la vidéo. C'est le cas des applications de visiophonie ou de vidéoconférence qui permettent de combiner la voix, la vidéo et le partage d'environnements de plus en plus complexes. Dans ce contexte s'inscrit notre thèse intitulée « Analyse et modélisation de la qualité perçue des applications de visiophonie ». Nous présenterons dans cet article un résumé du sujet de la thèse en mettant en évidence ses principaux axes et objectifs. En outre, nous fournirons les résultats du premier test subjectif d'évaluation de la qualité audiovisuelle que nous avons mené.

Mots clefs

Vidéoconférence, qualité audiovisuelle, qualité de service, dégradations réseau, test subjectif.

1 Introduction

Pour les services de téléphonie, des décennies de travaux de recherche, de normalisation et d'exploitation de réseaux ont permis de déterminer des métriques représentatives de la qualité perçue par l'utilisateur final (retard, qualité audio, écho, bruit, pertes d'information, etc.) [1]. Ces travaux ont permis également de développer des outils automatiques permettant de connaître la performance du réseau et son impact sur la qualité de bout en bout [2]. Néanmoins, les opérateurs de télécommunications sont en revanche démunis dès lors qu'il s'agit d'assurer la supervision de ces nouveaux services de visiophonie. En effet, il manque le recul de l'expérience pour déterminer les bonnes métriques représentatives et les seuils associés pour juger l'acceptabilité de la qualité d'un service [3, 4]. Or, la valeur ajoutée des services offerts par ces opérateurs réside en grande partie dans le fait qu'ils sont à qualité garantie. C'est pourquoi le développement de méthodes adéquates pour la mesure et la supervision de la qualité perçue des services conversationnels audiovisuels devient un défi majeur pour les opérateurs de télécommunications. Au-delà de la complexité d'accès aux données (séparation des flux de signalisation et de transport des données temps réel, sécurisation par cryptage des données) se pose la question de la pertinence des indicateurs réseau pour représenter les dégradations perçues par l'utilisateur final. Dans les outils de mesures relatifs à la téléphonie, une longue expérience a permis de voir l'intégration de métriques fiabilisées et maîtrisées par les équipes opérationnelles, à même d'aider au diagnostic et à la correction des problèmes. La situation n'est absolument pas la même pour les services conversationnels audiovisuels, et ce pour plusieurs raisons :

– Usage encore peu répandu en dehors du monde de l'Internet ;
– Complexité technique de la mesure de qualité vidéo. Le médium vidéo est autrement plus complexe que le son. Son codage et sa transmission dans les réseaux IP obéissent à des règles très particulières et beaucoup moins déterministes avec une dépendance au contenu [5] ;
– Multiplicité et complexité des terminaux et écrans utilisés (PC, smartphone, TV, etc.) ;

Or, les besoins opérationnels commencent à poindre. Les fabricants d'outils de test proposent depuis plusieurs années déjà des solutions dédiées à la supervision des services de streaming audiovisuels [6, 7] et tentent de les adapter à la problématique des services conversationnels [8]. Les difficultés techniques mentionnées ci-dessus font que, la plupart du temps, ces outils sont spécialisés sur un service disponible avec un format d'image donnée et sur un modèle de terminal donné. Cette thèse se focalise sur la modélisation de la qualité perçue de ces nouveaux services à partir de l'analyse approfondie d'informations techniques collectées au niveau des terminaux ainsi que des réseaux.

2 Objectifs de la thèse

L'objectif de la thèse est d'étudier et de proposer des métriques représentatives de la perception de la qualité des flux associés aux services de visiophonie et de vidéoconférence. Ces métriques seront à déterminer à partir d'informations issues du signal audiovisuel, mais aussi d'éléments d'analyse du fonctionnement du service. Ces données sont accessibles au niveau du terminal ou des équipements réseau. Pour mener à bien ces études, il faudra au préalable s'assurer un accès à l'information détaillée sur la perception client de ces services et des dégradations qui les caractérisent à travers :

– Des méthodes de tests subjectifs ;
– Des méthodes de collecte d'informations liées

aux usages.

Donc, en court terme, nous menons des tests subjectifs afin d'avoir des notes données par les utilisateurs des services sur la qualité perçue. Par la suite, nous appliquerons des calculs statistiques de corrélation avec des mesures objectives.

3 Expérience

La thèse s'appuie sur des méthodes dites "offline", pour recueillir et analyser des données d'utilisation et de perception des utilisateurs. L'une de ces méthodes c'est les tests subjectifs. Nous avons réalisé un test non conversationnel à travers duquel nous étudions l'influence des dégradations réseau (perte de paquets, gigue et retard) sur la qualité perçue d'un contenu audiovisuel. D'autre part, nous mettons à jour les limites d'acceptabilité de la désynchronisation audio-vidéo dans le contenu multimédia.

3.1 Plateforme de simulation

Afin de générer notre base des séquences de test, nous avons utilisé un logiciel de vidéoconférence appelé "eConf". Ce logiciel permet de partager des contenus multimédias entre deux utilisateurs. Avec "eConf" nous somme capable et de séparer les flux IP audio et vidéo ce qui permet d'appliquer les dégradations audio et vidéo indépendamment. Le protocole de communication adopté dans cette plateforme est le H.232. Pour simuler des dégradations réseau, nous avons utilisé le logiciel "NetDisturb" qui permet de perturber les flux sur le réseau IP. Nous avons inséré une machine équipée de NetDisturb entre nos deux PC "eConf" connectés sur un réseau local Ethernet. Une fois un appel est établi, l'appelant transmet les fichiers sources audio, vidéo et audiovisuels au destinataire (voir Fig. 1). Ensuite, nous contrôlons la transmission des paquets IP entre les deux PC en ajoutant du perte de paquet, de la gigue et du retard. Au niveau de la réception nous enregistrons les séquences dégradées et nous prenons des captures du flux IP (format pcap avec Wireshark). Pour assurer une lecture parfaite, toutes les séquences multimédias sont enregistrées sous le format YUV 4 :2 :0 pour le flux vidéo et sous le format non compressé Pulse Code Modulation (PCM) pour le flux audio.

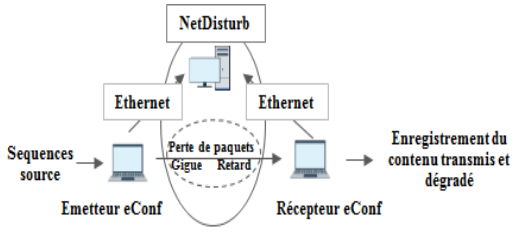


Figure 1 – *Conception de la plateforme de simulation.*

3.2 Séquences sources

Six séquences audiovisuelles de durée comprise entre 8 et 10 secondes définissant différents types de contenus ont été utilisées dans l'étude. La résolution des séquences sources est VGA (640×480) avec un taux de 15 images par seconde. Avant de simuler les dégradations réseau, le flux vidéo a été codé en H.264 à un débit binaire égal à 768 kbps. Le flux audio a été codé en AMR Wideband à 23,85 kbps. Ces codecs et débits ont été choisis vu leur utilisation fréquente dans les systèmes de vidéoconférence et vu la bonne qualité perçue qu'ils assurent.

3.3 Types de dégradation

Les dégradations que nous avons simulées reflètent les types de dégradation que peut subir un appel vidéo sur réseau IP. Le niveau de distorsion a été modifié pour générer des contenus multimédias à une large gamme de qualités. L'application de différents retards pour les flux audio et vidéo génère une désynchronisation entre eux. Une valeur négative de la désynchronisation signifie que le flux audio est retardé et une valeur positive signifie qu'il est avancé.

Perte de paquet vidéo VPL (%)	0, 0.5, 1, 2
Perte de paquet audio APL (%)	0, 2, 5, 20
Gigue audio (ms)	0, 60
Gigue vidéo (ms)	0, 30
Désynchronisation audio-vidéo (ms)	-400, -250, -150, 0, +50, +150, +400

Tableau 1 – *Paramètres expérimentaux*

3.4 Méthode de Test

Notre test a été organisé sous trois sessions séparées comme c'est détaillé dans le tableau 2. La méthode expérimentale que nous utilisons est l'évaluation par catégorie absolue (ACR). Le protocole de test proposé est basé sur les recommandations de l'UIT-T P.800 [9], l'UIT-T P.910 [10] et de l'UIT-T P.911 [11] pour le test audio seulement, le test vidéo seulement et le test audiovisuel respectivement. 30 sujets (13 hommes, 17 femmes) ont participé à l'expérimentation. Nous avons réalisé le test audio et vidéo dans la même session avec 15 sujets alors que le test audiovisuel a été réalisé avec les 15 autres sujets. Pour mesurer la qualité perçue nous utilisons une échelle MOS (note moyenne d'appréciation) à cinq niveaux : 5-Excellente , 4-Bonne , 3-Moyenne, 2-Médiocre, 1-Mauvaise. Pour la synchronisation nous utilisons une échelle MOS de dégradation à cinq niveaux [12] 5-Imperceptible, 4-Perceptible mais pas gênant, 3-Légèrement gênant, 2-Gênant, 1-Très gênant.

Test	Durée	Séquences	Conditions	Sorties
Audio	10min	36	5	MOS_A
Vidéo	10min	36	5	MOS_V
Audiovisuel	1h30	176	33	MOS_{AV} MOS_A^{AV} MOS_V^{AV} MOS_{synch}

Tableau 2 – *Organisation du test*

4 Résultats et Discussion

4.1 Impact des qualités audio et vidéo sur la qualité globale

Les résultats présentés dans la figure 2 sont la moyenne sur les six contenus. Une analyse des courbes révèle que pour un niveau de qualité audio donné, la décroissance de la qualité vidéo se traduit généralement par des notes audiovisuelles inférieures. L'impact des pertes de paquet vidéo sur la qualité audiovisuelle est plus important dans le cas d'une bonne qualité audio (0% et 2%APL) que dans le cas d'une médiocre et mauvaise qualité audio (5% et 20% APL). D'autre part, nous notons que, généralement, la qualité vidéo influe sur la qualité audiovisuelle globale plus que la qualité audio comme révélée dans [13, 14]. La corrélation linéaire entre MOS_V et MOS_{AV} est égale à 77% tandis que la corrélation linéaire entre MOS_A et MOS_{AV} est égale à 57%. De ces résultats on peut retenir que des dégradations visuelles influencent la qualité d'une communication vidéo même ci l'audio est de bonne qualité. Donc, les processus de correction intégrés dans les terminaux ou au niveau des réseaux doivent assurer une bonne configuration vidéo (frame rate, débit, résolution).

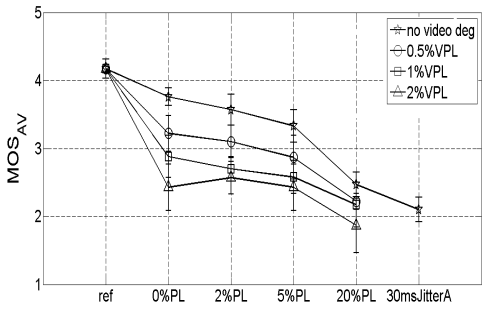


Figure 2 – *Impact de la qualité audio et vidéo sur la qualité audiovisuelle globale.*

4.2 Modèle de qualité audiovisuelle

Afin d'étudier l'influence des qualités audio et vidéo sur la qualité audiovisuelle, nous avons appliqué un modèle de régression linéaire. Le modèle général proposé dans d'autres études [15, 16] est le suivant :

$$MOS_{AV} = \alpha_0 + \alpha_1 MOS_A^{AV} + \alpha_2 MOS_V^{AV} + \alpha_3 MOS_A^{AV} \cdot MOS_V^{AV} \quad (1)$$

Nous avons calculé la corrélation linéaire R^2 entre le modèle additif ($MOS_A + MOS_V$), le modèle multiplicatif ($MOS_A \times MOS_V$) et le modèle général (Eq.1) et la note MOS_{AV}. Les résultats montrent que le modèle multiplicatif fournit ma meilleure corrélation égale à 95,8% et une erreur quadratique moyenne (RMSE) égal à 0,14. Ainsi, le modèle que nous proposons est le suivant :

$$MOS_{AV} = 1.277 + 0.174 MOS_A^{AV} \cdot MOS_V^{AV} \quad (2)$$

Pour comparaison, d'autres études ont suggéré des modèles multiplicatifs [11, 17] mais chacun d'eux était basé sur des contenus synchronisés. Cependant, le modèle que nous proposons ici tient compte de la désynchronisation et basé sur des dégradations réseaux. Ceci explique la différence entre les coefficients α_0 et α_3.

4.3 Synchronisation audio-vidéo

D'après les résultats du test, nous avons identifié les seuils d'acceptabilité de désynchronisation entre l'image et le son dans un contenu audiovisuel. Nous mettons le score MOS_{synch} égal à 4 comme la limite d'acceptabilité de désynchronisation. La figure 3 montre que les sujets commencent à être gênés lorsque l'audio est retardé de plus de 250 ms et lorsqu'il est avancé de plus de 150 ms. Ces résultats sont accord avec les limites reportés pour l'IPTV [18, 19, 20] et montrent que la résolution de l'écran n'a pas d'influence sur la perception de la synchronisation. En outre, la présence des pertes de paquets vidéo (et dans une moindre mesure audio), a un faible impact, mais pas significatif sur la perception de la synchronisation. Limités par le nombre de conditions et la durée du test, nous ne pouvons pas croiser les 6 valeurs différentes de désynchronisation avec tous les niveaux de dégradation réseau. Ceci explique l'absence de certains points sur la figure 3. En outre, la présence des pertes de paquets vidéo (et dans une moindre mesure, audio), a un impact faible mais non significatif sur la perception de la synchronisation.

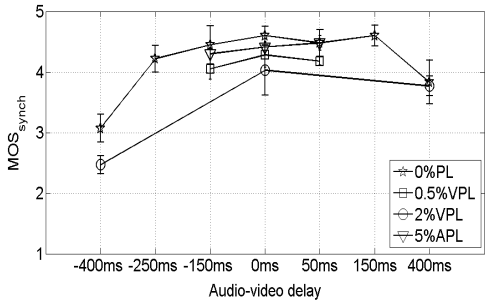


Figure 3 – *Limites d'acceptabilité de la désynchronisation*

5 Conclusion

Dans cet article, nous avons présenté le sujet de la thèse et les principaux travaux que nous avons menés. Les résultats montrent qu'à la fois les qualités audio et vidéo contribuent à la qualité audiovisuelle globale avec une domination générale de la qualité vidéo. Notre prochaine étape consiste à effectuer des mesures objectives de qualité sur nos séquences de test en vue de trouver une corrélation entre les mesures objectives et les résultats des tests subjectifs. Nous prévoyons également d'étudier davantage les questions de qualité audiovisuelle et de désynchronisation dans des contextes conversationnels.

Références

[1] ETSI Guide 202 057-2. Speech processing transmission and quality aspects (stq) : user related qos parameter definitions and measurements ; part2 : Voice telephony, group 3 fax, modem data services and sms, February 2009.

[2] Antony W Rix, John G Beerends, Michael P Hollier, et Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. Dans *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 749–752. IEEE, 2001.

[3] Quan Huynh-Thu et Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13) :800–801, 2008.

[4] Stefan Winkler et Praveen Mohandas. The evolution of video quality measurement : from psnr to hybrid metrics. *Broadcasting, IEEE Transactions on*, 54(3) :660–668, 2008.

[5] Recommendation ITU-T H.264. Advanced video coding for generic audiovisualservices. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*, February 2016.

[6] Michal Ries, Olivia Nemethova, et Markus Rupp. Video quality estimation for mobile h. 264/avc video streaming. *Journal of Communications*, 3(1) :41–50, 2008.

[7] Kandaraj Piamrat, Cesar Viho, Jean-Marie Bonnin, et Adlen Ksentini. Quality of experience measurements for video streaming over wireless networks. Dans *Information Technology : New Generations, 2009. ITNG'09. Sixth International Conference on*, pages 1184–1189. IEEE, 2009.

[8] Akira Takahashi, Atsuko Kurashima, et Hideaki Yoshino. Objective assessment methodology for estimating conversational quality in voip. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6) :1984–1993, 2006.

[9] Recommendation ITU-T P.800. Methods for subjective determination of transmission quality. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*, August 1996.

[10] Recommendation ITU-T P.910. Subjective video quality assessment methods for multimedia applications. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*, April 2008.

[11] Recommendation ITU-T P.911. Subjective audiovisual quality assessment methods for multimedia applications. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*, Dec. 1998.

[12] Recommendation ITU-R BT-500.13. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*, 2012.

[13] Junyong You, Ulrich Reiter, Miska M Hannuksela, Moncef Gabbouj, et Andrew Perkis. Perceptual-based quality assessment for audio–visual services : A survey. *Signal Processing : Image Communication*, 25(7) :482–501, 2010.

[14] JG Beerends et FE de Caluwe. Relations between audio, video and audiovisual quality. *Contr COM*, pages 12–19, 1997.

[15] Mike P Hollier, Andrew N Rimell, David S Hands, et Rupert M Voelcker. Multi-modal perception. *BT Technology Journal*, 17(1) :35–46, 1999.

[16] Stefan Winkler et Christof Faller. Perceived audiovisual quality of low-bitrate multimedia content. *Multimedia, IEEE Transactions on*, 8(5) :973–980, 2006.

[17] Benjamin Belmudez et Sebastian Möller. Audiovisual quality integration for interactive communications. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1) :1–23, 2013.

[18] T. Hayashi K. Yamagishi et A. Takahashi. Planning model for audiovisual communication services. *NTT Technical Review*, 7(4), Apr 2009.

[19] N. Egi K. Yamagishi et T. Tominaga. Monitoring the quality of iptv services. *NTT Technical Review*, 11(5), May 2013.

[20] Dominic W Massaro, Michael M Cohen, et Paula MT Smeele. Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America*, 100(3) :1777–1786, 1996.

Réduction de bruit multiplicatif dans les images ultrasons basée sur la Décomposition Multiplicative Multiresolution (MMD)

M.Outtas[1] A.Serir[1] O.Deforges[2] L.Zhang[2]

[1] USTHB-LTIR, Faculté d'Electronique et Informatique, Alger , ALGERIE

[2] IETR-INSA, UEB, UMR 6164 Rennes, 35708, FRANCE

Résumé

Cet article traite du débruitage d'images medicales ultrasons en utilisant une décomposition multiplicative multirésolution **MMD** de l'image. La méthode proposée est basée sur un seuillage des coefficients MMD. Les seuils sont automatiquement évalués à partir du contenu de la l'image transformée. La méthode proposée est testée sur des images bruitées de manière synthétique et sur des images médicales réelles ultrasons (échographies). Nous utilisons le PSNR et la NIQE comme métriques de qualité pour l'évaluation des résultats.

Mots clefs

Image ultrasons, échographie, Speckle, Bruit multiplicatif, MMD.

1 Introduction

L'ultrasonographie est l'une des techniques les plus répandues en imagerie médicale. Cependant cette technique présente un inconvénient majeur : la faible qualité des images, avec un caractère granuleux, due au bruit de speckle. Ce type de bruit, qui introduit des distortions de manière multiplicative, est lié au phénomène d'interférence entre les ondes incidentes et les ondes réfléchies [1]. De type hautes fréquences, **le speckle** fait disparaître les limites entre les régions représentatives de l'image ; ce qui peut rendre les tâches (détection de détails, mesure de lésions,...) difficiles pour les praticiens. Pour ces raisons, les méthodes de filtrage présentent un intérêt particulier pour les images ultrasons. Il existe trois grandes catégories pour le filtrage des images ultrasons : les méthodes de filtrage spatial, les méthodes de filtrage multi-échelle et les méthodes de filtrage anisotrope. Parmi les méthodes de filtrage spatial, nous pouvons citer les plus fréquentes à savoir : le filtre median [2], le filtre de Wiener, le filtre de lee [3], le filtre de frost[4] ou encore les filtres à moyennes non-locales comme le filtre (OBNLM) optimized Bayesian NL-means avec séléction de blocs proposées par Coupé [5].

Dans [6], un algorithme de réduction de speckle, basé sur un filtrage anisotrope, est proposé. La catégorie la plus importante d'algorithmes de réduction de speckle, qui utilisent des méthodes multi-échelles "la plus importante" est principalement basée sur la décomposition en ondelette. Les coefficients d'ondelettes sont alors débruités en leur appliquant un seuillage unique. Cette méthode appelée communément "visushrink" a été introduite par Donoho [7].

Etant donné la nature multiplicative du speckle, des auteurs proposent un filtrage homomorphique. Pour cela, on applique à l'image un opérateur non-linéaire comme le logarithme afin de transformer le bruit multiplicatif en bruit additif. L'image ainsi obtenue est traitée comme étant simplement distordue par du bruit additif [8]. Jain [9] introduit l'approche homomorphique pour la réduction du speckle en utilisant le filtre de Wiener. La méthode"visushrink" a elle aussi été appliquée par une approche homomorphique dans [10].

Même si beaucoup de travaux pertinents sur la réduction du speckle ont été réalisés, des améliorations sont néanmoins possibles. Dans cet article, nous proposons une nouvelle méthode d'analyse et de réduction de speckle dans les images ultrasons en exploitant la nature multiplicative du bruit considéré. La décomposition multiplicative multirésolution proposée par Serir et Belouchrani[11] appelée MMD sera utilisée à cet effet. Les coefficients de la MMD sont seuillés avec une valeur de seuil obtenue automatiquement à partir de l'analyse du speckle. La méthode proposée est testée sur des images bruitées de manière synthétique et sur des images médicales réelles ultrasons (échographies). Pour l'évaluation des résultats, nous utilisons comme métriques de qualité deux metriques : l'une avec référence (PSNR) et la seconde sans référence NIQE[12].

Le papier est organisé comme suit :la partie 2 présente brièvement la méthode de décomposition choisie. La partie3 traite de la méthode proposée. Enfin la partie 4 est dédiée aux résultats expérimentaux.

2 La décomposition Multiplicative Multirésolution

La décomposition Multiplicative Multirésolution (MMD) est une décomposition multi-échelle non linéaire [11]. 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 bruitées par du bruit multiplicatif. Les figures 1 et 2 illustrent la Décomposition Multiplicative Multirésolution (MMD) d'une image $I(m,n)$.
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 $I(n,m)$ sont définis par :

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

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

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

$h_{12} = \alpha h_{11}, h_{21} = \nu h_{11}, h_{22} = \gamma h_{11}$
Où α, ν et γ sont des scalaires positifs.
Ainsi, les filtres polyphases linéaires $h_{i,j}$ et $f_{i,j}$ sont définis par :

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

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

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

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

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

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

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 :

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

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

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

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

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 conrrespondent 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 :

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

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 [11] et [13].

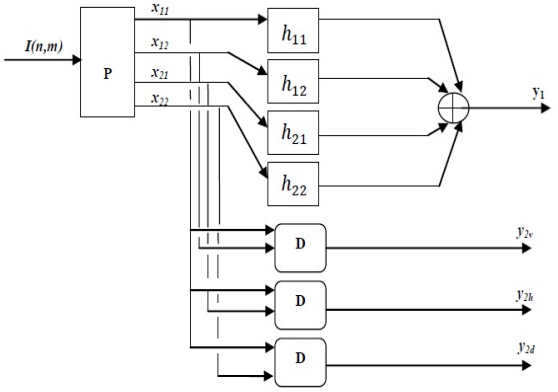


Figure 1 – *Schéma d'Analyse de la décomposition multiplicative multirésolution MMD*

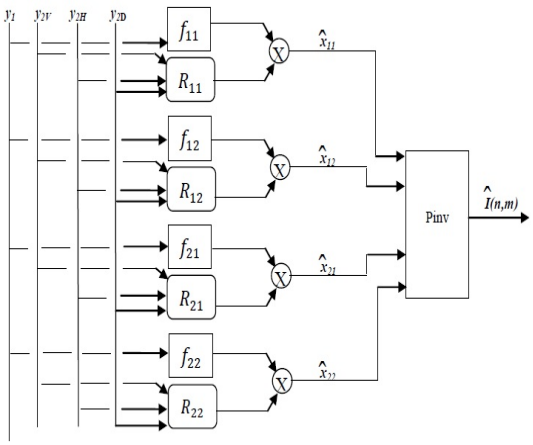


Figure 2 – *Schéma de synthèse de la décomposition multiplicative multirésolution MMD*

3 Méthode proposée

Dans ce papier, nous proposons une méthode de débruitage des images ultrasons. Les coefficients $y_2{}^j$ de l'image obtenues dans le domaine MMD sont seuillés. Le seuil est déterminé de façon automatique, et ce en effectuant une analyse du bruit dans le domaine MMD. La variance du bruit est extraite à partir d'une zone homogène de l'image. Une fenêtre glissante Ω parcours les images détails à la recherche d'une région homogène Ω_{Smooth} (dont les coefficients MMD sont proche de β). Le choix de la zone est basé sur la minimisation de la déviation médiane absolue DMA.

$$DMA = \frac{1}{n}\sum_{i=1}^{n}\left|y_2^{(1)}(i) - \beta\right|_{(i\in\Omega)} \tag{13}$$

$$\begin{cases} \sigma_h = \sqrt{\frac{1}{n}\left(y_{2h}^{(1)}(i) - \overline{y_{2h}^{(1)}}\right)^2} \\ \sigma_v = \sqrt{\frac{1}{n}\left(y_{2v}^{(1)}(i) - \overline{y_{2v}^{(1)}}\right)^2} \quad (i \in \Omega_{smooth}) \\ \sigma_d = \sqrt{\frac{1}{n}\left(y_{2d}^{(1)}(i) - \overline{y_{2d}^{(1)}}\right)^2} \end{cases} \tag{14}$$

On considère un seuil $0 < t_j < 1$, $t_j = C_n^j$ où C_n^j représente l'écart type normalisé du bruit (rapport de l'écart type à la moyenne).
Le seuillage faible "soft thresholding" dans le cas multiplicatif se fait comme suit :

$$\begin{cases} w_t^j = min\left\{\beta, w^j + t_j\right\} & si \quad w^j \leq \beta \\ w_t^j = min\left\{\beta, w^j - t_j\right\} & autrement \end{cases} \quad (i \in \Omega_{Smooth}) \tag{15}$$

où w^j représente les coefficients MMD des images détails y_2v, y_2h et y_2d à la résolution j, et w_t^j sont les coefficients MMD **seuillés** des images détails à la résolution j. Cependant, pour augmenter l'efficacité de notre filtre, il est nécessaire d'effectuer un seuillage plus fin. Lorsque $\left(1 - C_n^j\right)/4 < w_t^j < \left(1 + C_n^j\right)/4$, le coefficient représente une région homogène **bruitée** et le coefficient est alors ramené à β. Dans le cas contraire le coefficient w^j reflète un pixel transition. Afin de réduire le bruit présent dans ce pixel transition ou région hétérogène, on se propose d'ajouter ou de soustraire un offset comme suit :

$$\begin{cases} w_r^j = \nu w_t^j + \gamma si w^j \geq \frac{1+C_n^j}{4} \\ w_r^j = \nu w_t^j - \gamma si w^j \geq \frac{1-C_n^j}{4} \end{cases} \quad (i \in \Omega_{smooth}) \tag{16}$$

avec $\nu = \frac{1}{\sqrt{1+{c_n^j}^2}}$ et $\gamma = 1 - \frac{1}{\sqrt{1+{c_n^j}^2}}$.

ν, γ et w_r^j sont les coefficients MMD à partir desquels l'image sera reconstruite par synthèses successives de la résolution j à la résolution 1, donnant au final l'image débruitée.

4 Experimentations et résultats

Les expérimentations ont été conduites sous MATLAB en utilisant des images synthétiques comprenant du bruit type "speckle", et sur une image médicale ultrason de métastase hépatique. L'image synthétiquement bruitée est obtenue en distordant l'image Lena (512×512) par un bruit type speckle. Pour cela, nous utilisons la fonction MATLAB *imnoise*. L'image bruitée finale est la moyenne de trois images bruitées avec un même écart type.

(a) (b) (c) (d) (e) (f)

Figure 3 – *Performance des différentes méthodes de débruitage sur image Lena 3(a) Image Lena entachée de bruit d'écart type 0.2, 3(b)Wiener homomorphique , 3(c) Visushrink, 3(d) Median,3(e) Visushrink homomorphique,3(f) Méthode proposée*

La relation liant l'écart type normalisé C_n^j à différentes résolutions j est obtenue de façon expérimentale. Nous avons choisi une image de synthèse uniforme et avons calculé les C_n^j à travers les résolutions. Ces tests ont été confirmés sur une partie homogène d'une image bruitée. L'analyse des C_n^j à travers les résolutions, montre que C_n^j peut être approximé à partir de la connaissance de C_n^{j-1} : $C_n^j = 2C_n^{j-1}/3$.
Les résultats de la méthode proposée sont comparés avec le filtre homomorphique de Wiener [9], "Visushrink"[7], homomorphique"Visushrink" et ONBLM [5].
Pour évaluer les différentes méthodes de débruitage nous avons utilisé 2 critères :

— Pour les images synthétiques entachées de bruit : Le Peak Signal to Noise Ratio (PSNR)
— Pour les images médicales ultrasons : La Natural Image Quality Evaluator (NIQE)[12]. Cette métrique de qualité "aveugle" ne requiert pas d'image référence (Sans bruit), et il a été montré que la NIQE peut être utilisée pour évaluer la qualité des images ultrasons [14]. Pour la NIQE, une valeur supérieure indique une moins bonne qualité de l'image.

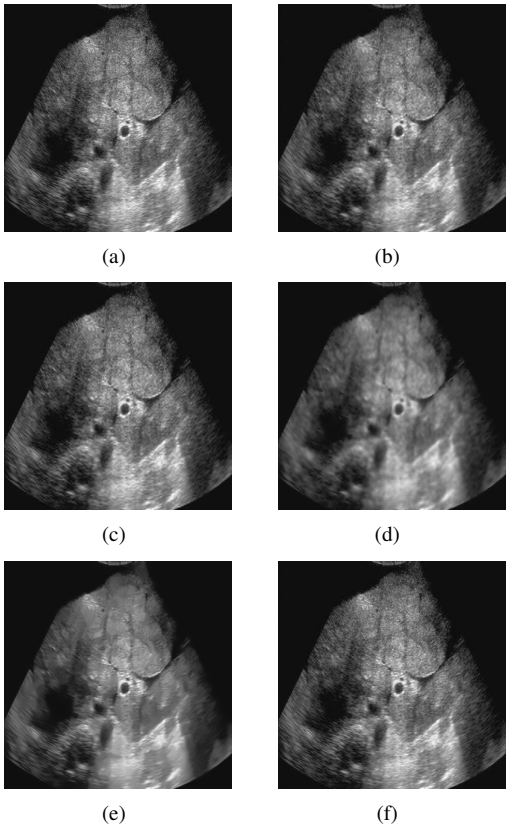


Figure 4 – *Méthodes de dé-bruitage appliquée à l'image échographique 4(a) Image originale, 4(b) Visushrink , 3(c) Visushrink, 4(c) Visushrink homomorphique,4(d) Wiener homomorphique ,4(e) OBNLM,4(f) Méthode proposée*

Le tableau 1 compare les performances en terme de PSNR des différentes méthodes de débruitage. Nous constatons que la méthode proposée est plus adaptée pour réduire le bruit multiplicatif. Dans le tableau 2, nous avons répertorié les scores obtenus par la métrique NIQE sur les images débruitées par différents algorithmes. La méthode proposée réalise une meilleure réduction de speckle sur les images échographiques. La figure 4 montre les images débruitées par ces différents algorithmes.

5 Conclusion

Dans cet article, nous proposons une nouvelle méthode de réduction de bruit "speckle". A travers la Décomposition

Ecart type	Median	Wiener Homo-mor-phique	Visu shrink	Visushrink Homo-mor-phique	Méthode proposé
0.2	29.60	29.41	29.16	29.78	30. 84
0.3	26.09	27.70	26.17	26.67	27.95
0.4	25.03	25.59	24.34	24.86	25.76
0.7	20.46	18.91	19.63	20.37	20.50

Tableau 1 – *Comparaison PSNR des images obtenues par différentes méthodes de débruitage sur l'image Lena contaminée de bruit pour différents écart type*

Méthode de dé-bruitage	NIQE
Wiener Homomorphique	6.97
Visushrink	6.83
Visushrink Homomorphique	6.21
OBNLM	6.90
Méthode proposé	6.19

Tableau 2 – *Comparaison NIQE des images obtenues par diffrentes méthodes de débruitage sur l'image échographique*

Multiplicative Multiresolution, les paramètres du speckle sont extraits automatiquement à partir de l'image bruitée. Les résultats obtenus montrent que la méthode proposée est plus performante que les méthodes de l'état de l'art considérées à savoir : Visushrink, Visushrink homomorphique, Wiener homomorphique et OBNLM.
Comme perspective, nous envisageons de caractériser dans des images échographiques, les structures les plus importantes dans l'établissement d'un diagnostic medical.

Références

[1] Alejandro César Frery, Hans-Jürgen Müller, Corina da Costa Freitas Yanasse, et Sidnei João Siqueira Sant'Anna. A model for extremely heterogeneous clutter. *IEEE T. Geoscience and Remote Sensing*, 35(3) :648–659, 1997.

[2] E. Ritenour, T. Nelson, et U. Raff. Applications of the median filter to digital radiographic images. Dans *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, volume 9, pages 251–254, Mar 1984.

[3] Jong-Sen Lee. Speckle analysis and smoothing of synthetic aperture radar images. *Computer graphics and image processing*, 17(1) :24–32, 1981.

[4] Victor S Frost, Josephine Abbott Stiles, K Sam Shanmugan, et J Holtzman. A model for radar images and its application to adaptive digital filtering of multiplicative noise. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2) :157–166, 1982.

[5] Pierrick Coupé, Pierre Hellier, Charles Kervrann, et Christian Barillot. Nonlocal means-based speckle fil-

tering for ultrasound images. *IEEE Transactions on Image Processing*, 18(10) :2221–2229, 2009.

[6] Gabriel Ramos-Llorden, Gonzalo Vegas-Sanchez-Ferrero, Marcos Martin-Fernandez, Carlos Alberola-Lopez, et S Aja. Anisotropic diffusion filter with memory based on speckle statistics for ultrasound images. 2014.

[7] David L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3) :613–627, 1995.

[8] Arash Vosoughi et Mohammad Bagher Shamsollahi. Speckle noise reduction of ultrasound images using m-band wavelet transform and wiener filter in a homomorphic framework. Dans *Proceedings of the 2008 International Conference on BioMedical Engineering and Informatics, BMEI 2008, May 28-30, 2008, Sanya, Hainan, China - Volume 2*, pages 510–515, 2008.

[9] Anil K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[10] Alin Achim, Anastasios Bezerianos, et Panagiotis Tsakalides. Novel bayesian multiscale method for speckle removal in medical ultrasound images. *IEEE Trans. Med. Imaging*, 20(8) :772–783, 2001.

[11] A. Serir et A. Belouchrani. Multiplicative multiresolution decomposition for 2d signals : application to speckle reduction in sar images. Dans *Image Processing, 2004. ICIP '04. 2004 International Conference on*, volume 1, pages 657–660 Vol. 1, Oct 2004.

[12] Anish Mittal, Rajiv Soundararajan, et Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3) :209–212, 2013.

[13] A. Serir, A. Beghdadi, et F. Kerouh. No-reference blur image quality measure based on multiplicative multiresolution decomposition. *Journal of Visual Communication and Image Representation*, 24(7) :911 – 925, 2013.

[14] Ju Zhang, Chen , et Yun Cheng. Comparison of despeckle filters for breast ultrasound images. *Circuits, Systems, and Signal Processing*, 34(1) :185–208, 2015.

Evaluation par observateur numérique basé tâche de la qualité d'IRM compressées

Christine CAVARO-MENARD[1], Mélanie SCHMIDT[1], ZHANG-GE Lu[2], Jean-Yves TANGUY[3], Patrick LE CALLET[4]

1- LARIS, EA 7315, Université Angers
2- IETR lab, UMR CNRS 6164, INSA de Rennes
3- Service de Radiologie, CHU d'Angers
4- IRCCyN lab, UMR CNRS 6597, Université de Nantes

Mail du correspondant : christine.menard@univ-angers.fr

Résumé

En imagerie médicale, l'évaluation objective de la qualité des images est un problème ouvert et complexe afin de garantir un diagnostic fiable. Quantifiant la performance d'un praticien à la réalisation d'une tâche diagnostique spécifique, les observateurs numériques basés tâches s'avèrent alors très pertinents. Le PCJO utilisé dans cette étude modélise les tâches de détection et de localisation de plusieurs signaux d'amplitude, d'orientation et de taille variables sur une image. Afin de mener à bien l'évaluation d'IRM cérébrales compressées, la notion de compression a du être insérée à plusieurs niveaux du modèle. Les résultats du modèle ont été comparés à ceux obtenus lors d'évaluations subjectives conduites avec 5 radiologues dont 1 expert en neuroradiologie. Cette étude montre que le modèle PCJO a une performance proche de celle de l'expert lors de l'évaluation d'images pas ou peu compressées (7.5:1). Pour un taux plus important (20:1), le PCJO doit intégrer un traitement simulant le processus mental effectué par le radiologue pour s'affranchir des artéfacts de compression visibles sur l'image.

Mots clefs

Evaluation basée tâche, Observateur Numérique, IRM

1 Introduction

Avec l'essor de nouveaux systèmes d'imagerie de plus en plus performants et offrant de multiples fonctionnalités, l'évaluation de la qualité des images obtenues en sortie de ces appareils est devenue primordiale. Cette qualité d'images médicales peut être définie par la performance d'un praticien à la réalisation d'une tâche diagnostique spécifique. Cette évaluation est alors dite «basée tâche» [1]. Cependant les évaluations subjectives de la qualité d'images réalisées avec des radiologues sont chronophages et fastidieuses. De plus, elles présentent une variabilité inter experts et intra expert qui peut biaiser les résultats finaux [2]. Néanmoins, cette évaluation de la qualité d'images médicales s'avère aujourd'hui indispensable pour pouvoir comparer et optimiser les différents systèmes d'imagerie médicale (systèmes d'acquisition, de visualisation ou de traitement d'images).

C'est dans ce contexte que des observateurs numériques ont été développés. Basés sur des modèles mathématiques, ils permettent de quantifier les performances d'un système d'acquisition, de visualisation ou de traitement d'images (reconstruction, compression, tatouage, etc.) et d'ainsi qualifier ce dernier pour un diagnostic fiable pour une tâche spécifique. Dans le cadre de la compression, il est alors possible de tester pour un même algorithme de compression une grande gamme de paramètres (taille des blocs, modèle de prédiction, etc.) et de définir les valeurs pour lesquelles la performance de la tâche à réaliser est proche de celle d'un expert.

Lors de l'interprétation d'un examen d'imagerie, le processus diagnostique peut se décomposer en différentes tâches dont les principales sont [3] :

- La tâche de détection qui se traduit par une note de confiance quant à la présence d'un signal (par exemple une lésion sur une image médicale) ;
- La tâche de localisation qui consiste à fournir l'emplacement du signal (par exemple de la lésion) ;
- La tâche de caractérisation qui permet d'obtenir des données plus complexes sur le signal, telles que sa texture, son contour ou encore sa régularité.

Ces trois tâches combinées à l'expertise du médecin permettent d'aboutir à un diagnostic. La modélisation de ces précédentes tâches est essentielle pour obtenir un observateur numérique qui fournira des résultats proches de ceux des radiologues.

Modélisant les tâches de détection et de localisation, l'observateur numérique PCJO (Perceptually relevant Channelized Joint Observer) fournit des résultats proches de ceux des radiologues pour la détection et la localisation de plusieurs signaux d'amplitude, d'orientation et de taille variables sur une image [4]. Le modèle a été validé sur des images IRM présentant plusieurs lésions de sclérose en plaques (SEP).

Peu d'études concernent l'adaptation et l'application des observateurs numériques pour l'évaluation basée tâche d'images médicales compressées [5,6,7,8,9]. Bien que ces études prouvent l'intérêt des observateurs numériques pour optimiser le paramétrage d'un algorithme de compression, les observateurs numériques étudiés sont limités à la tâche de détection.

L'étude présentée a pour objectif d'adapter l'observateur numérique PCJO pour l'évaluation basée tâches de détection et de localisation d'images IRM compressées en JPEG2000. Une évaluation subjective menée par des radiologues a permis de valider le modèle pour la tâche de détection-localisation de lésions de SEP.
Le modèle PCJO et l'insertion de la notion de compression dans le modèle seront présentés dans la section Matériel et Méthodes. Les résultats obtenus par le modèle seront comparés à ceux de l'évaluation objective (section Résultats) puis discutés.

2 Matériel et Méthodes

2.1 Images traitées

Les IRM cérébrales utilisées dans cette étude sont des coupes transversales de séquences volumiques T2 FLAIR de 10 sujets sains (âge moyen : 25 ans). Afin d'avoir une vérité terrain (gold standard) pour le modèle et l'étude subjective, des lésions de SEP sont simulées et insérées dans les zones de matière blanche des coupes d'IRM (la SEP étant caractérisée par des hypersignaux dans la substance blanche). La fonction de simulation est une fonction elliptique gaussienne dont les paramètres de forme, de taille, d'échelle et d'amplitude peuvent être contrôlés par l'équation suivante :

$$[x]_p = a \exp\left(-\frac{1}{2}\,(p-q)^t B^t D^{-1} B (p-q)\right)$$

où $[x]_p$ représente la valeur de l'intensité du signal ajouté à la coordonnée p, a l'amplitude du signal centré en q, D la matrice diagonale spécifiant l'échelle et la forme de l'ellipse et B la matrice de rotation.
Les images avec et sans lésions simulées ont été compressées selon le standard JPEG2000 encapsulé DICOM (standard largement utilisé dans les PACS) via le logiciel OsiriX aux taux de compression 7.5 :1 et 20 :1.

2.2 L'observateur numérique PCJO

Le PCJO est un modèle SKS (Signal Known Statistically) qui se rapproche plus de la routine clinique, comparé à un modèle SKE (Signal Known Exactly). En effet, les médecins n'ont pas de connaissance exacte des paramètres de la lésion recherchée avant d'effectuer leur diagnostic. Le PCJO s'inspire dans son fonctionnement de l'observateur numérique CHO (Channelized Hotelling Observer) du fait de la décomposition en canaux perceptuels [1,4]. Le PCJO modélise les deux étapes qui composent le processus de vision humaine : la sensation et la perception. L'étape de sensation correspond à un ensemble de procédés biochimiques et neurologiques qui interviennent suite au contact d'un stimulus sur les photorécepteurs de l'œil. L'étape de perception est plus complexe puisqu'il s'agit d'un processus cognitif d'interprétation du stimulus présenté. Cette dernière étape est difficile à modéliser par les observateurs numériques. Elle correspond en effet dans le cadre de l'évaluation d'images médicales à la modélisation de l'expérience du radiologue, mais aussi de la connaissance de la modalité d'imagerie et de la pathologie étudiée [10].
Le PCJO est composé d'une phase d'entraînement du modèle, d'une phase de détection des blocs d'image candidats à présenter une lésion et une phase de calcul de la figure de mérite qui permet de quantifier la performance de la tâche de détection-localisation des lésions.
La phase d'entraînement permet au modèle d'acquérir de l'expérience comme pourrait l'avoir un radiologue. L'entraînement est réalisé sur un grand nombre de blocs images présentant pour la moitié une lésion dont les paramètres (amplitude, forme et orientation) sont définis aléatoirement et qui est centrée dans le bloc.
Afin de détecter les blocs de l'image candidats à présenter une lésion, le modèle VDP (Visible Difference Predictor) de Daly [11] calcule la carte des différences perceptuelle entre une image de référence (ici une coupe IRM sans lésion) et l'image à évaluer (ici l'image de référence sur laquelle des lésions simulées sont ajoutées et ayant été ou non compressée). Le VDP permet de modéliser le système visuel humain (adaptation aux luminances, sensibilité aux contrastes, décomposition spatiofréquentielle, effet de masquage) en tenant compte des conditions d'observation (ambiance lumineuse, écran, distance d'observation). Cette étape modélise le processus de localisation.
La figure de mérite, quantifiant la performance du modèle dans la tâche de détection-localisation de lésions, s'obtient après application du CJO (Channelized Joint detection and estimation Observer) [4]. Le CJO est fondé sur le principe de détection-estimation jointes des lésions. L'estimation des paramètres α de la lésion ainsi que l'hypothèse de présence ou non d'une lésion H_k $(k=0,1)$ sur un bloc de test sont choisies conjointement pour maximiser la probabilité a posteriori jointe $P(\alpha, H_k|g)$ où g correspond à l'image considérée. La validation d'une des deux hypothèses se fait donc au travers de l'estimation de la probabilité maximum a posteriori (MAP) d'un signal dont les paramètres sont inconnus selon l'équation :

$$(\widehat{\alpha, H_k}) = arg \max_{\alpha, H_k} P(\alpha, H_k | g) = arg \max_{\alpha, H_k} P(g|\alpha, H_k) P(\alpha) P(H_k)$$

Le problème d'estimation cherche à trouver les paramètres estimés du signal qui maximisent cette probabilité :

$$\hat{\alpha} = arg \max_{\alpha} P(\alpha|g)$$
$$= arg \max_{\alpha} \{P(g|x_\alpha, H_1) P(\alpha) P(H_1), P(g|H_0) P(\alpha) P(H_0)\}$$

Une fois les paramètres estimés obtenus, la détection consiste à choisir l'une des deux précédentes hypothèses selon l'image considérée.

$$P(H_k|\hat{\alpha}, g) = \frac{P(H_k) P(\hat{\alpha}) P(g|H_k, \hat{\alpha})}{P(g|\hat{\alpha}) P(\hat{\alpha})} \propto P(H_k)\, P(g|H_k, \hat{\alpha}) \qquad k = 0,1$$

Le CJO intègre une décomposition en canaux perceptuels afin de réduire la dimension des matrices à traiter. Dans le cadre du CJO, deux types de canaux orientables ont été choisis : des canaux modifiables pour l'orientation et des canaux modifiables pour l'échelle du signal.

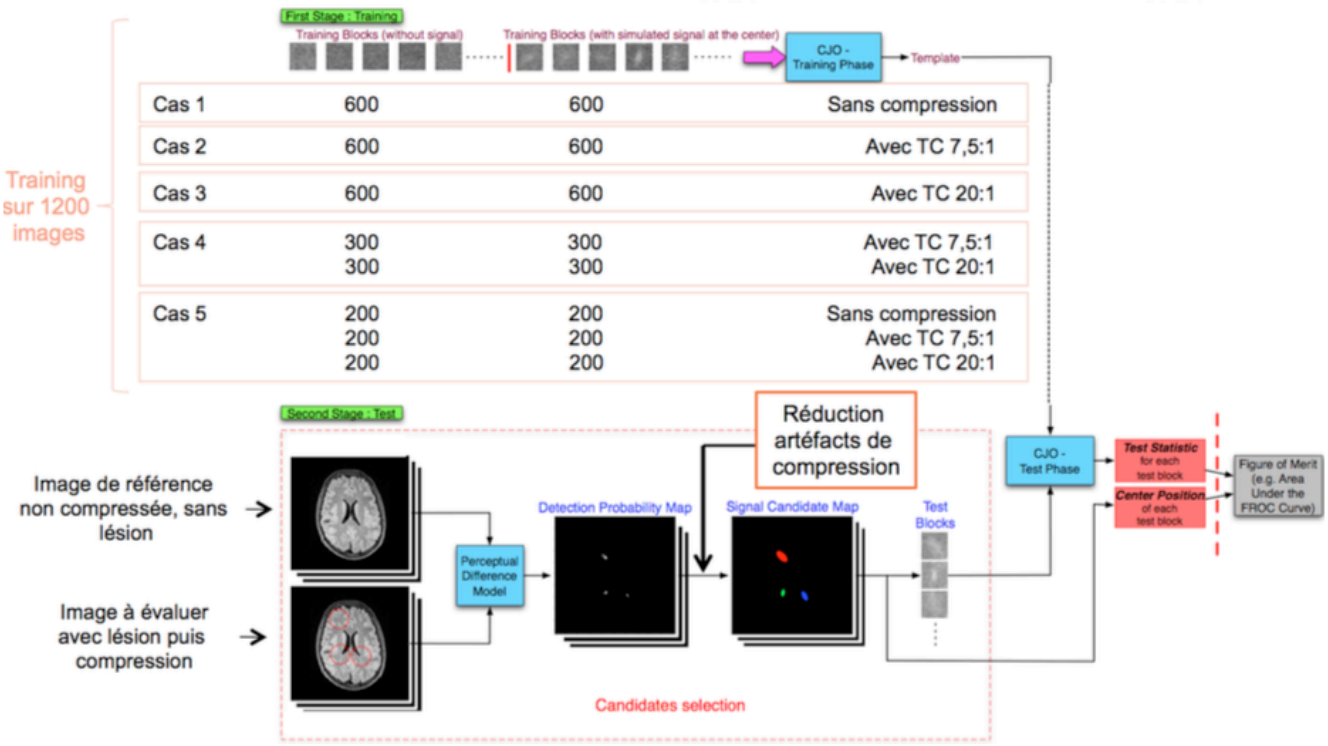


Figure 1 : *Schéma du PCJO adapté à l'évaluation d'images compressées*

Le CJO de la phase d'entraînement permet d'obtenir deux variables dans l'espace des canaux perceptuels : un signal référence x_0 et un template w. Le template correspond au poids donné à chaque pixel dans l'espace des canaux perceptuels par l'observateur réalisant la tâche de détection. Le test statistique correspondant à la note de confiance du modèle quant à la présence d'une lésion dans un bloc candidat est alors obtenu grâce à l'équation suivante :

$$\lambda = \max_{a,\theta,\sigma}(\lambda_{a,\theta,\sigma}) = \max_{a,\theta,\sigma}\left(\frac{w^{t}}{\left\|\mathrm{U}\left(\mathrm{A}'_{a,\theta,\sigma}\right)^{t}\right\|_F^2}\left(\mathrm{A}'_{a,\theta,\sigma}g' - \frac{1}{2}\widehat{x'_0}\right)\right)$$

où g représente l'image à tester, w le template, x_0 le signal référence, U et A deux matrices pour passer dans le domaine des canaux perceptuels. La dénotation « ' » indique que les paramètres sont définis dans le domaine des canaux perceptuels.

Ce processus est répété sur un certain nombre de couples d'images où des lésions plus ou moins subtiles sont ajoutées.

Une figure de mérite, obtenue grâce aux tests statistiques et aux centres des lésions supposées de chacun des blocs de test, permet finalement de quantifier les performances du modèle. La quantification des résultats de détection-localisation du modèle PCJO ainsi que des radiologues qui ont participé à l'étude subjective a été réalisée en utilisant la figure de mérite JAFROC1 [12].

2.3 Insertion de la notion de compression

L'insertion de la notion de compression a dû être effectuée à plusieurs niveaux du modèle : au niveau de la préparation des images, au niveau de la phase d'entraînement et de la phase de détection des blocs candidats (figure 1). Trois taux de compression ont été étudiés : 1:1 (c.-à-d. sans compression), 7.5:1 et 20:1. Au niveau de la phase de test (détection des blocs candidats), les images à évaluer correspondent aux images compressées après insertion des lésions simulées.

Différents cas d'entraînement ont été testés afin de voir lequel modélise au mieux la performance des radiologues en présence de compression. Sur une base de 1200 blocs d'entraînement, les différents cas sont :

Cas 1 : Non-prise en compte de la compression (avec 600 blocs sans lésion et 600 blocs avec lésion non compressés)

Cas 2 : Prise en compte de la compression au taux de 7.5:1 (avec 600 blocs sans lésion et 600 blocs avec lésion, compressés au taux de 7.5:1)

Cas 3 : Prise en compte de la compression au taux de 20:1 (avec 600 blocs sans lésion et 600 blocs avec lésion, compressés au taux de 20:1)

Cas 4 : Prise en compte de la compression aux taux de 7.5:1 et 20:1 (avec 300 blocs sans lésion et 300 blocs avec lésion, compressés au taux de 7.5:1 + 300 blocs sans lésion et 300 blocs avec lésion, compressés au taux de 20:1)

Cas 5 : Prise en compte de l'ensemble des taux de compression (avec 200 blocs sans lésion et 200 blocs avec lésion, non compressés + 200 blocs sans lésion et 200 blocs avec lésion, compressés au taux de 7.5:1 + 200 blocs sans lésion et 200 avec lésion, compressés au taux de 20:1)

Les blocs compressés avec lésion sont extraits de coupes d'IRM compressées après insertion des lésions simulées.

Après un premier essai, la phase de détection des blocs candidats a dû être adaptée afin de pouvoir être appliquée à l'évaluation d'images compressées. En effet, la carte VDP des différences perceptuelles entre l'image de référence (sans lésion et non compressée) et l'image à évaluer avec lésion(s) simulées et compressée présente de nombreuses différences perceptibles comme le montre la figure 2. De ce fait de nombreux blocs de tests sont générés, beaucoup ne correspondant à aucune lésion.

.Les différences sont situées principalement dans le fond de l'image (non utile au diagnostic). A fort taux de compression (20 :1) des artéfacts de compression (de type ringing artifacts) génèrent des différences perceptibles dans le cerveau généralement près des contours.

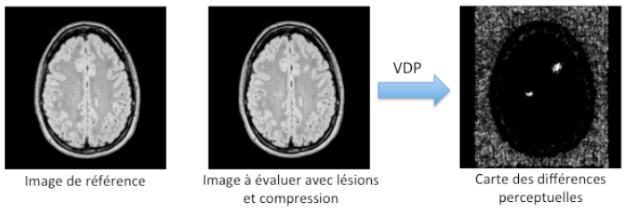


Figure 2 : *Effet de la compression sur la carte VDP*

Pour résoudre le problème dû aux artéfacts du fond de l'image, les images sources ont été segmentées manuellement afin de ne traiter que les pixels contenant une information diagnostique (suppression des pixels du fond et de la boîte crânienne). Un filtrage morphologique (ouverture puis fermeture avec un élément structurant en forme de disque de rayon 4 (PT1 : post traitement 1), 5 (PT2 : post traitement 2) ou 6 (PT3 : post traitement 3)) a été réalisé sur les cartes VDP afin de supprimer les artéfacts de taille inférieure à la plus petite lésion simulée. La carte de différences perceptuelles s'avère alors exploitable comme le montre la figure 3.

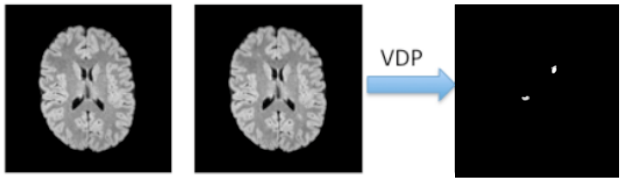


Figure 3 : *Carte des différences perceptuelles après suppression des zones non utiles au diagnostic et filtrage morphologique*

2.4 Evaluation subjective

Cinq radiologues ont participé à l'étude subjective qui s'est déroulée dans une salle de test au sein du CHU d'Angers. Le premier radiologue (Expert 1) est un expert en neuroradiologie avec 24 ans d'expérience dans le domaine. Les radiologues 2 à 5 sont des internes en radiologie ayant respectivement 4 ans, 5 ans, 3 ans et demi et 6 ans d'expérience en radiologie.

Chacun des radiologues a participé à trois sessions d'évaluation d'images. Chaque session se déroule sans limites de temps et en double aveugle. La luminosité de la salle d'expérience est similaire aux conditions de travail des radiologues, et doit être inférieure à 15 lux. La console de visualisation des images est le Positoscope de KEOSYS calibré DICOM-GSDF. La distance de visualisation à l'écran est de 40 cm et l'angle visuel est de 42 pixels par degré.

162 images sont présentées à chacune des sessions d'évaluation (54 images non compressées, 54 images compressées à 7.5 :1 et 54 images compressées à 20 :1 toutes différentes - hauteur de coupe et/ou examen différents). A noter que ces images sont identiques à celles testées par le PCJO. Les images sont présentées aux radiologues de manière aléatoire.

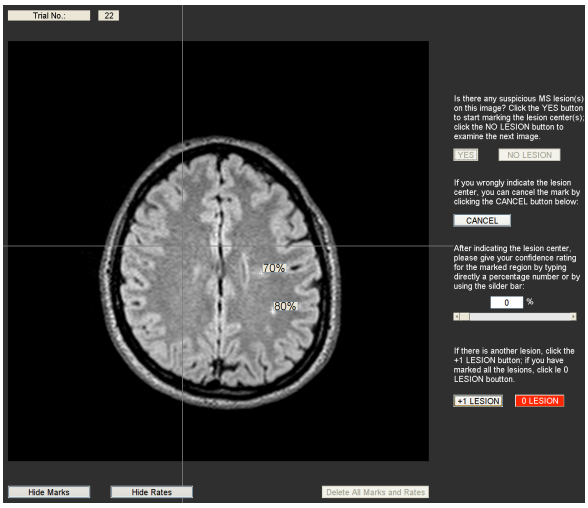


Figure 4 – *Interface graphique de test*

La figure 4 montre une capture d'écran de l'interface de test. La tâche du radiologue est de détecter sur chaque image les lésions de sclérose en plaques (SEP) dont la taille, la position, l'amplitude et l'orientation sont inconnues. Les radiologues sont informés que les images contiennent entre 0 et 4 lésions. Si le radiologue détecte une lésion sur la coupe présentée, il clique sur le bouton *YES* et un curseur apparait afin de lui permettre de cliquer sur le centre de la lésion repérée. La lésion est ainsi marquée d'une croix et le curseur disparaît. Un bouton *CANCEL* permet d'annuler la lésion marquée. Puis, le radiologue indique une note de confiance (en %) de la présence de cette lésion. Si le radiologue détecte une autre

lésion, il clique sur « +1 LESION » et réitère le marquage et la note de confiance. Dans le cas contraire, le radiologue clique sur « O LESION » afin de passer à la coupe suivante. Chaque coupe est examinée de la même façon. Les cordonnées des lésions marquées ainsi que la note de confiance définie par le radiologue sont enregistrées pour chaque radiologue.

3 Résultats

3.1 Résultats de l'étude subjective

Le tableau 1 présente la figure de mérite JAFROC1, évaluant la performance de la détection- localisation des lésions, pour chaque radiologue et pour les images compressées aux taux 1:1, 7.5:1 et 20:1.

Tableau 1 : *Performance de détection-localisation pour chacun des radiologues et chaque taux de compression*

	JAFROC1 FOM Détection et localisation		
Taux de compression	1	7.5	20
Expert 1	0.7840	0.8110	0.7717
Radiologue 2	0.6679	0.6810	0.6197
Radiologue 3	0.6219	0.6258	0.6064
Radiologue 4	0.6212	0.6959	0.6450
Radiologue 5	0.6740	0.6619	0.5977

On note que la performance de détection-localisation est meilleure pour les images compressées au taux de 7.5:1 (Expert 1 et Radiologues 2 à 4), et qu'elle est la plus faible pour le taux de 20:1 (hors Radiologue 4). Le radiologue 5 présente un résultat assez similaire pour les taux de compression 1:1 et 7,5:1 avec un résultat légèrement meilleur pour les images non compressées. Le radiologue 4 montre de moins bons résultats pour les images non compressées. Ce résultat peut s'expliquer par la faible expérience de ce radiologue. Les résultats de l'expert en radiologie (Expert 1) sont nettement meilleurs que ceux des radiologues internes.

3.2 Résultats du PCJO

La figure de mérite JAFROC1 et l'aire sous la courbe ROC ont été également calculées avec le PCJO sur les mêmes images que les radiologues (Tableau 2).

Tableau 2 : *Résultats du PCJO pour les 5 cas de training du PCJO et pour différents taux de compression et post-traitement des cartes VDP (SPT=sans post-traitement)*

	Cas 1	Cas 2	Cas 3	Cas 4	Cas 5
Sans compression	0.848	0.846	0.844	0.846	0.846
Jpg2k 7.5 :1	0.859	0.856	0.856	0.856	0.857
Jpg2k 20:1 (SPT)	0.432	0.427	0.428	0.426	0.430
Jpg2k 20 :1 (PT1)	0.460	0.453	0.452	0.452	0.457
Jpg2k 20 :1 (PT2)	0.636	0.627	0.624	0.626	0.632
Jpg2k 20 :1 (PT3)	0.837	0.838	0.834	0.835	0.838

Nous pouvons noter au vu des résultats que le cas de training n'influe que très peu sur les résultats (quasi constants pour chacun des cas d'images à évaluer).
Dans le cas où les cartes de différences perceptuelles ne sont pas post-traitées (cas : Taux 20:1 sans traitement), les résultats JAFROC1 du PCJO sont très mauvais puisque l'aire sous la courbe AFROC est de 0.43 en moyenne (quel que soit le cas de training). Ceci confirme que le modèle ne peut pas quantifier correctement la performance de détection-localisation des lésions lorsque les images sont compressées à un taux de 20:1.
Concernant les images sans compression et compressées à un taux de 7.5:1 non post-traitées, le PCJO montre une meilleure performance de détection-localisation pour les images compressées au taux de 7.5:1 suivi de près par les images sans compression. Une légère compression permet ainsi tout comme pour les radiologues une meilleure détection-localisation de lésions de SEP.
Le post-traitement choisi a une très grande influence sur les résultats finaux du modèle. Un post-traitement très efficace, tel que PT2 ou PT3, permet d'obtenir des résultats proches de ceux de l'expert pour PT3 ou des autres radiologues pour PT2.

3.3 Comparaison des résultats

L'analyse de p-values obtenues grâce au logiciel JAFROC permet d'étudier la significativité des différences au niveau de la figure de mérite JAFROC1 entre chaque radiologue et le PCJO, et ce pour les taux de compression 1:1, 7.5:1, 20:1 sans post-traitement, 20:1 avec post-traitement PT2 et 20:1 avec post-traitement PT3.
Les différences sont significatives entre les résultats de l'Expert et des Radiologues 2 à 5 quel que soit le taux de compression, entre les résultats du PCJO et des Radiologues 2 à 5 pour les taux de compression 1:1, 7.5:1, 20:1 et 20:1 avec post-traitement PT3 et entre le PCJO et l'Expert 1 pour les taux de compression 20:1 et 20:1 avec post-traitement PT2. Les autres différences ne sont pas significatives.

4 Discussion

Les résultats de cette étude montrent la possibilité d'appliquer le modèle PCJO pour l'évaluation d'images compressées en JPEG 2000 à faible taux de compression. Les résultats du modèle sont pour les images non compressées et compressées à 7.5:1 validés par l'étude subjective menée avec cinq radiologues.
L'évaluation des images compressées à 20:1 est particulière. En effet, à un tel taux de compression, les images sont très artéfactées. Si les radiologues, du fait de leur expérience et leur connaissance des lésions de SEP, ont relativement bien su faire abstraction des artéfacts lors de l'évaluation des images, le modèle n'a quant à lui dans sa configuration d'origine pas réussi à le faire. L'ajout d'un post-traitement des cartes VDP de différences

perceptuelles permet d'intégrer dans le modèle la différentiation artéfacts-lésions. En supprimant une grande partie des artéfacts, le modèle donne des résultats proches de ceux des radiologues. C'est finalement l'ajout d'un post-traitement qui fait que le PCJO peut être appliqué aux images compressées à 20:1. Néanmoins, le post-traitement choisi influe directement sur les performances du modèle. Nous pouvons ainsi considérer qu'à l'image du modèle qui réussit à détecter-localiser les lésions sur les images à fort taux de compression grâce au post-traitement, les radiologues opèrent eux aussi un post-traitement mental pour supprimer les artéfacts de l'image.

Nous avons également pu noter que l'expertise de l'observateur influençait la performance de détection-localisation lors des évaluations subjectives. Les résultats de l'expert sont apparus comme meilleurs comparés aux autres radiologues, et ce quel que soit le niveau de compression. L'expert a d'ailleurs indiqué durant les sessions d'évaluation qu'il avait acquis avec les années une expérience suffisante pour pouvoir comparer chaque coupe avec son double sans lésion et sans compression, évitant plus de faux positif. L'expert est donc capable de comparer les images qu'il visionne avec un atlas du cerveau qu'il s'est forgé avec l'expérience. Ce concept est celui du PCJO.

Autre remarque importante, la performance de la détection-localisation de lésions de SEP est apparue meilleure pour les images compressées au taux de compression 7.5:1 pour l'ensemble des radiologues (hormis le Radiologue 5) ainsi que pour le modèle. Ceci va en faveur du fait qu'un léger lissage de l'image favoriserait la détection de lésions par un radiologue et par le modèle. Néanmoins, la différence des résultats avec les images non compressées reste faible et l'on peut considérer qu'un taux de compression JPEG 2000 de 7.5:1 est très largement envisageable pour étudier des images d'IRM en routine clinique.

Enfin, l'ensemble des radiologues s'est accordé sur le fait que l'évaluation de coupes simples d'IRM n'était pas optimale pour la détection-localisation de lésions de SEP. En effet, l'évaluation des coupes volumiques leur permettrait de naviguer entre les coupes adjacentes et d'éviter de confondre un fond de sillon ou un prolongement de corne occipitale de ventricule latéral avec une lésion.

5 Conclusion

Les résultats de cette étude montrent l'intérêt des observateurs numériques basés tâches pour évaluer la performance d'une tâche diagnostique sur des images compressées. Le modèle PCJO étudié a une performance proche de celle de l'expert dans l'évaluation des taux de compression 1:1 et 7.5:1. Dans le cas de l'évaluation d'images compressées à 20:1, le choix du post-traitement influence les performances du modèle.

Références

[1] H. Barrett and K. Myers, *Foundations of image science.* John Wiley and Sons, Inc., Hoboken, New Jersey, USA, 2004.

[2] C. Cavaro-Ménard, L. Zhang and P. Le Callet. Diagnostic quality assessment of medical images: challenges and trends, Dans *Visual Information Proceedings (EUVIP)*, pp. 277-284, Paris, 2010.

[3] C. Cavaro-Ménard, J.Y. Tanguy and P. Le Callet. Eye-position recording during brain MRI examination to identify and characterize steps of glioma diagnosis. Dans *Proceedings Medical Imaging, Image Perception, Observer Performance, and Technology Assessment*, San Diego, 2010.

[4] L. Zhang, C. Cavaro-Ménard, P. Le Callet and J.Y. Tanguy. A Perceptually relevant Channelized Joint Observer (PCJO) for the detection-localization of parametric signals. *IEEE Transactions on Medical Imaging*, vol. 31, pp. 1875-1888, 2012.

[5] Y. Zhang, B.T. Pham and M.P. Eckstein. Evaluation of JPEG 2000 encoder options: Human and model observer detection of variable signals in X-Ray coronary angiograms. *IEEE Transactions on Medical Imaging*, vol. 23, n° 5, pp. 613-632, 2004.

[6] Y. Zhang, B.T. Pham and M.P. Eckstein. Task-based model/human observer evaluation of SPIHT wavelet compression with human visual system-based quantization. *Academic Radiology*, vol. 12, n° 3, pp. 324-336, 2005.

[7] S. Suryanarayanan, A. Karellas, S. Vedantham, S. M. Waldrop and C.J. D'Orsi. Detection of simulated lesions on data-compressed digital mammograms. *Radiology*, vol. 236, n°1, pp. 31-36, 2005.

[8] M.P. Eckstein, C.K. Abbey and J.L. Bartroff. Model observer optimization of JPEG image compression. *SPIE Proceedings*, Vol. 3981, pp. 106-115, 2000.

[9] J.G. Brankov, Y. Yang, L. Wei, I. El Naqa and M.N. Wernick. Learning a channelized observer for image quality assessment. *IEEE Transactions on Medical Imaging*, Vol. 28, n° 7, pp. 991-999, 2009.

[10] G. Mather, *Foundations of Sensation and Perception*. Hove and New York, USA: Psychology Press Ltd., Talor & Francis Inc., 2009.

[11] S. Daly, « The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity », *Digital images and human vision*, pp. 179-206, 1993.

[12] D. P. Chakraborty, « New developments in observer performance methodology in medical imaging », *Seminars in Nuclear Medicine*, Vol. 41, pp. 401-418, 2011.

Analyse d'empreintes digitales à partir de paramètres structurels calculés sur une référence réduite de l'image.

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

L'analyse d'empreintes digitales est un domaine important en biométrie. Généralement, cette analyse est basée sur une étude des minuties associées. Ces dernières servent également lorsque l'on compare plusieurs empreintes entre elles. Dans ce papier, nous proposons plusieurs paramètres basés sur la triangulation de Delaunay appliquée aux minuties pour l'analyse des empreintes digitales. Nous montrons l'utilité de ces paramètres pour l'analyse de bases de données d'empreintes digitales existantes en particulier pour reconnaître le type d'empreinte digitale sans avoir accès à l'image associée. Les résultats obtenus montrent la pertinence de l'approche proposée.

Mots clefs

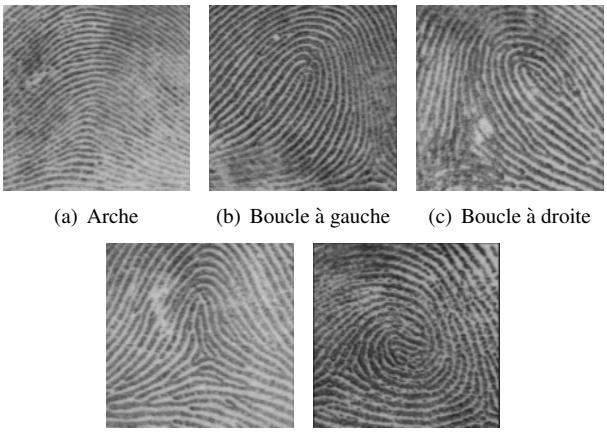
Triangulation de Delaunay, Template de minuties, Empreinte digitale.

1 Introduction

Depuis plusieurs années, l'utilisation de données biométriques est devenue incontournable, que se soit avec les passeports lors des contrôles aux frontières, le contrôle d'accès sur un smartphone, etc. En 2013, le premier smartphone avec lecteur d'empreinte digitale a été déployé sur le marché par Apple™avec l'Iphone®5S. Bien que cela facilite l'usage des utilisateurs, les données biométriques sont des données très sensibles car non révocables. C'est la raison pour laquelle les templates biométriques (qui sont une représentation réduite d'une empreinte digitale) sont sauvegardés dans des éléments sécurisés (SE). Étant donné que la capacité de stockage est limitée sur un SE, ce template ne pourra contenir qu'un nombre restreint d'informations. Une image ne peut donc pas être contenu dans ce template, c'est pourquoi des points caractéristiques sont extraits de l'image, ce sont les minuties. Ces minuties sont stockées dans le SE au format ISO Compact Card II [1]. Ce format est alors utilisé pour la comparaison entre la référence (enregistrée dans le SE) et le template extrait d'une nouvelle empreinte acquise par un capteur biométrique.

Une empreinte digitale peut être catégorisée en cinq classes en fonction de son type, à savoir 1) Arche, 2) Boucle à gauche, 3) Boucle à droite, 4) Tente et 5) Spirale. La figure 1 donne un exemple de chacune des cinq classes. Même si de nombreux travaux portant sur la reconnaissance du type d'une empreinte digitales existent, ces derniers utilisent systématiquement l'image de l'empreinte [2, 3] et pas seulement le template de minuties. A notre connaissance, il n'existe pas de travaux utilisant exclusivement ce template.

Ce papier est organisé de la façon suivante : tout d'abord, nous présentons les paramètres utilisés pour caractériser la structure d'une empreinte. Ensuite, nous proposons deux contributions. La première permet de comparer une empreinte synthétique avec une empreinte réelle, la seconde permet de reconnaitre le type des empreintes digitales exclusivement basée sur l'extraction de caractéristiques à partir du seul template de minuties. En conclusion, nous discutons de la pertinence de notre approche et des perspectives à cette étude.



(a) Arche (b) Boucle à gauche (c) Boucle à droite

(d) Tente en arche (e) Spirale

Figure 1 – *Les 5 grands types d'empreintes définis par Henry*

2 Extraction des paramètres de l'empreinte

Dans cette étude, nous allons utiliser des bases de données d'empreintes digitales, qui sont composées initialement d'images. Nous générons le template de minuties à partir de l'image grâce à un extracteur de minuties. La figure 2 montre comment les minuties sont extraites à partir de l'image et comment elles se composent.

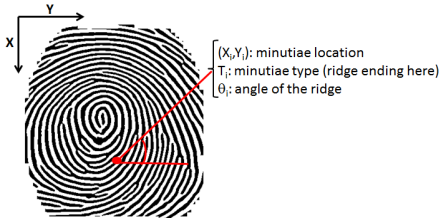


Figure 2 – *Schéma montrant comment sont extraites les minuties à partir d'une image*

Ces minuties sont enregistrées dans un template binaire (ce qui signifie qu'aucun accès à l'image originale n'est possible). Chaque minutie est codée sur 3 octets et contient un ensemble de quatre valeurs $(x_i, y_i, T_i, \theta_i)$, $i = 1 : N_j$ où les coordonnées (x_i, y_i) correspondent à la localisation des minuties dans l'image, T_i correspond au type de minutie (bifurcation, fin de crête, ...), θ_i représente l'orientation de la minutie (relative à la crête) et N_j le nombre de minuties pour l'échantillon j de l'utilisateur.
Comme décrit précédemment, lorsque l'on utilise un template de minuties, nous avons au final seulement quatre types de paramètre caractérisant une minutie, ce qui peut s'avérer trop peu pour modéliser une empreinte digitale dans son ensemble. Afin d'étoffer le nombre d'attributs par minutie, nous avons utilisé la triangulation de Delaunay [4, 5] dans laquelle les minuties de l'empreinte correspondent aux sommets des triangles obtenus. Ceci nous permet d'extraire six nouveaux paramètres par minutie, tel que décrit dans la section 2.1 et synthétisé par la Figure 3.

2.1 Triangulation de Delaunay

La triangulation de Delaunay est souvent utilisée dans différents domaines comme par exemple en géométrie algorithmique [6], ou la reconstruction de surface [7, 8]. Dans notre cas, l'hypothèse formulée est la suivante : à l'instar des minuties qui sont spécifiques à une empreinte, la triangulation de Delaunay, dont les sommets des triangles sont les minuties extraites, doit également pouvoir être spécifique à une empreinte. Ainsi, une caractérisation de cette triangulation doit pouvoir être réalisée par des statistiques. De plus, la triangulation de Delaunay présente l'avantage de s'affranchir des problèmes de translation et de rotation du template de minuties mais aussi de faire une abstraction du template de minuties. En effet, à partir de la triangulation obtenue, il est possible d'extraire plusieurs

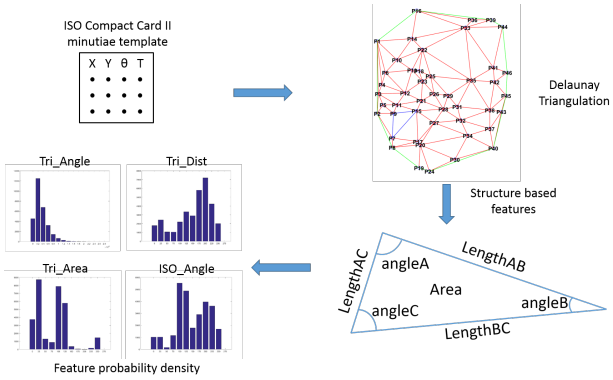


Figure 3 – *Schéma général utilisé pour le calcul des paramètres à partir d'une représentation par triangulation de Delaunay du template de minuties d'une empreinte digitale.*

paramètres de forme tels que la longueur de l'enveloppe convexe, la valeur des angles d'un triangle, l'aire des triangles, le périmètre d'un triangle, etc.
Pour chaque template, nous calculons la triangulation de Delaunay basée sur les minuties (la Figure 4 montre un exemple de triangulation obtenue). Pour chacun des triangles, sept paramètres sont calculés et sont sauvegardés dans un vecteur nommé **TriInf**, tel que :

- **angleA** : sa mesure associée en degré
- **angleB** : sa mesure associée en degré
- **angleC** : sa mesure associée en degré
- **LengthAB** : La longueur entre le point A et B du triangle
- **LengthAC** : La longueur entre le point A et C du triangle
- **LengthBC** : La longueur entre le point B et C du triangle
- **Area** : L'aire du triangle

Un 8ème paramètre est ajouté à la structure de données **TriInf**. Ce dernier correspond à l'orientation contenu initialement dans le template de minuties. Pour résumer, le vecteur d'attribut TriInf_{jk} est généré pour le template j de l'individu k et se compose de quatre groupes principaux de caractéristiques :

$$\begin{aligned}\text{TriInf}_{j,k} = \big[&\{\text{AngleA}_{jkl}, \text{AngleB}_{jkl}, \text{AngleC}_{jkl}\},\\ &\{\text{LengthAB}_{jkl}, \text{LengthAC}_{jkl}, \text{LengthBC}_{jkl}\},\\ &\{\text{Area}_{jkl}\}, \{\text{Orientation}_{jki}\}\big],\\ &\forall l \in [1; M_j]; i \in [1; N_j]\end{aligned} \quad (1)$$

où $\{\text{AngleA}_{ijk}, \text{AngleB}_{ijk}, \text{Angle C}_{ijk}\}$ est le vecteur des données relatives à la valeur des angles du triangle M_j du template j, $\{\text{LengthAB}_{jkl}, \text{LengthAC}_{jkl}, \text{LengthBC}_{jkl}\}$ représente le vecteur de données relatives aux longueurs calculées du triangle M_j du template j, $\{\text{Area}_{jkl}\}$ correspond au vecteur de données relatif au triangle M_j du template j et $\{\text{Orientation}_{jki}\}$ correspond au vecteur de

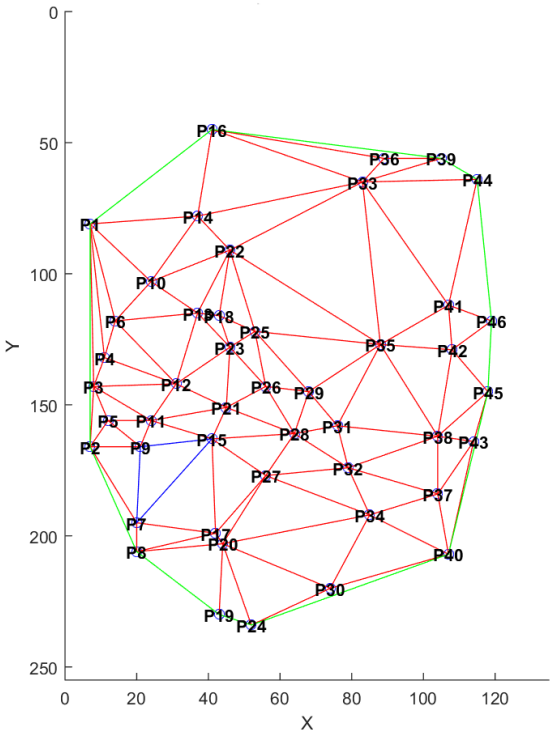


Figure 4 – *Triangulation de Delaunay pour un template de minuties*

données contenant l'angle ISO du template N_i de la minutie j.
À partir de ce vecteur de caractéristiques, et pour chacun des ensembles associés, un histogramme normalisé est calculé afin de prendre en compte la densité de chaque ensemble de paramètres. Finalement, le vecteur d'attributs $\text{TemplateStruct}_{jk}$ associé au template j de l'individu k est défini comme suit :

$$\begin{aligned}\text{TemplateStruct}_{jk} = \{&\text{Type}_{jk}, \text{HistoAngle}_{jk},\\ &\text{HistoDistance}_{jk}, \text{HistoArea}_{jk},\\ &\text{HistoISOAngle}_{jk}, \text{NbrMinutiae}_{jk}\}\end{aligned} \quad (2)$$

où HistoAngle_{jk}, $\text{HistoDistance}_{jk}$, HistoArea_{jk} et $\text{HistoISOAngle}_{jk}$ sont les histogrammes calculés à partir de leurs sous-vecteurs associés TriInf_{jk}. Ces histogrammes sont générés avec un nombre N de niveaux, de manière à pouvoir affiner la précision de la distribution des paramètres.

3 Empreinte synthétique vs. empreinte réelle

Dans cette première contribution, nous nous intéressons à savoir si une empreinte digitale synthétique présente ou non des différences structurelles majeures avec une empreinte réelle. Nous avons donc utilisé une base générée avec SFinge (FVC2004DB4) et des bases d'empreintes réelles (FVC2002DB2, FVC2004DB1, FVC2004DB2, FVC2004DB3) pour valider ou non l'utilisation de SFinge pour notre seconde contribution.

3.1 Protocole

Bases de données. Dans cette expérimentation, nous avons utilisé un jeu de données de la base FVC2002 [9] et quatre sous-bases de la base FVC2004 [10]. Chacune de ces bases de données est composée de 100 individus et 8 templates par individu, soit un total de 800 images par base. Le détail de chaque base est donné dans le tableau 1.

DB	Sensor	Dim.	Resolution
02DB2A	Optical	296×560	569dpi
04DB1A	Optical	640×480	500dpi
04DB2A	Optical	328×364	500dpi
04DB3A	Thermal	300×480	512dpi
04DB4A	SFinGe	288×284	about 500dpi

Tableau 1 – *Spécifications de chaque base FVC.*

Comme reporté dans le tableau 1, on observe que la taille des images pour chaque base est différente et que la résolution est approximativement de 500dpi pour chacune des bases considérées dans cette étude.
Comme ces bases sont constituées d'images, nous devons extraire les minuties, nous avons utilisé un extracteur du NIST nommé MINDTCT [11] pour créer des bases de template de minuties. Pour chaque élément présent dans les bases de données testées, nous avons calculé les caractéristiques normalisées de chaque template en utilisant des histogrammes à 64 niveaux ($N = 64$). Nous avons utilisé 64 niveaux car la valeur les angles présents dans le template de minuties.
Pour les bases de données FVC, étant donné que nous n'avons aucune information sur le type du template, le premier paramètre de la structure est fixé à zéro.
Pour chaque template, on obtient au final une structure contenant 258 paramètres ($4 \times 64 + 2$).

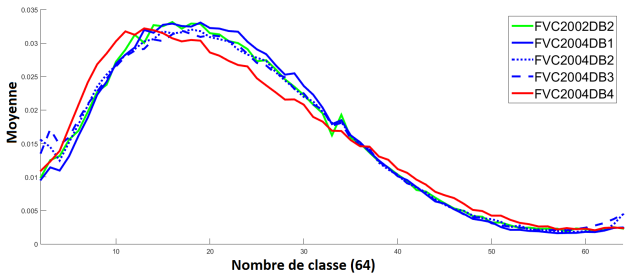
3.2 Résultats

L'évolution de chaque paramètre est montré dans la Figure 5. Nous pouvons observer une même évolution concernant la distribution du paramètre HistoAngle qui représente les angles des Triangles (Figure 5(a)) et la distribution du paramètre HistoArea concernant l'aire des Triangles (Figure 5(c)).
En revanche, lorsque l'on considère le paramètre HistoDistance qui est associé à la longueur des arêtes des triangles (Figure5(b)), nous pouvons observer que distribution asociée à la base de donnée FVC2002DB2 diffère des distributions associées aux quatre autres bases. Cette constatation peut également être formulée pour le paramètre HistoISOAngle qui correspondant à la distribution de l'orientation des minuties (Figure 5(d)). En effet, on

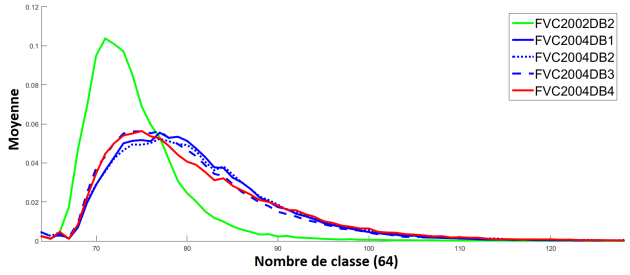
observe qu'il y a seulement un peu moins de 220 valeurs différentes pour les orientations des minuties lorsque l'on considère les empreintes de la base FVC2002DB2. Cela peut être partiellement expliqué par le fait que la résolution des images pour cette base est plus petite que dans les autres bases de données.
Pour toutes les autres bases de données, nous remarquons la même évolution et ce quel que soit le paramètre retenu.
Il est néanmoins intéressant de noter que parmi toutes les bases tests, seule la base FVC2004DB4 contient des données synthétiques. Nous pouvons effectivement observer sur toutes les sous-figures de la Figure 5 que la courbe rouge correspondant aux données extraites de la base FVC2004DB4 a une évolution très similaire à l'évolution des courbes associées aux autres bases de données contenant des images d'empreintes réelles.
Cette observation confirme les résultats présentés précédemment lors la compétition FVC2004 [12] à savoir que SFinge [13] permet de générer des empreintes digitales synthétiques réalistes pour lesquelles il est possible d'atteindre des niveaux de performances, en terme de taux de reconnaissance, identiques à ceux obtenus à partir de données réelles. Ainsi, en calculant les paramètres du vecteur TemplateStruct$_{jk}$ sur une base de données d'images d'empreintes synthétiques, la distribution des paramètres ne sera pas affectée par le fait que les données sont synthétiques, ce qui nous permettra d'étendre nos conclusions sur les bases d'images d'empreintes réelles.
Nous avons pour ces raisons choisi d'utiliser SFinge pour créer des bases de données d'empreintes digitales pour lesquelles le type sera connu. Ces bases serviront pour notre seconde contribution sur la reconnaissance du type d'empreinte à partir d'un template de minutie sans avoir d'accès à l'image.

4 Reconnaissance du type d'empreinte

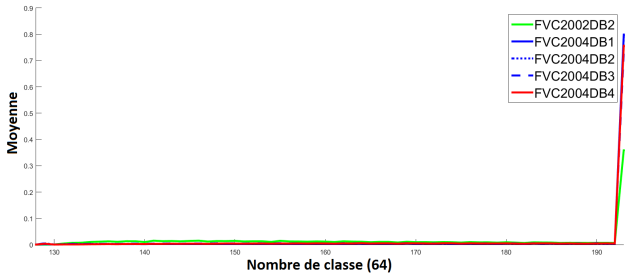
La contribution précédente nous à permis grâce à la triangulation de Delaunay de démontrer que des bases de données synthétiques se comportent comme des bases réelles. Cela nous permet, pour cette seconde contribution, de pouvoir utiliser SFinge pour générer des bases de données avec connaissance du type de l'empreinte. La plupart des méthodes de classification d'images d'empreintes digitales se basent sur la classification de Henry qui les catégorisent en cinq grandes familles : Arches, Boucle à gauche, Boucle à droite, Tente et Spire.
Dans cette contribution, nous voulons prédire le type du template d'empreinte lorsque nous n'avons aucun accès à l'image originale. Pour ce faire, nous avons utilisé un SVM (Support Vector Machine) pour créer un modèle afin de déterminer le type du template de minuties. Dans un premier temps, nous expliquons les paramètres de l'expérimentation, ensuite nous présentons la corrélation entre les attributs et le type d'empreinte. Pour finir, nous étudions l'influence du nombre de niveaux utilisés lors la construction du vecteur TemplateStruct$_{jk}$ sur le taux de reconnaissance du type de l'empreinte.



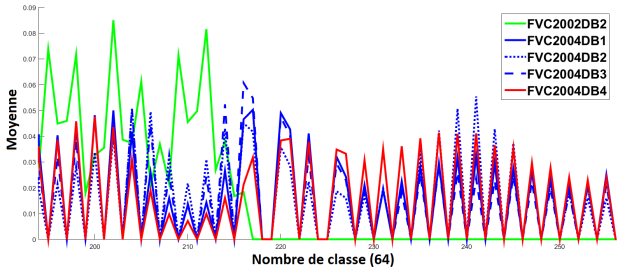
(a) Angles du triangle



(b) Longueur des arêtes du triangle



(c) Aire du triangle



(d) Angle des minuties

Figure 5 – *Evolution des bases FVC pour tous les attributs*

4.1 Base de données SFinge

Les bases des données FVC ne possède aucune information explicite portant sur le type d'empreinte associée aux images. Comme nous l'avons rappelé précédemment, les bases de données générées avec SFinge ont des caractéristiques très similaires aux bases d'empreintes digitales réelles et possède l'avantage de conserver cette in-

formation du type de l'empreinte. En se basant sur ce constat, nous avons alors généré cinq bases de données avec SFinge, une pour chaque type d'empreinte (voir tableau 2). Chaque base de données ainsi générée contient 800 templates et nous avons calculé pour chaque template d'empreinte le vecteur d'attributs TemplateStruct$_{jk}$.

Type d'empreinte	Valeur
1	Arche
2	Boucle à gauche
3	Boucle à droite
4	Tente
5	Spirale

Tableau 2 – *Table du type d'empreinte*

4.2 Corrélation entre tous les attributs

Après le calcul de tous les attributs des bases de données SFinge, nous calculons la corrélation entre les différents critères et le type de template d'empreinte. La Figure 6 montre que la corrélation entre les attributs ne sont pas très importante. Nous pouvons néanmoins noter que la paramètre HistoDistance lié à la longueur des arêtes permet d'améliorer le niveau de corrélation avec le type d'empreinte. Il convient également de noter que le paramètre d'orientation HistoISOAngle présente le plus haut niveau de corrélation avec le type d'empreinte.

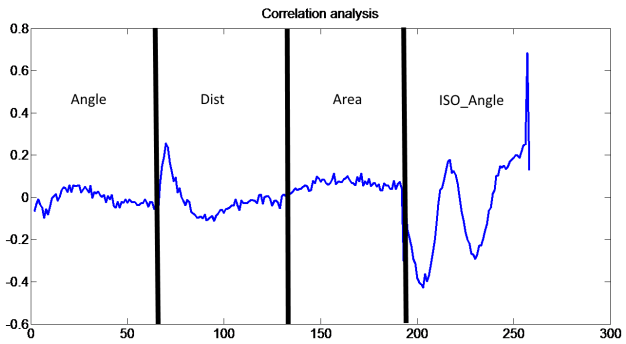


Figure 6 – *Corrélation entre tous les attributs*

4.3 Classification - SVM

Dans ce papier, nous avons utilisé les techniques de classification basées sur les SVM (Support Vector Machine) principalement en raison de leurs fortes capacités en généralisation. Cette méthode a été développée par Vapnik [14]. Le but est alors de classer un objet x à l'aide d'une marge maximale associée à un sous-ensemble de la base d'apprentissage dont les élements sont les vecteurs de support et d'une fonction noyau. Cette dernière permet d'opérer un changement de repère dans un espace de plus grande dimension afin d'arriver à un problème de séparation linéaire des données, lorsque initialement les données ne sont pas linéairement séparables.
Soit un ensemble d'apprentissage $A = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ composé de m couples (vecteur d'attributs, classe) avec $x_i \in \mathcal{R}^n$ et $y_i \in \{-1, +1\}$. L'algorithme des SVM projette les vecteurs x_i dans un espace de travail $\mathbf{H}$ à partir d'une fonction non linéaire $\phi : \mathcal{R}^n \rightarrow \mathbf{H}$. L'hyperplan optimal de séparation des deux classes dans l'espace $\mathbf{H}$ est ensuite recherché. Cet hyperplan $(\mathbf{w}, b)$ matérialise la frontière de séparation entre les deux classes. La classe y d'un nouvel exemple $\mathbf{x}$ est définie par :

$$y = < \mathbf{w}, \phi(\mathbf{x}) > + b = \sum_{i \in SV} \alpha_i^* y_i K(\mathbf{x_i}, \mathbf{x}) + b \qquad (3)$$

avec $\alpha_i^* \in \mathbb{R}$ et $K(.,.)$ est la fonction noyau.
L'hyperplan de séparation est optimal s'il maximise la distance qui le sépare des exemples lui étant le plus proche. Cette distance est appelée marge du classifieur. Les valeurs α_i^* maximisant le critére d'optimalité sont calculés en maximisant

$$-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) + \sum_{i=1}^{m} \alpha_i$$

sous les contraintes $\forall_{i=1}^m : \sum_{i=1}^m y_i \alpha_i = 0\,,\ 0 \leq \alpha_i \leq C$ avec C le coefficient de pénalisation.
L'algorithme SVM est initialement conçu pour des problèmes de classification à 2 classes. Dans cet article, étant donné que nous avons 5 classes, nous avons utilisé l'approche $1-$contre-1 avec le critére de vote majoritaire pour la sélection de la classe finale.

4.4 Reconnaissance du type d'empreinte par SVM

Pour application la classification par SVM, nous avons besoin de créer une base de données dédiée à l'apprentissage et une autre au test. Concernant l'apprentissage, nous avons testé quatre taux (1%, 30%, 50%, 90%). Ces taux définissent le pourcentage de la base de données utilisé pour l'apprentissage du type d'empreintes ; le reste étant utilisé pour le test de reconnaissance. Pour chaque taux d'apprentissage, nous obtenons un taux de reconnaissance du Type et nous pouvons ainsi observer son évolution. Pour tous les taux d'apprentissage, nous avons calculé 10 fois le taux de reconnaissance et fait la moyenne pour faire de la validation croisée.

4.5 Résultats

Pour la reconnaissance de type, nous avons testé différentes valeurs du niveau de quantification lors de la construction des histogrammes (8, 16, 32 and 64) afin de comparer les taux de reconnaissance avec différentes précisions. Pour la première étape, nous regardons seulement la reconnaissance aux quatre taux avec 50% d'apprentissage sur la base de données. Les résultats sont présentés dans le tableau 3.

Nbr de classe par attribut	Taux de reconnaissance (%)
8	79.54
16	79.79
32	78.91
64	76.67

Tableau 3 – *Tableau sur la reconnaissance du type d'empreinte avec un taux d'apprentissage de 50%*

Nous pouvons observer que le meilleur taux de reconnaissance correspond à l'utilisation de 16 niveaux de quantification pour les histogrammes. Nous expliquons ce résultats par le fait que l'utilisation de 64 niveaux engendre une résolution trop haute ce qui induit de nombreuses valeurs à zéro. Avec seulement 66 informations ($4 \times 16 + 2$), nous obtenons un taux de reconnaissance du type d'empreinte de 79% avec des paramètres standards pour le SVM .
Le Tableau 4 montre l'évolution du taux de reconnaissance du Type d'empreinte pour les quatre valeurs d'apprentissage sur la base de données. Nous pouvons observer que le taux de reconnaissance est vraiment très proche entre 30% et 90% d'apprentissage (moins de 3%) et atteint un taux de 82% de reconnaissance sur la base de test.
Ce taux est d'autant très bon qu'il a été obtenu en calculant les paramètres sur le seul template de minuties de l'empreinte digitales sans aucun accès à l'image d'origine associée. A titre de comparaison, Jain *et al.* ont développé une méthode de reconnaissance du type de l'empreinte en s'appuyant sur l'image d"origine et obtiennent un taux de reconnaissance de 89% [15]. Cela montre que l'approche retenue est prometteuse.

Taux d'apprentissage(%)	Taux de reconnaissance(%)
1	43.13
30	78.70
50	79.79
90	81.45

Tableau 4 – *Tableau contenant le taux de reconnaissance du type d'empreinte pour différents ratios d'apprentissage.*

5 Conclusion et perspectives

Dans ce papier, nous avons présenté un vecteur d'attributs nous permettant de décrire un template de minuties. Ces attributs sont extraits de la triangulation de Delaunay. Nous avons montré que les bases de données FVC et SFinge sont uniformes sur tous les attributs. Nous avons étudié la corrélation entre tous les critères et nous avons défini que seuls quelques-uns sont très corrélés avec le Type d'empreintes digitales. Nous avons étudié avec succès la reconnaissance du Type d'empreintes digitales. Nous avons obtenu environ 80% de taux de reconnaissance quand le tiers de la base de donnée est utilisé pour l'apprentissage. Ces premiers résultats sont très prometteurs pour poursuivre notre recherche sur l'utilisation de la triangulation de Delaunay pour aider à la caractérisation des minuties d'empreintes digitales.

Références

[1] ISO/IEC 19794-2. information technology - biometric data interchange format format - part 2 : Finger minutiae data, 2011.

[2] 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.

[3] Qinzhi Zhang and Hong Yan. Fingerprint classification based on extraction and analysis of singularities and pseudo ridges. *Pattern Recognition*, 37(11) :2233–2243, 2004.

[4] Franz Aurenhammer. *ACM Computing Surveys (CSUR)*, 23(3) :345–405, 1991.

[5] Peter Su and Robert L Scot Drysdale. A comparison of sequential delaunay triangulation algorithms. In *Proceedings of the eleventh annual symposium on Computational geometry*, pages 61–70. ACM, 1995.

[6] Jonathan Richard Shewchuk. Delaunay refinement algorithms for triangular mesh generation. *Computational geometry*, 22(1) :21–74, 2002.

[7] M Gopi, Shankar Krishnan, and Cláudio T Silva. Surface reconstruction based on lower dimensional localized delaunay triangulation. In *Computer Graphics Forum*, volume 19, pages 467–478, 2000.

[8] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[9] Fvc2002. http ://bias.csr.unibo.it/fvc2002/download.asp.

[10] Fvc2004. http ://bias.csr.unibo.it/fvc2004/databases.asp.

[11] C. I. Watson, M. D. Garris, E. Tabassi, C. L. Wilson, R. M. Mccabe, S. Janet, and K. Ko. Technical report, NIST, 2007.

[12] 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.

[13] Sfinge software. http ://biolab.csr.unibo.it/sfinge.html.

[14] V. N. Vapnik. *Statistical Learning Theory*. New York, wiley edition, 1998.

[15] Anil K Jain, Salil Prabhakar, , and Lin Hong. A multichannel approach to fingerprint classification. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, volume 24, pages 248–359, 1999.

Recherche Rapide du Mode de Prédiction Optimal basée Apprentissage via une Comparaison des Modes Intra/Inter en H.264/AVC

M. Bichon[†,‡] J. Le Tanou[†] W. Hamidouche[‡]

[†] Envivio France

[‡] INSA de Rennes, Institut d'Électronique et de Télécommunications de Rennes - UMR6164

maxime.bichon@ericsson.com, julien.le.tanou@ericsson.com, wassim.hamidouche@insa-rennes.fr

Résumé

Dans un contexte de codage vidéo hybride comme définit par la norme H.264/AVC, l'efficacité maximale de compression est atteinte par l'optimisation d'un ensemble de paramètres de codage. Parmi les plus cruciaux figure le choix du mode de prédiction. Ce dernier est communément sélectionné suivant un critère débit-distorsion RDO [1] *au cours d'une recherche exhaustive, d'une complexité de calcul importante notamment due à l'étape d'estimation de mouvement (ME). La solution proposée ici adapte un algorithme de décision rapide basé analyse d'image, de manière à limiter la complexité introduite par la ME. En effet, la mise en concurrence des corrélations spatiales et temporelles relatives au macrobloc permet une décision rapide et fiable du mode de prédiction, évitant les MEs inutiles et coûteuses en calcul. Basé sur un apprentissage hors-ligne, le modèle proposé dans cet article permet d'accélérer systématiquement l'encodeur temps-réel x264, de 11,29% en moyenne et jusqu'à 20% selon les séquences, pour une perte en efficacité de codage négligeable (<1%).*

Mots clefs

Choix de Mode Rapide, Apprentissage, Corrélation Spatio-Temporelle, H.264/AVC.

1 Introduction

La norme H.264/AVC [1, 2] est le standard de compression vidéo majoritairement déployé sur les services vidéos. Communément, la sélection du mode optimal du macrobloc (MB) se fait par une recherche exhaustive complexe basée sur l'évaluation d'un critère débit-distorsion (*RDO*[3]). L'objectif est de minimiser une distorsion D sous une contrainte de débit $R \leq R_t$. Cette minimisation sous contrainte se traduit par le calcul du coût débit-distorsion (coût R-D), noté J, que l'on cherche à minimiser:

$$J = D + \lambda R \quad (1)$$

avec λ le multiplicateur de Lagrange [4] permettant de réécrire le problème sans contrainte. Le coût R-D réel de chaque mode s'obtient par un processus de reconstruction, étape de l'encodage qui simule le décodage afin d'estimer précisément le coût R-D. Certaines méthodes accélèrent cette étape grâce à des métriques de distorsion sous-optimales telles que la SAD [2] ou SATD [3], mais entraînent un biais due à l'approximation de la distorsion réelle. Une solution plus souple consiste à introduire une hiérarchie entre les modes. L'analyse *RDO* d'un sous-ensemble de modes, construit selon un critère robuste, est théoriquement plus rapide au détriment de pertes R-D négligeables. Trois catégories d'algorithmes de décision rapide utilisant cette stratégie se distinguent:

1. L'accélération de l'analyse des modes Intra a été largement traitée dans la littérature [5, 6]. Les algorithmes se basent généralement sur les caractéristiques spatiales du MB telles que le gradient [5] ou la variance de ses pixels [6]. Les gains, généralement faibles, s'expliquent par une analyse des modes Inter beaucoup plus complexe à cause de l'estimation de mouvement (ME). Cette étape, occupe à elle seule jusqu'à 30% du temps d'encodage total.

2. L'accélération de l'analyse Inter est populaire parmi les algorithmes de décision rapide en raison de la complexité et de l'efficacité de codage de la ME. [7, 8] proposent deux accélérations possibles de l'estimation elle-même. En limitant les tailles de partitions à analyser et donc l'usage de la ME, la complexité peut aussi être réduite de manière non négligeable, au détriment de l'efficacité R-D. Deux exemples définissent une hiérarchie des partitions Inter, préconisant une analyse plus fine, en fonction des Coded Block Pattern CBP [9] ou d'une pseudo-ME [10].

3. L'accélération des analyses Intra/Inter, consiste en un algorithme de décision rapide sur l'ensemble de la recherche exhaustive. [11] propose une solution où les modes sont hiérarchisés en fonction de l'activité temporelle. Une approche innovante [12] considère les modes Intra et Inter comme deux classes distinctes. En mesurant les corrélations spatiales et temporelles relatives au MB courant, les

1. Rate Distortion Optimization

2. Sum of Absolute Differences

3. Sum of Absolute Transformed Differences

auteurs utilisent la notion de risque Bayésien pour éliminer avec fiabilité l'ensemble des modes Intra ou Inter des choix possibles. Cependant, le modèle repose sur le choix de seuils fixe, méthode jugée trop rigide.
Dans cet article, nous proposons une évolution de [12] en un modèle simple et flexible. L'objectif est de palier aux cas où la ME est un "poids mort" pour l'encodage, en particulier dans le cas de séquences à mouvements complexes, p.ex. *Touchdown Pass* et *Pedestrian Area*. La suite de l'article est structurée de la manière suivante. Une rapide revue de l'algorithme initialement proposé est donnée dans la Section 2. Les nouveaux critères de décisions seront présentés dans la Section 3. La Section 4 fournit les explications justifiant la flexibilité du nouveau modèle. Les simulations et résultats seront discutés dans la Section 5, avant de conclure sur cette étude dans la Section 6.

2 Choix de mode Intra versus Inter basé analyse d'image

Le principe de [12] est d'estimer en amont de la *RDO* l'efficacité des modes Intra et Inter. Le calcul d'un descripteur de bloc, composé de trois critères robustes f_{intra}, f_{inter} et $|MV|$, permet d'estimer si l'ensemble des modes Intra ou Inter peuvent être éliminés des candidats potentiels, selon une notion de risque Bayésien. Le domaine de représentation du descripteur définit trois zones:

- La zone "sans risque": Les modes Intra ou Inter peuvent être ignorés de la *RDO* avec un impact faible sur l'efficacité de codage
- La zone "à risque tolérable": Les pertes R-D attendues du mode qui minimise le risque Bayésien sont acceptables, donc les modes Intra ou Inter sont ignorés de la *RDO*
- La zone "à risque intolérable": Les pertes attendues sont trop élevées, il faut appliquer la *RDO* sur l'ensemble de ces modes

La délimitation de ces zones se fait par un apprentissage supervisé hors-ligne, utilisant une base de MBs encodées en RDO, afin d'anticiper les pertes R-D en cas de mauvaise décision. De plus, l'estimation du risque Bayésien nécessite également un apprentissage au préalable. f_{intra} est une mesure de l'amplitude de la corrélation spatiale, calculée comme suit:

$$f_{intra} = \sum_{k=1}^{16} \min_{m} \{SATD(m_k)\} \quad (2)$$

avec k l'indice du bloc 4x4 du MB courant, m le mode de prédiction à déterminer parmi cinq des neuf disponibles en H.264/AVC et $SATD(.)$ la mesure de distorsion, définie comme une erreur de prédiction dans le domaine transformé.
f_{inter} mesure l'amplitude de la corrélation temporelle. Afin d'être comparable à f_{intra}, elle se mesure par une SATD résultant d'une ME rapide "$MVFAST$" [13].

$$f_{inter} = \min_{MV} \{SATD(m_{MV})\} \quad (3)$$

avec MV le vecteur de mouvement cherché selon [13] et m_{MV} la compensation en mouvement dépendant du vecteur MV. La comparaison des deux variables détermine si le MB est plus fortement impacté par la corrélation spatiale ou temporelle et, a fortiori, quels modes sont à prioriser entre l'Intra et l'Inter.
La variable $|MV|$ concerne l'amplitude du mouvement estimé par la ME. Elle est utilisée comme critère de confiance de la ME et de f_{inter}. Son utilisation comme troisième dimension permet un découpage plus fiable du domaine de représentation du descripteur.
Les modifications que nous proposons sont présentées dans la suite de cet article. Les changements effectués pour la prise de décision, c'est-à-dire les nouvelles mesures de corrélations et de précision, sont exposées dans la Section 3. Dans la Section 4, l'espace de représentation est redéfini, selon deux frontières ($Bounder_{intra}$ et $Bounder_{inter}$) dépendantes de plusieurs paramètres de flexibilité.

3 Nouveaux Critères

Les critères utilisés pour mesurer l'amplitude des corrélations sont discutables, notamment due au biais induit par certaines approximations. D'une part, l'utilisation de la *SATD* comme mesure de la corrélation ne tient pas compte de la distorsion réelle et du débit. D'autre part, les variables utilisées sont obtenues grâce à des analyses rapides (et potentiellement imprécises) des modes Intra et Inter. Nous proposons ici plusieurs modifications des critères précédemment utilisés.

3.1 Corrélations Spatiale et Temporelle

Le calcul de la variable f_{intra} est réalisé par une analyse *RDO* du mode Intra 16x16, l'équation (2) devient alors:

$$f_{intra} = \min_{m} J_{SSD}(m) = D_{SSD}(m) + \lambda R_{SSD}(m) \quad (4)$$

avec J_{SSD}, D_{SSD} et R_{SSD} correspondant au coût R-D réel, à la distorsion et au débit associé au mode m. En remplaçant la *SATD* initiale par une distorsion mesurée sur le MB reconstruit et en se basant sur le coût R-D plutôt que sur la distorsion, le critère f_{intra} est une mesure plus fiable de la corrélation spatiale.
La variable f_{inter} est définie comme le coût R-D d'une analyse *RDO* du mode Inter 16x16.

$$f_{inter} = \min_{MV} J_{SSD}(MV) = D_{SSD}(MV) + \lambda R_{SSD}(MV). \quad (5)$$

Le principal changement apporté par ces deux nouvelles variables est la prise en compte de la distorsion réelle et du débit réel. f_{intra} et f_{inter} sont respectivement des mesures plus fiables de l'efficacité de codage des modes Intra et Inter, suivant un critère R-D.

3.2 Paramètres de Précision

Afin de contrôler la précision du modèle, l'algorithme initial utilise l'amplitude du mouvement $|MV|$. Néanmoins, dans le cas d'un mouvement important mais uniforme, la ME peut être précise malgré une forte amplitude. En effet, l'algorithme de Block-Matching utilisé pour la ME recherche le bloc le plus similaire dans une fenêtre de recherche autour d'une "graine". Classiquement, cette graine correspond simplement à la position du bloc courant dans l'image de référence, mais dans l'encodeur x264 elle correspond à un vecteur de mouvement prédit calculé comme suit:

$$pMV(D) = Median[MV(A), MV(B), MV(C)] \quad (6)$$

avec $pMV(.)$ la fonction estimant la graine pour la ME, $Median[.]$ le vecteur médian parmi ceux proposés et $MV(.)$ le vecteur mouvement utilisé par le MB concerné. Le voisinage (A, B et C) utilisé pour ce calcul correspond respectivement aux MBs gauche, supérieur, et supérieur-droite du MB courant D. Sur la Figure 1, on peut voir la fenêtre obtenue avec ou sans graine.

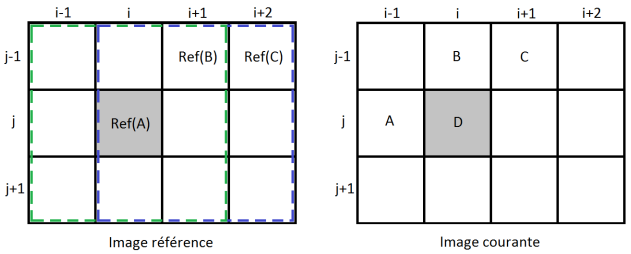


Figure 1 – *Fenêtres de recherche de la ME avec vecteur prédit (zone bleue) et sans (zone verte)*

L'intérêt de ce vecteur prédit est une convergence plus rapide du résultat optimal de la ME. De plus, cette prédiction est aussi utilisée dans la norme H.264/AVC afin d'obtenir un gain, en termes de débit, en ne transmettant que l'écart MVD entre le vecteur médian et le vecteur final. Nous proposons d'utiliser $|MVD|$ au lieu de $|MV|$ comme mesure de confiance de la ME.

4 Frontières Flexibles

Les différentes zones proposées dans l'article initial recouvrent de manière non-uniforme l'espace de représentation du descripteur partitionné en plusieurs cellules. Le découpage de cet espace se fait en fonction des données d'apprentissage qui servent également à évaluer les pertes R-D envisagées en cas d'erreur pour chaque cellule. Les cellules dont les pertes R-D estimées sont inférieures à 0.5% appartiennent à la zone sans risque. Ce choix de seuil arbitraire implique une certaine rigidité dans le modèle. Nous proposons à la place de calculer deux frontières *flexibles*, $Bounder_{intra}$ et $Bounder_{inter}$, définies ci-après par les équations (8) et (9). Quatre observations sont faites et intégrées au modèle pour le rendre adaptatif:

1. On observe que le coût R-D est une valeur qui varie en fonction des caractéristiques de la source. Un MB sans texture et avec peu de mouvement aura un coût R-D relativement faible. Cela est due au fait qu'une forte corrélation spatiale et temporelle entraîne une prédiction de très bonne qualité. Inversement, si un MB est fortement texturé ou affecté par un mouvement complexe, la prédiction sera de mauvaise qualité et le coût R-D élevé. Une variable intermédiaire est donc calculée:

$$df_{ratio} = \frac{|f_{intra} - f_{inter}|}{MAX(f_{intra}, f_{inter})} \quad (7)$$

où df_{ratio} est une valeur comprise dans l'intervalle $[0, 1[$ qui estime l'importance de l'écart entre les deux corrélations.

2. Remarqué par [12] lors des expérimentations, commettre une erreur en choisissant de privilégier l'Intra plutôt que l'Inter est plus critique que l'inverse. Afin de prendre en compte cet effet, deux frontières $Bounder_{intra}$ et $Bounder_{inter}$ sont définies. La variable df_{ratio} présentée ci-dessus sera donc comparée, pour chaque MB, à la borne correspondant à la classe de plus faible coût.

3. Expliquée dans la section 3.2, la précision de la variable f_{inter} est dépendante des résultats de la ME. La variable $|MVD|$ est introduite dans les calculs de frontière afin de tenir compte de cette précision. Basés sur une étude expérimentale, nous proposons les modèles suivants:

$$Bounder_{intra} = a * ln(|MVD|) + b. \quad (8)$$

$$Bounder_{inter} = c * ln(|MVD|) + d. \quad (9)$$

Les constantes a, b, c et d sont estimées par un apprentissage hors-ligne et sont différentes selon le type d'image auxquels le modèle est appliqué: P, B ou B référence.

4. Un paramètre de quantification (QP) élevé entraîne une distorsion plus élevée mais un débit plus faible et peut donc modifier les coûts R-D d'une manière non négligeable. La valeur de QP est donc introduite dans le modèle. Au travers d'observations, les relations suivantes ont pu être établies entre les différentes constantes des modèles précédemment introduites (a, b, c et d) et QP:

$$a = \alpha_0 QP^2 + \beta_0 QP + \gamma_0. \quad (10)$$

$$b = \alpha_1 QP^2 + \beta_1 QP + \gamma_1. \quad (11)$$

$$c = \alpha_2 QP + \beta_2. \quad (12)$$

$$d = \alpha_3 QP + \beta_3. \quad (13)$$

Quand la ME est suffisamment précise, c.à.d. $|MVD| = 0$, les équations (8) et (9) sont remplacées par deux fonctions affines de QP. On visualise sur la Figure 2 que si df_{ratio} est supérieure (zone verte) à la frontière calculée, alors la précision du modèle est suffisante pour éliminer un ensemble de mode. Ces deux frontières sont vues comme

"flottantes" dans l'intervalle de df_{ratio}, selon le couple ($|MVD|$,QP) et les équations (8∼13).

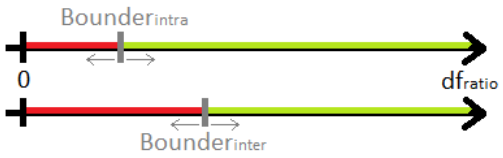


Figure 2 – *Frontières en fonction de* df_{ratio}

5 Expériences et Résultats

5.1 Environnement Expérimental

L'encodeur de référence JM [14], associé à la norme H.264/AVC, permet la comparaison d'outils sur une base commune. L'importante complexité de calcul du JM en fait une plateforme non-réaliste pour le développement et l'évaluation d'algorithmes d'accélération. L'efficacité et la stabilité reconnues de l'encodeur x264 [15] en font, a contrario, une plateforme cohérente pour ces travaux.
Nous utilisons la configuration suivante: une I par seconde; 3 B hiérarchique; la ME a une précision au 1/4 de pixel, recherche hexagonale; aucune partition Inter inférieure au 8x8. Le x264 est configuré en *RDO*.
Un ensemble de séquences aux formats 4K, 1080p et 720p, sont utilisées pour évaluer l'algorithme. Trois séquences 720p non-standards sont ajoutées pour enrichir l'ensemble de test: *Danse*, *PubSony* et *Home*. Des images de ces séquences sont données Figure 3.

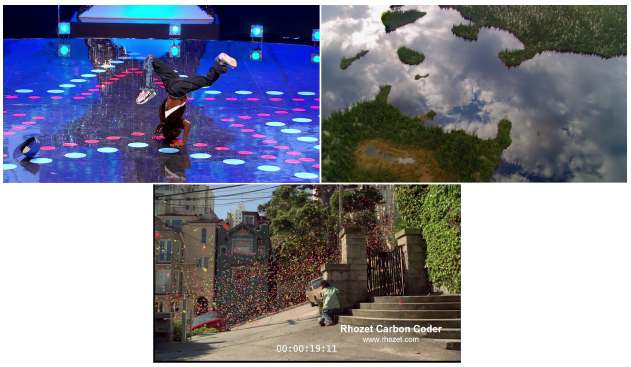


Figure 3 – *En haut Danse(720p) et Home(720p); en bas PubSony(720p)*

Les trois séquences 4K (4096x2160), *SquareTimelapse*, *TunnelFlag* et *ToddlerFountain*, sont fournies par Netflix et publiquement disponible [16]. Les séquences *Cactus*, *MobileCalendar* et *Home* sont définies comme ayant des mouvements simples à prédire. A contrario, les séquences restantes ont des mouvements complexes.
L'évaluation de l'algorithme se fait selon deux critères. Le premier est la réduction relative de complexité donnée par:

$$T_{\%} = \frac{Time(x264_{rapide}) - Time(x264_{original})}{Time(x264_{original})} \quad (14)$$

avec $T_{\%}$ le temps relatif, $Time(.)$ le temps d?encodage et $x264_{original}$ et $x264_{rapide}$, respectivement le x264 avec et sans l?algorithme que nous proposons.
Le second critère représente les pertes en efficacité de codage par la mesure de Bjontegaard [17], qui consiste a mesurer la différence d'aire entre deux courbes R-D. Elle est ici basée sur une métrique PSNR pour cinq valeurs de QP (25,28,31,34,37). Les scores négatifs représentent une économie de débit et donc une amélioration de l'efficacité de codage.

5.2 Résultats

Le Tableau 1 montre la réduction de complexité induite par l'algorithme. Les résultats sont beaucoup plus probants sur les séquences dont la complexité du mouvement est plus forte. Les séquences *Cactus*, *MobileCalendar* et *Home* avec des mouvements simples présentent les gains les plus faibles (<6%). En effet pour ce type de séquence, la ME converge très rapidement vers le résultat optimal, la réduction potentielle de complexité étant ainsi faible.

Séquences de test		QP			
		25	31	37	Moy
720p	Danse	-15.57%	-19.09%	-21.54%	-18.73%
	PubSony	-6.93%	-9.50%	-10.8%	-9.08%
	MobileCalendar	-1.72%	-3.03%	-4.81%	-3.19%
	Home	-4.51%	-5.62%	-6.16%	-5.43%
	Moyenne	**-7.18%**	**-9.31%**	**-10.83%**	**-9.11%**
1080p	PedestrianArea	-12.37%	-14.22%	-15.24%	-13.94%
	TouchdownPass	-10.33%	-10.58%	-10.9%	-10.60%
	RedKayak	-18.24%	-21.53%	-21.32%	-20.36%
	Cactus	-4.16%	-5.61%	-5.67%	-5.15%
	Moyenne	**-11.28%**	**-12.99%**	**-13.28%**	**-12.51%**
4K	SquareTimelapse	-10.77%	-14.89%	-16.55%	-14.07%
	TunnelFlag	-7.68%	-16.94%	-20.66%	-15.09%
	ToddlerFountain	-4.28%	-9.5%	-11.73%	-8.5%
	Moyenne	**-7.58%**	**-13.78%**	**-16.31%**	**-12.56%**
Tous	**Moyenne**	**-8.78%**	**-11.86%**	**-13.22%**	**-11.29%**
	Maximum	**-18.24%**	**-21.53%**	**-21.54%**	**-20.36%**
	Minimum	**-1.72%**	**-3.03%**	**-4.81%**	**-3.19%**

Tableau 1 – *Complexité relative au x264*

Avec une réduction de complexité maximale de 21.54% (pour *RedKayak*) et une moyenne de 11.29% , l'algorithme réponds aux objectifs. Les gains maximums sont atteints pour les séquences jugées difficiles pour la ME. Les gains minimums sont observés sur *MobileCalendar* avec une réduction moyenne de complexité de 3.19%. Autre point remarquable, les gains tendent à augmenter pour un QP élevé, s'expliquant par l'augmentation de la complexité des décisions relativement à la complexité de codage de l'information résiduelle.
Les résultats en efficacité de codage sont présentés dans le Tableau 2. Les pertes R-D sont minimes (0.91% dans le pire cas) pour l'ensemble des séquences testées.

Séquences de test		BD-Rate
720p	Danse	0.83%
	PubSony	0.36%
	MobileCalendar	0.18%
	Home	0.26%
	Moyenne	**0.41%**
1080p	PedestrianArea	0.91%
	TouchdownPass	0.71%
	RedKayak	0.35%
	Cactus	0.39%
	Moyenne	**0.61%**
4K	SquareTimelapse	-0.3%
	TunnelFlag	-0.66%
	ToddlerFountain	-0.01%
	Moyenne	**-0.32%**

Tableau 2 – *BD-Rate en comparaison au x264*

6 Conclusion

La solution que nous proposons permet de définir des seuils de décision de manière adaptative ou *flexible*, plutôt qu'arbitrairement comme proposé par [12]. Les résultats sont satisfaisants au vu de l'objectif initial, qui était d'accélérer les décisions des modes de codage dans un contexte *RDO*, en particulier en compensant la complexité d'une ME mise en échec. Les gains systématiques observés, atteignent jusqu'à 20% sur certaines séquences, pour des pertes R-D négligeables (<1%). Ces gains dans un encodeur quasi temps-réel tel que le x264 démontrent l'efficacité de l'algorithme proposé. Néanmoins, certaines séquences peuvent présenter des gains limités, dans le cas où l'estimation de mouvement converge rapidement. L'approche proposée peut-être efficacement couplée avec d'autres algorithmes de décisions rapides des modes Intra ou Inter [5, 9]. Enfin, les hypothèses avancées lors de la conception de l'algorithme sont valables en HEVC [18], la norme qui succède à H.264/AVC. Son adaptation au standard suivant est donc la prochaine étape à considérer.

References

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, et A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, Juillet 2003.

[2] I. E. Richardson. *The H.264 Advanced Video Compression Standard*. Wiley publications, 2010.

[3] G. J. Sullivan et T. Wiegand. Rate-distortion optimization for video compression. *Signal Processing Magazine, IEEE*, 15(6):74–90, Novembre 1998.

[4] T. Wiegand et B. Girod. Lagrange multiplier selection in hybrid video coder control. Dans *Image Processing, 2001. Proceedings. 2001 International Conference on, IEEE*, pages 542–545, Thessaloniki, Octobre 2001.

[5] A. Elyousfi. An improved fast mode decision method for H.264/AVC intracoding. *Advances in Multimedia*, 2014(7):1–8, Janvier 2014.

[6] Y. H. Huang, T. S. Ou, et H. H. Chen. Fast decision of block size, prediction mode and intra block for H.264 intra prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(8):1122–1132, Juillet 2010.

[7] A. M. Tourapis, O. C. Au, et M. L. Liou. Predictive motion vector field adaptive search technique (pmvfast): enhancing block-based motion estimation. Dans *Proc. Visual Communications and Image Processing 2001*, pages 883–892, 2001.

[8] I. Ahmad, W. Zheng, J. Luo, et M. Liou. A fast adaptive motion estimation algorithm. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(3):420–438, Mars 2006.

[9] Z. Shi, W. A. C. Fernando, et A. M. Kondoz. A hybrid coded block patterns based fast mode decision in H.264/AVC. Dans *Multimedia and Expo Workshops, 2012 IEEE International Conference on*, pages 13–18, Melbourne, VIC, Juillet 2012.

[10] W. Zhang, Y. Huang, et J. Peng. Fast mode decision for H.264/AVC based on local spatio-temporal coherency. Dans *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific, IEEE*, pages 1–6, Kaohsiung, Octobre 2013.

[11] H. Zeng, C. Cai, et K.K. Ma. Fast mode decision for H.264/AVC based on macroblock motion activity. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(4):491–499, Avril 2009.

[12] C. Kim et C.-C. J. Kuo. Feature-based intra-/intercoding mode selection for H.264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(4):441–453, Avril 2007.

[13] P.I. Hosur et K.K. Ma. Motion vector field adaptive fast motion estimation. Dans *Second International Conference on Information, Communications and Signal Processing (ICICS)*, pages 7–10, Singapore, Décembre 1999.

[14] ITU-T et ISO/IEC JTC. Reference software for advanced video coding. https://www.itu.int/rec/T-REC-H.264.2, Juin 2015.

[15] VideoLAN. x264 software revision 2495. http://www.videolan.org/developers/x264.html, Avril 2014.

[16] Video Test Media [derf's collection]. https://media.xiph.org/video/derf/, Février 2016.

[17] G. Bjontegaard. Calculation of average PSNR differences between RD-curves. Dans *ITU-T VCEG, Texas, USA, Proposal VCEG-M33*, Austin, Texas, Avril 2001.

[18] High efficiency video coding. Dans *Rec. ITU-T H.265 and ISO/IEC 23008-2*, Sapporo, JP, Janvier 2013.

IDENTIFICATION DE SCANNERS X À PARTIR D'EMPREINTES DE CAPTEURS

Anas Kharboutly[*]*, William Puech*[*]*, Gérard Subsol*[*]*, Denis Hoa*[†]

[*] LIRMM, Université de Montpellier/CNRS, Montpellier, France
[†] IMAIOS, MIBI, Montpellier, France

Résumé

Les scanners X produisent une énorme quantité d'images médicales. Dans cet article, nous présentons un nouveau défi dans le domaine de la criminalistique des images médicales, particulièrement sur l'identification des scanners X, basée sur le bruit de capteurs de scanners X. Nous avons construit un modèle de bruit de référence (RPN), puis, nous utilisons ce RPN comme une empreinte unique pour identifier un scanner X. Grâce aux propriétés des images médicales, nous proposons un nouveau concept d'extraction d'empreintes et nous fournissons trois couches d'empreintes. Nous séparons les images médicales en trois couches : l'air, les tissus et les os. Puis à partir de chaque couche, nous construisons une empreinte unique pour le scanner X. Nous testons après la présence de cette empreinte dans une nouvelle image pour l'authentifier. La méthode proposée est robuste et donne une grande précision d'identification en utilisant le concept des trois niveaux RPN.

Mots clefs

Criminalistique des images médicales, identification d'appareils, bruit de capteurs, empreintes de capteurs

1. INTRODUCTION

Le scanners X créent des images médicales 3D [1]. Ces images fournissent une représentation de haute qualité pour n'importe quelle partie anatomique du corps. Les images de scanners X sont stockées dans un format standard des images médicales, le DICOM [2]. Les Fichiers DICOM sont composés de deux parties : les méta-données et les données images. Les méta-données contiennent toutes les informations du patient et le contenu de l'image. En leur absence, ou si les fichiers des méta-données sont modifiés, nous ne pouvons pas authentifier leur information ou le contenu de l'image. Notre objectif est d'identifier l'acquisition d'un scanner X à partir du contenu de l'image sans utiliser les méta-données. La criminalistique numérique est la technologie utilisée pour accomplir cet objectif [3].

Comme pour tous les systèmes d'imagerie, la production de bruit est toujours un sujet de discussion. La réduction de bruit est le premier besoin de tous les utilisateurs finaux d'un système d'imagerie. Mais pour nous, l'objectif est d'analyser le bruit. Dans les différents appareils d'imagerie, l'analyse du bruit du capteur est la technique de base de la criminalistique numérique. Il a été proposé pour la première fois par [4] pour l'identification des appareils numériques. Ils ont fondé leur étude sur la photo-réponse non-uniforme (PRNU) qui est un bruit multiplicatif. Ils ont estimé une empreinte digitale de l'appareil photo numérique (APN) à partir d'un ensemble d'images. Puis, ils ont détecté sa présence dans toutes les nouvelles images par corrélation. Dans ce contexte, des nombreux travaux qui se basent sur le PRNU existent déjà [5, 6, 7]. Une amélioration des méthodes de PRNU ont été proposées dans [8, 9], ils ont essayé d'améliorer le taux d'identification de l'appareil et ont réduit l'effet des détails du contenu de l'image. De plus, des améliorations ont été proposées par [10, 11], ils ont fourni un facteur de pondération comme un moyen d'éliminer certaines informations dans les hautes fréquences. Tous ces procédés sont limités à l'image numérique et l'identification d'APN.

Dans le domaine médical, nous ne trouvons pas encore ce genre d'études sur l'identification d'appareils. Dans [12], les auteurs comparent les propriétés de bruit entre deux scanners X de différents fabricants. Une méthode d'identification d'appareil médical a été proposée par [13], mais il concerne les primitives des images 2D en radiographie à rayons X. En ce qui concerne l'identification du scanner X, nous présentons dans [14] une première analyse du problème du scanner X. Nous avons créé une méthode dérivée de celle présentée dans [4] pour extraire l'empreinte du scanner X puis détecter sa présence dans une nouvelle image par corrélation. La limitation de base de l'application de ces méthodes sur les images de scanner X est la grande variation entre les valeurs de l'image. L'inhomogénéité de l'image continue laisse des traces dans les hautes fréquences et il est difficile de l'isoler du bruit. Dans [15], nous avons proposé une amélioration en masquant les traces dans les hautes fréquences. Deux limitations ont été découvertes, un grand nombre d'informations à hautes fréquences a été supprimé, ainsi que la diminution de la précision de l'identification des données réelles. Dans [16], un travail sur l'identification du scanner X est proposé, ils sont capables de séparer les scanners en se basant sur leur

algorithme de reconstruction. Ils caractérisent le bruit radial généré par le scanner X dans une empreinte, qui représente la corrélation entre les 180 projections des bruits composantes et ses moyennes. Le vecteur résultat de l'empreinte est présenté en entrée d'un SVM. Ils ont fondé leur solution sur [4] pour extraire la composante de bruit. La limitation de base est l'extraction du bruit en utilisant la méthode classique de [4], qui va laisser des traces à hautes fréquences en plus du bruit, comme nous l'expliquions dans [14].

Dans cet article, nous présentons un nouveau concept d'extraction d'empreinte dans des images médicales. Nous séparons trois couches homogènes des images du scanner X : la couche os, la couche tissus et la couche air. Chaque scanner X possède donc trois couches d'empreintes. Ensuite, pour tester une nouvelle image, nous vérifions la présence de l'empreinte du scanner X dans chacune de ses couches par corrélation. Contrairement à la méthode PRNU améliorée [15], nous utilisons les informations à hautes fréquences dans la méthode d'identification. Nous avons travaillé sur des images réelles de scanner X, et à partir d'une seule image en 2D, nous avons pu identifier la source du scanner X.

Le reste de ce papier est organisé comme suivent. Dans la section 2, nous présentons les préliminaires du bruit d'acquisition de scanners X. Dans la section 3, nous décrivons notre méthode proposée. Dans la section 4, nous présentons des résultats expérimentaux. Dans la section 5, nous concluons et proposons notre vision pour le futur.

2. COMPRÉHENSION PRÉLIMINAIRE

Le scanner X a un détecteur indirect. Il se compose de deux parties : un scintillateur et d'une photodiode au silicium. Le scintillateur est la partie qui absorbe les rayons X, les envoie de la source et les transmet au travers du patient. Il convertit les photons de rayons X reçus à la lumière. Cette lumière est ensuite recueillie par une photodiode de silicium, puis à son tour, convertit la lumière en un courant électrique. Enfin, le courant électrique est transmis à un convertisseur analogique/numérique [17]. Deux types de bruit existent dans le détecteur du scanner X, le bruit multiplicatif quantique et le bruit additif électronique. Le bruit quantique (Q) est le résultat d'un nombre aléatoire de photons envoyés par le tube à rayons X, et l'inhomogénéité du matériau de silicium de la photodiode. Le bruit électrique (EN) est ajouté au signal de l'image pendant l'étape de conversion analogique/numérique. Contrairement au modèle de bruit de l'appareil photo numérique [4], nous ne disposons pas d'un bruit de motif fixe (FPN), car le système d'acquisition CT expose tous les capteurs à la lumière lors de l'acquisition de l'image. La seule composante de bruit du modèle est donc la Photo-Réponse Non-Uniforme (PRNU). Elle est causée par le bruit Q et le bruit EN. Basé sur la classification de bruit du capteur dans [4, 7] et la simplification du concept de la méthode proposée, nous supposons que le bruit PN du détecteur est :

$$PN = FPN + PRNU, \tag{1}$$

et que la forme de l'image de sortie I du système d'acquisition du scanner X est :

$$I = I_0 + I_0.Q + EN, \tag{2}$$

$$PRNU = I_0.Q + EN, \tag{3}$$

où I_0 est l'image contenue sans bruit, Q est le facteur PNU quantique et EN est le bruit électrique supplémentaire. Le PNU sert une empreinte unique pour l'appareil d'acquisition. Ses propriétés en font une base pour l'identification des appareils [5, 18, 19] :

– son bruit est unique et les images du même appareil en héritent.
– la robustesse à différentes opérations de traitement, et sa stabilité au cours du temps.

Le système d'acquisition du scanner X garde toujours le bruit EN moins que le bruit Q associé avec seulement quelques rayons, donc nous ne pouvons pas compter sur les faibles fréquences de la composante EN, car elle ne caractérise pas le bruit du détecteur. Comme il est impossible d'accéder à la sortie bruitée des détecteurs du scanner X, nous proposons d'extraire la composante Q de l'image de sortie qui est reconstruite et de l'utiliser comme une empreinte du scanner X.

3. LA MÉTHODE PROPOSÉE

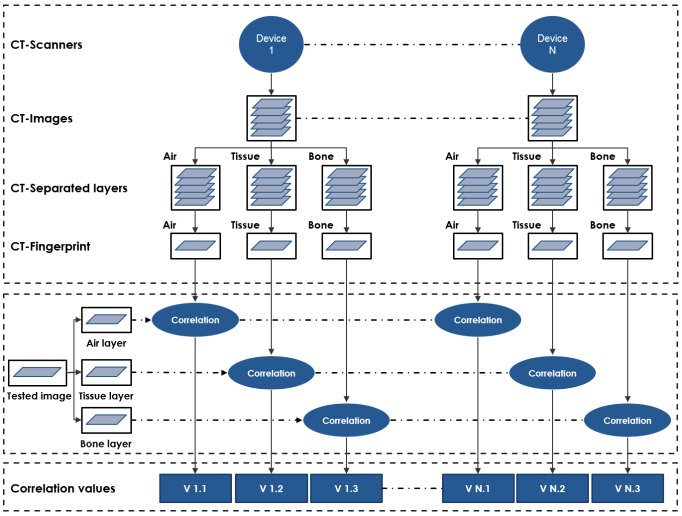


Fig. 1. Vue d'ensemble de la méthode.

Le cœur de notre méthode d'identification consiste à extraire l'empreinte du scanner X. Comme il est impossible d'accéder aux données à la sortie du détecteur du scanner X, on extrait les empreintes du scanner X à partir de ses images reconstruites. Nous construisons un RPN pour chaque scanner X. Ce RPN se compose de trois couches séparées qui représentent l'empreinte de ce scanner X. Pour construire le RPN, nous séparons d'abord les images étudiées en trois couches : 1. Couche d'air, 2. Couche de tissu et 3. Couche d'os. Cette séparation est appliquée en utilisant le seuillage. Ensuite,

nous extrayons le bruit de chaque couche pour construire une image 2D qui représente le RPN des couches. Enfin, une étape de correction de post-traitement est appliquée sur le RPN. Cette étape de correction est importante pour enlever toutes les traces des hautes fréquences qui persistent. La figure 1 illustre une vue d'ensemble de la méthode.

3.1. La séparation en couches

Nous travaillons avec des images médicales de scanner X, qui sont codées en 16 bits. L'idée est de séparer l'image originale du scanner X en couches, où l'intensité sera plus homogène. Cette séparation nous aide à éviter les variations des hautes fréquences qui faussent l'extraction et l'analyse de bruit. Contrairement aux images numériques, dans la plupart des images du scanner X, nous pouvons distinguer trois types de zones. Les zones de haute intensité qui sont les os ou les structures anatomiques au contraste renforcé, les zones de faible intensité qui contiennent des structures anatomiques rempli d'air comme les poumons et les zones de moyenne intensité qui correspondent aux tissus mous comme les muscles. Nous avons séparé chaque image du scanner X en trois couches en utilisant trois gammes d'intensité : l'air, le tissus et l'os. La séparation est appliquée en utilisant une technique de seuillage comme suit, après la proposition que I est une image du scanner X :

$$A(i,j) = I(i,j) | I(i,j) \in [a,b], \tag{4}$$

$$T(i,j) = I(i,j) | I(i,j) \in]b,c], \tag{5}$$

$$B(i,j) = I(i,j) | I(i,j) \in]c,d], \tag{6}$$

A est l'image de l'air, T est l'image des tissus et B est l'image de l'os. Cela sépare les intensités dans l'intervalle de valeur de l'air [a, b], de valeur des tissus gamme] b, c] et de valeur de l'os] c, d] respectivement. Comme dans les images CT on a une échelle calibrée d'intensité (Hounsfield) [20], on peut déterminer des seuils fixes pour certains structure anatomique comme l'os ou les tissus mous, comme présenté dans la section 4.

A partir de (2), et (4) nous pouvons étendre la forme de l'image du scanner X :

$$I = B + Q_B.B + T + Q_T.T + A + Q_A.A + EN. \tag{7}$$

avec $I_1 = B$, $I_2 = T$ et $I_3 = A$, on peut en déduire :

$$I = \sum_i (I_i + Q_i.I_i) + EN, \tag{8}$$

$i \in \{A, T, B\}$.

3.2. Extraction de bruit

Dans la méthode proposée, nous travaillons avec trois couches d'intensité : l'air, les tissus et les os. Cela signifie que nous pouvons extraire le bruit dans chaque couche séparément. Le point fondamental dans l'extraction de bruit est d'appliquer un filtre de débruitage $F()$ sur l'image I_i, puis extraire la composante de bruit par soustraction :

$$RN_i = I_i - F(I_i), \tag{9}$$

où $i \in \{A, T, B\}$, RNi est la composante de bruit pré-traitée. En plus du bruit, il existe des artefacts dans les basses fréquences. Ces basses fréquences sont le résultat du débruitage. Nous convertissons la couche originale dans un masque binaire B_i, puis, nous l'appliquons sur la composante de bruit pré-traitée pour extraire la composante de bruit finale n_i. L'application de ce masque supprime les faibles fréquences des artéfacts et conserve uniquement les informations de bruit :

$$n_i = B_i.RN_i. \tag{10}$$

Pour extraire la composante de bruit pré-traitée de chaque couche, nous utilisons un filtre de débruitage $F()$ qui est basé sur une transformation en ondelettes [21, 22].

3.3. Référence pattern de bruit

Pour construire un RPN pour une utilisation générale, nous sélectionnons un groupe d'images. Ensuite, nous extrayons leur composante de bruit. Enfin, nous appliquons un opérateur de moyennage. Dans la méthode proposée, nous travaillons sur trois couches séparées, nous devons donc garder ces trois couches séparées même dans l'extraction RPN, afin de construire un RPN différente de chaque couche :

1. Sélectionner un groupe d'images, ces images couvrent presque toutes les parties anatomiques du corps. La variation des images est un point important dans la construction du RPN, afin d'inclure toutes les variations de bruit concernant les différentes parties anatomiques.
2. Séparer ces images en trois couches, selon 3.1.
3. Extraire la composante de bruit de chaque image de chaque couche, selon 3.2.
4. Afin de renforcer la composante Q fixe du bruit de modèle de capteur et éliminer les traces de basses fréquences EN, nous appliquons une opération moyenne sur le groupe de chaque couche d'image, pour extraire un RPN de chacun :

$$RPN_i = \frac{1}{N_i} \sum_{h=1}^{N_i} n_{(i,h)}, \tag{11}$$

où $i \in \{A, T, B\}$, N_i est le nombre d'images utilisées dans la création du RPN et n_i est la composante de bruit.

Le RPN est une composante de bruit amélioré. En plus du bruit, il y a encore quelques artefacts des hautes fréquences qui ont été produits pendant le moyennage. Ces artefacts existent à cause de quelques traces supplémentaires de hautes fréquences. Si les artéfacts de hautes fréquences restent dans le RPN, cela affectera l'identification de l'appareil, il rend la discrimination du niveau de bruit au niveau de contenu anatomique.

Pour supprimer ces artefacts de hautes fréquences, nous passons dans le domaine fréquentiel par une transformation de

Fourier rapide. En utilisant un filtre passe-bas de Wiener, nous sommes en mesure d'éliminer tous les pics du signal RPN.

3.4. Les critères d'identification

Pour identifier le scanner X d'une image testée, la composante de bruit de l'image testée N_t doit être corrélée au RPN du scanner X. La corrélation entre N_t et RPN est mesurée par l'énergie de corrélation crête (PCE), qui est une métrique de similitude stable proposée par [7, 23], qui représente le rapport entre la hauteur du pic du plan de corrélation et son énergie totale, étant donné que le plan de corrélation est la corrélation croisée entre les deux signaux de bruit :

$$PCE(N_t, RPN_i) = \frac{E_p(N_t, RPN_i)}{E_{cp}(N_t, RPN_i)}, \qquad (12)$$

N_t est la composante de bruit de la couche testée, RPN_i est le bruit de pattern de référence, $i \in \{A, T, B\}$, E_p est la hauteur du pic de la corrélation plan et E_{cp} est l'énergie totale du plan de corrélation.
Nous calculons le PCE entre la composante de bruit de chaque couche séparée de l'image testée et le RPN de la même couche sur tous les appareils étudiés. Pour décider si une image appartient à un scanner X spécifique, il doit avoir la majorité de ses couches (dans notre méthode proposée, au moins 2) comme provenant du même dispositif.

4. RÉSULTATS EXPÉRIMENTAUX

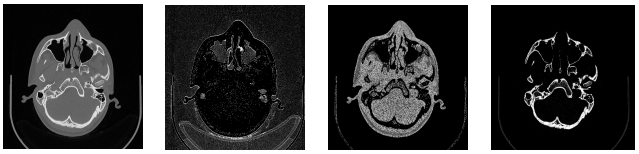
Nos expériences ont été appliquées sur 60 images 3D à partir de trois scanners différents (deux de Siemens S1, S2 et un de General Electric GE). Un nombre total de 20,939 images a été utilisé pour construire le RPN des trois appareils et de valider la précision de l'identification. Ces images couvrent presque toutes les parties anatomiques du corps. Tous ont les mêmes paramètres d'acquisition (faisceau d'énergie : (100 120 140) KV, la valeur Pitch : (0.5,1), Reconstruction : (soft, hard)). Le tableau 1 illustre les propriétés des images expérimentales.

	S1	**S2**	**GE**
Nb of 3D volumes	20	20	20
Nb of images	7572	7279	5088
Size (pixels)	512x512	512x512	512x512
Bits per pixel	16	16	16
Nb of images of RPN	3363	3756	2092
Nb of tested images	4209	4523	2996

Table 1. Caractéristiques des images expérimentales.

4.1. Pré-traitement de la base de données

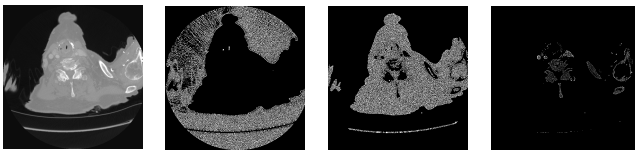
Pour préparer la base de données des images, nous avons séparé chaque image étudiée en trois couches. Les valeurs de la gamme de séparation ont été testées expérimentalement sur la base de 9000 images, où les valeurs de la gamme de l'air : [-990, -200], tissus :]-200, 200] et d'os :]200, 1500], correspondant à des valeurs standard pour l'échelle de Hounsfield. On pourrait aussi utiliser des methods automatique pour fixer les seuils. Toutes les valeurs de moins de -990 ou au-dessus du 1500 ont été ignorées, car elles contiennent des artefacts d'acquisition. La figure 2 illustre un exemple d'une image originale de scanner originale d'une tête et de ses trois couches séparées, issue d'un volume 3D acquis par un appareil Siemens. Ensuite, on extrait la composante de bruit de chaque



(a) Originale (b) Air (c) Tissue (d) Bone

Fig. 2. Image originale d'une tête et de ses trois couches.

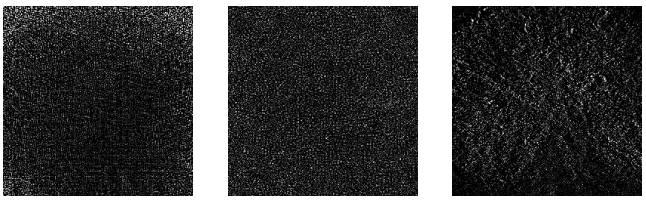
image comme nous l'avons expliqué dans la section 3.2. La figure. 3 illustre un exemple d'une image originale d'un thorax et le bruit de ses trois couches. Cette image a été acquise à partir d'un volume 3D par un appareil General Electric.



(a) Original (b) Air (c) Tissue (d) Bone

Fig. 3. Image originale d'un thorax et le bruit de ses trois couches.

Enfin, à partir de chaque appareil et chaque couche, nous sélectionnons un groupe d'images pour construire le RPN d'après (12). La figure, 4 illustre trois RPNs partiels de trois scanners utilisant différentes couches.

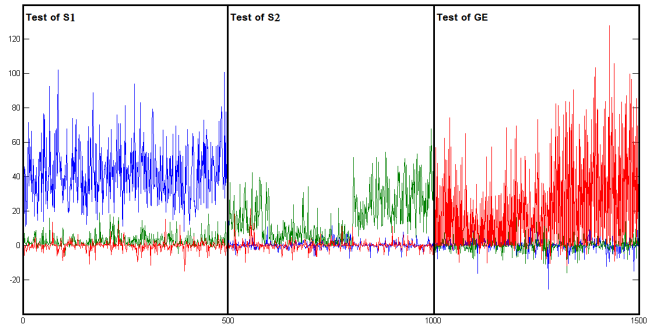


(a) S1 : Air RPN (b) S2 : Tissu RPN (c) GE : Os RPN

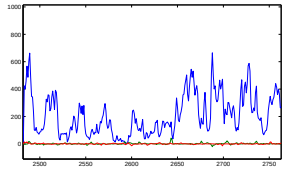
Fig. 4. RPNs de trois scanners différents pour différentes couches.

4.2. Les résultats quantitatifs

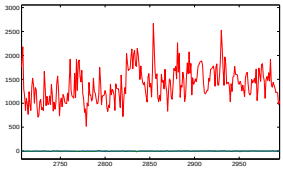
Nous calculons la PCE entre la composante de bruit de chaque couche et le RPN de chaque appareil de la même couche. La figure 5 illustre des exemples des trois couches différentes, où l'axe x représente le nombre de tranches et l'axe y la valeur du PCE.
En Figure. 5.a un exemple de la corrélation entre 1500 images (500 de chaque appareil, respectivement) et les trois RPNs,



(a) PCEs de la couche de tissu des trois scanners



(b) PCEs from the bone of S1



(c) PCEs from the air of GE

Fig. 5. PCEs des images testées à partir de 3 scanners et 3 couches différentes.

en considérant que la couche de tissu seulement. Nous notons que :

– 0-500 : images de S1. Elles sont classées correctement selon les valeurs de position les plus élevées (courbe bleue).
– 501-1000 : images de S2, elles sont classées correctement en fonction de l'endroit avec RPN du deuxième scanner Siemens (courbe verte).
– 1001-1500 : images de GE, il est tout à fait clair que les valeurs les plus élevées de PCE représentent également la classification correcte de ces images (courbe rouge).

Dans la Fig. 5.b, la corrélation de la couche d'os seulement, entre un groupe partiel de tests de S1 et les trois RPNs des trois appareils. Nous pouvons identifier la valeur maximale de PCE selon le RPN du S1. Dans la figure 5.c la corrélation de la couche d'air seulement, entre un groupe partiel testé de GE et les trois RPNs des trois appareils. Les valeurs les plus élevées de PCE avec RPN de GE classent correctement ces images.

Après avoir défini la source du scanner X de chaque couche pour toutes les images testées. Selon le tableau 1, nous pouvons continuer nos critères d'identification pour identifier le scanner X de chaque image. Le tableau 2 illustre la matrice de confusion de la précision de l'identification.

	Siemens 1	**Siemens 2**	**GE**	**No ID**
Siemens 1	**81.23** %	9.29 %	3.23 %	6.25 %
Siemens 2	4.75 %	**83.63** %	4.24 %	7.38 %
GE)	5.27 %	4.03 %	**81.81** %	8.89 %

Table 2. Confusion matrice de précision de l'identification.

Dans le tableau 2, chaque ligne représente la précision de l'identification de chaque appareil, la dernière colonne représente le pourcentage des images qui ne sont pas classées du tout, pour lesquelles au moins deux couches ne sont pas bien classées.

5. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons proposé une nouvelle méthode pour l'identification de scanners X. Elle est basée sur un nouveau concept d'extraction d'empreinte du capteur. Nous avons exploité les propriétés des images médicales pour présenter une empreinte basée sur trois couches. Notre travail a été appliquée à des données réelles à partir des images de patients. La méthode proposée avec la nouvelle technique de RPN montre une bonne performance d'identification, et nous avons été en mesure d'identifier des scanners X partir d'une seule image 2D.
Dans le travail à venir, nous allons étudier la possibilité d'attaquer ce type d'empreinte, d'examiner l'influence de la modification de l'image sur la méthode proposée et d'étudier l'influence de la compression de l'image sur notre méthode.

6. REFERENCES

[1] J. T. Bushberg et J. M. Boone. *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 2011.

[2] K. D. Toennies. *Guide to Medical Image Analysis - Methods and Algorithms*. Advances in Computer Vision and Pattern Recognition. Springer, 2012.

[3] H. T. Sencar et N. Memon. *Digital Image Forensics : There is More to a Picture Than Meets the Eye*. Springer, 2013.

[4] J. Lukas, J. Fridrich, et M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2) :205–214, 2006.

[5] M. Chen, J. Fridrich, M. Goljan, et J. Lukás. Determining image origin and integrity using sensor noise. *Information Forensics and Security, IEEE Transactions on*, 3(1) :74–90, 2008.

[6] T. Filler, J. Fridrich, et M. Goljan. Using sensor pattern noise for camera model identification. Dans *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 1296–1299. IEEE, 2008.

[7] M. Goljan, J. Fridrich, et T. Filler. Large scale test of sensor fingerprint camera identification. Dans *IS&T/SPIE Electronic Imaging*, pages 72540I–72540I. International Society for Optics and Photonics, 2009.

[8] X. Kang, Y. Li, Z. Qu, et J. Huang. Enhancing source camera identification performance with a camera reference phase sensor pattern noise. *Information Forensics and Security, IEEE Transactions on*, 7(2) :393–402, April 2012.

[9] C. T. Li. Source camera identification using enhanced sensor pattern noise. *Trans. Info. For. Sec.*, 5(2) :280–287, 2010.

[10] C. Shi, N.-F. Law, H.-F. Leung, et W.-C. Siu. Weighting optimization with neural network for photo-response-non-uniformity-based source camera identification. Dans *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pages 1–7, Dec 2014.

[11] L.-H. Chan, N.-F. Law, et W.-C. Siu. A two dimensional camera identification method based on image sensor noise. Dans *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 1741–1744, March 2012.

[12] J. B. Solomon, O. Christianson, et E. Samei. Quantitative comparison of noise texture across CT scanners from different manufacturers. *Medical physics*, 39(10) :6048–55, October 2012.

[13] Y. Duan, G. Coatrieux, et H. Shu. Identification of digital radiography image source based on digital radiography pattern noise recognition. Dans *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 5372–5376. IEEE, 2014.

[14] A. Kharboutly, W. Puech, G. Subsol, et D. Hoa. Ct-scanner identification based on sensor noise analysis. Dans *Visual Information Processing (EUVIP), 2014 5th European Workshop on*, pages 1–5. IEEE, 2014.

[15] A. Kharboutly, W. Puech, G. Subsol, et D. Hoa. Improving sensor noise analysis for ct-scanner identification. Dans *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2411–2415. IEEE, 2015.

[16] Y. Duan, G. Coatrieux, et H. Shu. Computed tomography image source identification by discriminating ct-scanner image reconstruction process. Dans *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 5622–5625. IEEE, 2015.

[17] G. C. Kagadis et S. G Langer. *Informatics in medical imaging*. CRC Press, 2011.

[18] H. B. Costa, R. F. Zampolo, D. M. Carmo, A. R. Castro, et E. P. Santos. On the practical aspects of applying the prnu approach to device identification tasks. *International Conference on Multimedia Forensics, Surveillance and Security*, September 2012.

[19] M. Chen, J. Fridrich, et M. Goljan. Digital imaging sensor identification (further study. Dans *In Security, Steganography, and Watermarking of Multimedia Contents IX. Edited by Delp, Edward J., III ; Wong, Ping Wah. Proceedings of the SPIE, Volume 6505*, 2007.

[20] G. N. Hounsfield. Computed medical imaging. *Journal of computer assisted tomography*, 4(5) :665–674, 1980.

[21] E. Jerhotová, A. Procházka, et J. Švihlík. *Biomedical image volumes denoising via the wavelet transform*. INTECH Open Access Publisher, 2011.

[22] N. Jacob et A. Martin. Image denoising in the wavelet domain using Wiener filtering. *Unpublished course project*, 2004. [Online], Project Report, Available : http ://homepages.cae.wisc.edu/∼ece533/project/f04/jacob_martin.pdf.

[23] A. Alfalou et C. Brosseau. Understanding correlation techniques for face recognition : from basics to applications. *Face Recognition*, pages 353–380, 2010.

Fusion et biométrie douce pour la dynamique de frappe au clavier

S.Z. Syed Idrus[1] E. Cherrier[2] C. Rosenberger[2]

[1] University Malaisia Perlis, Malaisie

[2] Normandie Univ, ENSICAEN, UNICAEN, UMR 6072 GREYC, France syzul@unimap.edu.my, {estelle.cher

Résumé

La dynamique de frappe au clavier (DDF) est une modalité biométrique comportementale. Moins invasive que les empreintes digitales, cette modalité présente néanmoins de moins bonnes performances que les systèmes biométriques morphologiques. Plusieurs pistes peuvent être envisagées pour augmenter les performances de la DDF : la multibiométrie (la DDF est combinée avec d'autres modalités), l'évaluation de la qualité des données biométriques ou encore la biométrie douce. Cet article présente une approche originale de biométrie douce pour la dynamique de frappe au clavier. Le système proposé est capable d'établir un profil de l'utilisateur, plus précisément de reconnaître : le nombre de mains utilisées pour taper au clavier, le sexe, la catégorie d'âge de l'utilisateur et sa latéralité (droitier/gaucher). Les taux de reconnaissance obtenus sont satisfaisants, et sont améliorés par des processus de fusion des données. Le système est évalué sur une base de données spécialement créée pour cette étude, à la fois en France, et en Norvège.

Mots clefs

Biométrie douce, fusion de données, SVM, profilage

1 Introduction

Un système biométrique est un système d'authentification d'un individu, à l'aide de caractéristiques physiques (visage, empreintes digitales, iris…), physiologiques (ADN, odeur…) ou comportementales (dynamique de frappe au clavier, dynamique de signature, démarche…). La biométrie utilise des techniques de traitement du signal et de reconnaissance de formes. En ce qui concerne le traitement du signal, on peut mentionner : la capture de la donnée (par un appareil-photo, un capteur dédié), le pré-traitement de la donnée (échantillonnage, filtrage pour atténuer le bruit), l'extraction des caractéristiques (temporelles ou spectrales, par exemple dans le cas de la reconnaissance du locuteur). Les techniques de reconnaissance de formes sont largement utilisées pour classifier les données, créer un modèle de l'utilisateur, comparer les données stockées et les données présentées au capteur. Aujourd'hui, il existe de multiples usages des systèmes biométriques : contrôle d'accès physique, contrôle de présence, paiement électronique, etc…Cet article s'intéresse au profilage d'individus à partir de biométrie douce pour la dynamique de frappe au clavier. La notion de *biométrie douce* a été introduite par Jain et al. [1]. Les auteurs définissent les *traits de biométrie douce* de la manière suivante : *un caractère fournissant de l'information sur l'individu, mais manquant d'unicité et de permanence pour différencier suffisamment deux individus*. Ce sont donc des caractéristiques qui ne sont pas suffisantes pour authentifier un individu, mais peuvent aider à la construction d'un profil. L'article [2] cite en example : la couleur de la peau, la couleur des cheveux, la couleur des yeux, la présence d'une barbe, plus généralement des informations extraites à partir du visage, mais aussi le sexe, l'âge, la taille, le poids, la démarche, différentes mesures du corps. De façon naturelle, la biométrie douce permet une recherche plus rapide dans une base de données (par élimination de données non conforme au profil) ainsi qu'une amélioration des performances du système d'authentification seul (par fusion des scores de biométrie douce et du score d'authentification). La biométrie douce est également considérée comme non invasive (par rapport à des modalités comme les empreintes digitales ou le visage), sans risque d'usurpation d'identité, avec une mise en oeuvre à faible coût.

La dynamique de frappe au clavier est une modalité biométrique qui consiste à mesurer les rythmes qui se dégagent lorsqu'on tape sur un clavier d'ordinateur [3], [4], [5]. En ce sens, c'est une modalité biométrique *comportementale*, de même que la dynamique de signature, la démarche ou la voix. Parmi les avantages de la dynamique de frappe au clavier par rapport à d'autres modalités, nous pouvons mentionner son faible coût et sa facilité d'usage : en effet, en dehors d'un clavier, aucun capteur ni dispositif supplémentaire n'est nécessaire et les utilisateurs sont habitués à taper un mot de passe. En contrepartie, la dynamique de frappe présente de plus faibles performances que les autres modalités biométriques comme les empreintes digitales, le visage, l'iris. Cela peut s'expliquer par une variabilité intra-classe élevée. Une façon de gérer cette variabilité est de prendre en compte des informations supplémentaires dans le processus de décision. Cela peut être fait de différentes manières :

— en fusionnant la dynamique de frappe au clavier avec une autre modalité biométrique (multibiométrie)

— en optimisant l'étape d'enrôlement (une donnée biométrique est exploitée pour la génération de la

référence seulement si le niveau de qualité est suffisant)
— en combinant la dynamique de frappe au clavier avec des caractéristiques de biométrie douce

Dans cet article, nous considérons les caractéristiques de biométrie douce suivantes pour la dynamique de frappe (en abrégé DDF) : le nombre de mains (l'utilisateur peut taper au clavier avec une ou deux main(s)), le sexe, l'âge (plus ou moins de 30 ans), la latéralité (droitier ou gaucher). Pour réaliser cette étude, nous avons créé une nouvelle base de données. Deux cas sont considérés : des mots de passe statique et du texte libre. En utilisant des techniques d'apprentissage statistique et des processus de fusion des données, les résultats obtenus sont encourageants.

L'article est organisé comme suit. La partie 2 présente la méthodologie adoptée pour acquérir les données, pour constituer une nouvelle base de données et les méthodes d'apprentissage et de fusion retenues. Les résultats obtenus font l'objet de la partie 3. La partie 4 conclut cet article et présente quelques perspectives.

2 Méthodologie

Cette partie présente la méthodologie adoptée pour collecter la base de données biométrique et pour traiter les données.

2.1 Capture des données de DDF

L'authentification par DDF ne requiert généralement qu'un clavier d'ordinateur et une application dédiée pour récupérer les actions sur ce clavier (appui ou relâchement de touches). Chaque capture est stockée dans une base de données dédiée et comporte les informations suivantes :

— touche concernée
— type d'événement (appui ou relâchement)
— temps d'exécution de l'événement

A partir de ces données brutes, on extrait le modèle de l'utilisateur constitué de différentes données temporelles, comme illustré à la figure 1 :

— pp : intervalle de temps entre deux pressions successives
— rr : intervalle de temps entre deux relâchements successifs
— rp : intervalle de temps entre l'appui sur une touche et le relâchement de la suivante
— pr : intervalle de temps entre un relâchement et l'appui sur la touche suivante

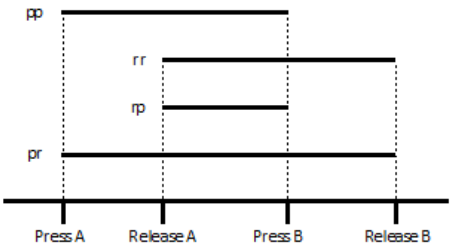


Figure 1 – *Caractéristiques extraites de la DDF*

Ces données sont concaténées dans un vecteur V.

2.2 Description de la base de données

Une base de données conséquente (110 utilisateurs) a été créée pour étudier la biométrie douce pour la dynamique de frappe au clavier. La collecte des données a eu lieu à la fois en France et en Norvège, auprès de volontaires (étudiants ou enseignants, chercheurs, personnels de laboratoires de recherche, grand public...), originaires de 24 pays différents. Au total, 70 personnes en France et 40 en Norvège ont participé à l'expérimentation. En plus des données de DDF, nous avons des informations sur les volontaires : leur sexe, leur âge, leur latéralité. . .Plus de détails sur cette base peuvent être trouvés dans l'article [6]. On peut mentionner d'autres études du même genre, comme les références [7], [8]. La figure 2 donne quelques statistiques sur le profil des utilisateurs, avec des détails en fonction du pays de capture (France ou Norvège).

User	70 (France); 40 (Norway)
Gender	78 males (47 from France, 31 from Norway); 32 females (23 from France, 9 from Norway)
Age Category **(between 15 and 65 years old)**	$<$ 30 years old (37 men, 14 women); $\geq$ 30 years old (41 men, 18 women)
Handedness	98 right-handed (70 men, 28 women); 12 left-handed (8 men, 4 women)

Figure 2 – *Statistiques sur le profil des utilisateurs*

D'après les travaux de Giot *et al.* [9], les meilleures performances pour l'authentification à base de DDF sont obtenues avec des mots de passe connus : cela signifie que taper au clavier un mot connu est plus représentatif du comportement que de taper une suite de lettres ou de chiffres sans signification . Puisque notre expérimentation se passe en France et en Norvège, nous avons choisi des mots de passe connus dans les deux pays, notés $P1$ à $P5$ (Password 1 à Password 5) :

— P1 : *leonardo di caprio*
— P2 : *the rolling stones*
— P3 : *michael schumacher*
— P4 : *red hot chili peppers*
— P5 : *united states of america*

Ces mots de passe comportent entre 17 et 24 caractères (espaces inclus). Les volontaires doivent taper chaque mot de passe 10 fois, d'abord avec une main (au choix main droite ou main gauche, en fonction de la latéralité de l'utilisateur), puis 10 fois de nouveau avec deux mains. Pour capturer les données, nous avons utilisé le logiciel développé au Laboratoire GREYC, que l'on peut librement télécharger à l'adresse suivante : `http://www.ecole.ensicaen.fr/~rosenber/keystroke.html`. Ce logiciel est décrit plus en détails dans l'article [5]. La base de données collectées comporte ainsi 11000 données (5 mots de passe $\times$ 10 captures $\times$ 2 façons de taper $\times$ 110 utilisateurs).

A partir des données collectées, nous définissons 4 traits de biométrie douce, répartis chacun en deux classes C_1 et C_2 :

- La façon de taper
  C_1 = une main (droite ou gauche en fonction de la latéralité), C_2 = deux mains
- Le genre
  C_1 = homme, C_2 = femme
- L'âge
  C_1 = moins de 30 ans, C_2 = plus de 30 ans
- La latéralité
  C_1 = droitier, C_2 = gaucher

Remarque : les données correspondant à la frappe avec une seule main sont exploitées uniquement pour le premier critère, à savoir la façon de taper (une ou deux mains). Pour tous les autres critères, nous utilisons exclusivement les données correspondant à la frappe avec deux mains (même pour le dernier critère, la latéralité), qui est plus naturelle (et donc plus caractéristique du comportement) pour les personnes volontaires.

Nous avons choisi ces traits de biométrie douce en association avec la dynamique de frappe au clavier pour les raisons suivantes : la façon de taper (une ou deux mains) et la latéralité sont directement liés à la façon de taper sur un clavier ; le genre a déjà été étudié dans l'article [8], nous avons donc repris ce critère ; quant à la catégorie d'âge (de même que le genre), il pourrait être utilisé ultérieurement dans une application de recherche de pédophile dans un tchat pour mineurs (la façon de taper au clavier pourrait indiquer si l'utilisateur est bien une adolescente de 14 ans, et pas un homme plus âgé par exemple).

A partir de cette base de données biométriques de dynamique de frappe au clavier, nous présentons les méthodes que nous allons avons choisies pour prédire le profil de chaque utilisateur.

2.3 Apprentissage et fusion

Pour chaque utilisateur, nous choisissons de ne pas tenir compte des trois premières captures, pour chaque mot de passe : cela correspond au temps nécessaire pour apprendre à taper correctement chaque mot de passe et permet de supprimer des hésitations qui ne révèlent pas la véritable façon de taper.

Pour chaque caractéristique de biométrie douce, pour chaque mot de passe, un Séparateur à Vastes Marges (SVM) [10] est entraîné à reconnaître les deux classes C_1 et C_2, voir la figure 3.

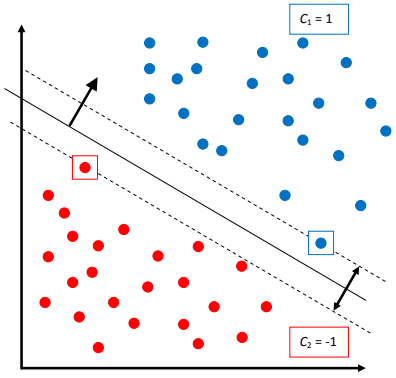


Figure 3 – *Classification : séparateur à vastes marges*

On utilise la bibliothèque LibSVM [11] avec un noyau gaussien radial (RBF : Radial Basis Function), et les valeurs de paramètres suivantes : $\gamma = 0.125$, le coefficient de pénalisation $C = 128$. Pour l'apprentissage, une partie de la base de données est conservée (entre 1% et 90%), le reste de la base sert aux tests et à l'évaluation de la performance, c'est-à-dire le taux de correcte reconnaissance de chaque critère de biométrie douce. Par conséquent, plus il y a de données réservées à l'apprentissage, meilleur sera le taux de reconnaissance de chaque classe : on peut supposer que la reconnaissance d'un critère est plus efficace si on apprend au SVM à le reconnaître sur 50 exemples différents que sur 10 exemples. Cette remarque sera vérifiée dans la section 3 consacrée à l'analyse des résultats. On peut noter qu'il existe dans notre base des classes déséquilibrées (hommes-femmes, droitiers-gauchers, voir la figure 2) : pour ces critères, on sélectionne au hasard dans la classe surnuméraire un nombre de données correspondant au nombre total de données de l'autre classe : par exmple, le nombre total de femmes est 32, on sélectionne donc aléatoirement les données de 32 hommes à chaque expérience parmi les 78 possibles.

Pour améliorer la confiance dans les résultats, pour chaque critère, pour chaque mot de passe, pour chaque pourcentage fixé de données d'apprentissage, on répète 100 fois l'entraînement d'un SVM (différent à chaque fois), et la performance de la classification selon ce critère est calculée comme le taux moyen de reconnaissance sur ces 100 essais. Pour valider les résultats obtenus, des intervalles de confiance sont calculés (voir les références [12] et [13] pour plus de détails). L'équation (1) donne la formule de

l'indice de confiance à 95% :

$$CI = m(taux) \pm 1.96\frac{\sigma(taux)}{\sqrt{N}} \quad (1)$$

où $m(taux)$ et $\sigma(taux)$ sont respectivement la moyenne et l'écart-type du taux de reconnaissance sur $N = 100$ itérations.

Pour améliorer les performances du système de biométrie douce, c'est-à-dire la reconnaissance de chaque trait de biométrie douce, un processus de fusion peut être appliqué. Les techniques de fusion des données sont des techniques classiques de reconnaissance de formes, très souvent utilisées en biométrie pour définir des systèmes multibiométriques [14], [15], [16]. Le principe est le suivant : pour chaque critère de biométrie douce, au lieu de considérer séparément le résultat de chaque classifieur (un par mot de passe $P1$ à $P5$), ces cinq résultats vont être fusionnés. Il existe plusieurs façons de fusionner les données issues de plusieurs systèmes biométriques : on peut appliquer une fusion de captures, de caractéristiques, de score, de rang, de décision (pour plus de détails, voir [15], [16]). Plus la fusion a lieu proche de la capture des données, plus la quantité d'information à fusionner est importante : de quelques Mo pour la fusion de captures (par exemple fusion de deux images d'empreintes digitales), à 1 bit pour la fusion de décision (décision finale = oui ou non, 1 ou 0...). Dans cet article, nous nous intéressons au vote majoritaire et à la fusion de scores, à partir de la décision prise par les 5 classifieurs et de la probabilité associée. Les deux processus de fusion sont illustrés à la figure 4.

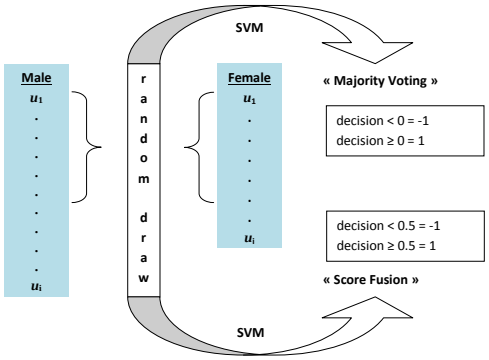


Figure 4 – *Fusion des données pour le critère* genre

Pour plus de clarté, nous détaillons ces deux procédés de fusion sur le critère du genre. Rappelons que l'on doit sélectionner les données de 32 hommes (car il y a 32 femmes dans la base de données) parmi les 78 disponibles. Nous gardons ici (dans la partie sur la fusion) le même sous-ensemble d'hommes pour chaque mot de passe. Un SVM est entraîné pour le critère du genre, à partir des données sélectionnées (un SVM différent est entraîné pour chaque critère de biométrie douce). Pour éviter l'influence de la sélection du sous-ensemble d'hommes, tout le processus (de l'extraction du sous-ensemble à la fusion) est répété 100 fois, avec un nouveau sous-ensemble à chaque fois : les résultats présentés sont la moyenne de ces 100 expériences. Dans cette partie, on ne s'intéresse plus à l'influence de la quantité de données réservées à l'apprentissage, par conséquent on fixe le pourcentage de données à 50% pour l'apprentissage et 50% pour les tests. La première méthode de fusion étudiée est le vote majoritaire. Pour cela, on exploite le label +1 ou -1 donné par LibSVM pour chaque mot de passe, correspondant à la classe C_1 ou C_2 du critère considéré (ici le genre). Puisqu'il y a 5 mots de passe, par simple addition de ces labels on obtient un score qui permet de procéder à un vote majoritaire. Pour appliquer la seconde méthode de fusion, on exploite non seulement le label donné par LibSVM, mais également la probabilité associée (appartenant à l'intervalle [0,1]). Pour les 5 mots de passe, on dispose donc de 5 labels et de 5 probabilités : on multiplie les labels (1 ou -1) par les probabilités correspondantes pour obtenir un score et on calcule la moyenne de ces 5 scores pour décider de la classe C_1 ou C_2.
Dans les deux cas (vote majoritaire ou fusion de scores), la décision finale est prise à partir d'une plus grande quantité d'information : 5 mots de passe au lieu d'un seul. On peut donc espérer que la performance de reconnaissance du genre (ainsi que de tous les autres critères) sera augmentée. La partie suivante présente les résultats obtenus pour les différents processus d'apprentissage et de fusion.

3 Résultats

3.1 Apprentissage séparé pour chaque mot de passe

On s'intéresse dans un premier temps à l'évolution du taux moyen (la moyenne est calculée sur 100 itérations) de reconnaissance de chaque critère de biométrie douce, pour chaque mot de passe. Les figures 5 à 8 montrent les résultats obtenus.

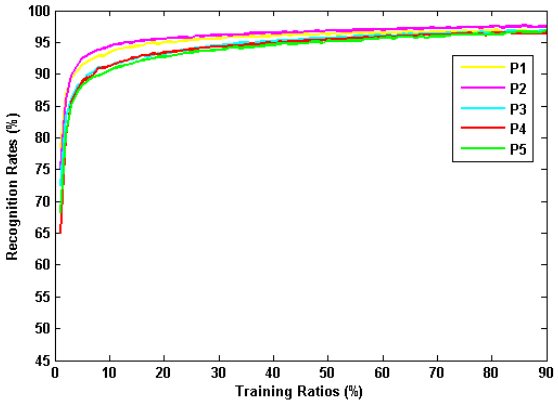


Figure 5 – *Taux de reconnaissance du nombre de mains utilisées*

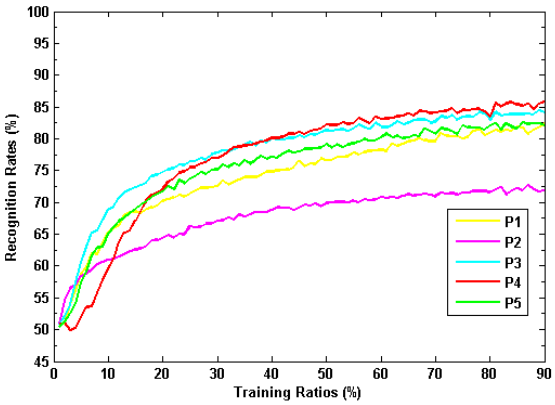


Figure 6 – *Taux de reconnaissance du sexe de l'utilisateur*

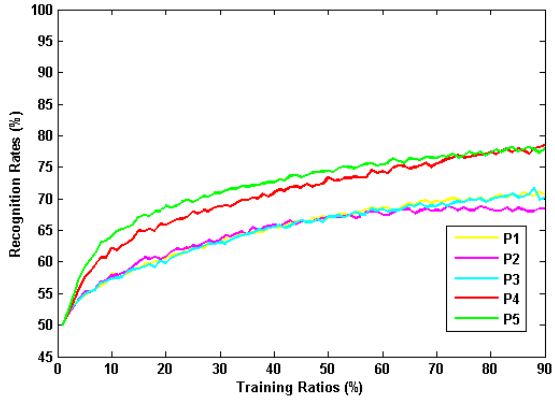


Figure 7 – *Taux de reconnaissance de la classe d'âge*

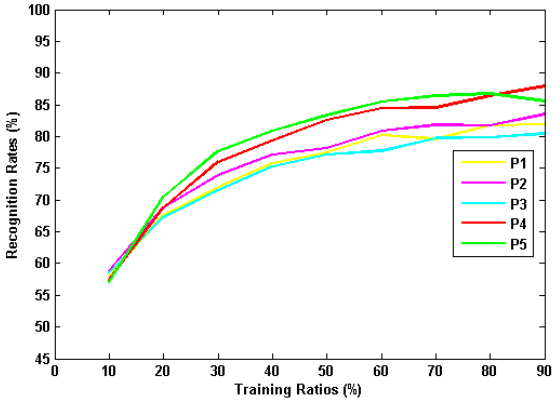


Figure 8 – *Taux de reconnaissance de la latéralité*

Les résultats obtenus sont également présentés dans le tableau 1 avec les intervalles de confiance correspondants, dans le cas où 50% des données disponibles sont utilisées pour la phase d'apprentissage. On constate que les performances sont comprises entre 63% et 96%.

3.2 Amélioration des performances par fusion

Dans cette partie, nous étudions l'influence de la fusion des 5 mots de passe sur le taux de reconnaissance de chaque critère de biométrie douce. Le tableau 2 présente les résultats obtenus pour les deux processus de fusion décrits dans la partie 2.3.
On constate que la fusion de scores est plus efficace que la fusion de décision par vote majoritaire : les performances atteignent au maximum 100% de reconnaissance pour le nombre de mains utilisées, et au minimum 86% pour la catégorie d'âge. Cette situation peut s'expliquer très classiquement par le fait que la fusion de scores utilise plus d'informations (label et probabilité associée) que le vote majoritaire (label seul). De manière générale, il est préférable d'utiliser une fusion de scores plutôt qu'une fusion de décision, mais cela n'est pas toujours facile (dans le cas de scores hétérogènes, qui nécessitent une étape supplémentaire de normalisation des scores), ni possible (par exemple, certains capteurs biométriques ne donnent accès qu'à la décision finale, et pas au score obtenu par l'utilisateur).

4 Conclusion et perspectives

Cet article s'intéresse à la biométrie douce pour la dynamique de frappe au clavier. Les critères de biométrie douce ne sont pas suffisants pour authentifier un utilisateur, mais permettent d'en établir un profil. Ce profil peut ensuite être utilisé à de nombreuses fins : amélioration des performances de recherche dans une grande base de données, amélioration des performances d'un système biométrique classique : le profil vient renforcer la confiance dans le score de décision. En ce qui concerne la dynamique de frappe au clavier, les critères étudiés sont les suivants : nombre de mains utilisées, sexe, catégorie d'âge et latéralité. Les taux de reconnaissance obtenus en sortie des SVM entraînés pour chaque mot de passe sont compris entre 63% et 96%. La fusion des données disponibles pour les 5 mots de passe permet d'atteindre des performances au-delà de 86%. Nos travaux futurs concerneront l'exploitation des informations de profilage pour améliorer les performances d'un système d'authentification par DDF.

		Taux de reconnaissance et intervalle de confiance				
Critère	**Nombre de données**	P_1	P_2	P_3	P_4	P_5
Nb mains	770 par classe	$96\% \pm 0.1\%$	$96\% \pm 0.1\%$	$95\% \pm 0.1\%$	$94\% \pm 0.1\%$	$94\% \pm 0.1\%$
Sexe	224 par classe	$74\% \pm 0.3\%$	$69\% \pm 0.3\%$	$70\% \pm 0.2\%$	$78\% \pm 0.2\%$	$76\% \pm 0.2\%$
Age	357 par classe	$64\% \pm 0.2\%$	$64\% \pm 0.2\%$	$63\% \pm 0.2\%$	$69\% \pm 0.2\%$	$69\% \pm 0.2\%$
Latéralité	84 par classe	$72\% \pm 1.2\%$	$73\% \pm 1.2\%$	$72\% \pm 1.2\%$	$72\% \pm 1.3\%$	$73\% \pm 1.2\%$

Tableau 1 – *Taux de reconnaissance avec intervalles de confiance pour un apprentissage avec* 50% *des données*

		By fusing	
Critère	**Sans fusion**	**Vote majoritaire**	**Fusion de score**
Nb mains	94%	100%	100%
Sexe	63%	86%	92%
Age	55%	87%	86%
Latéralité	62%	85%	92%

Tableau 2 – *Comparaison des performances sans et avec fusion pour un apprentissage avec* 50% *des données*

Références

[1] A. Jain, S. Dass, et K. Nandakumar. Soft biometric traits for personal recognition systems. *Biometric Authentication*, pages 1–40, 2004.

[2] Antitza Dantcheva, Carmelo Velardo, Angela D'angelo, et Jean-Luc Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2) :739–777, 2011.

[3] L.C.F. Araujo, Jr. Sucupira, L.H.R., M.G. Lizarraga, L.L. Ling, et J.B.T. Yabu-Uti. User authentication through typing biometrics features. *Signal Processing, IEEE Transactions on*, 53(2) :851–855, 2005.

[4] Daniele Gunetti et Claudia Picardi. Keystroke analysis of free text. *ACM Trans. Inf. Syst. Secur.*, 8(3) :312–347, 2005.

[5] R. Giot, M. El-Abed, et C. Rosenberger. Greyc keystroke : A benchmark for keystroke dynamics biometric systems. Dans *Biometrics : Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, 2009.

[6] S.Z.S. Idrus, E. Cherrier, C. Rosenberger, et P. Bours. Soft biometrics database : A benchmark for keystroke dynamics biometric systems. Dans *Biometrics Special Interest Group (BIOSIG), 2013 International Conference of the*, 2013.

[7] M. Bertacchini, C. Benitez, et P. Fierens. User clustering based on keystroke dynamics. Dans *In XVI Congreso Argentino de Ciencias de la Computacion (CACIC 2010)*, 2010.

[8] Romain Giot et Christophe Rosenberger. A new soft biometric approach for keystroke dynamics based on gender recognition. *International Journal of Information Technology and Management*, 11(1) :35–49, 2012.

[9] Romain Giot, Alexandre Ninassi, Mohamad El-Abed, et Christophe Rosenberger. Analysis of the acquisition process for keystroke dynamics. Dans *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*, pages 1–6. IEEE, 2012.

[10] V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[11] C.-C. Chang et C.-J. Lin. Libsvm : A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), 2011.

[12] A. Mayoue. Biosecure tool - performance evaluation of a biometric verification system. `http://svnext.it-sudparis.eu/svnview2-eph/ref_syst/Tools/PerformanceEvaluation/doc/howTo.pdf`, 2007.

[13] R.M. Bolle, N.K. Ratha, et S. Pankanti. Evaluation techniques for biometrics-based authentication systems. Dans *15th Internat. Conf. Pattern Recogn.*, 2000.

[14] A. Ross et A. Jain. Information fusion in biometrics. *Pattern recognition letters*, 24(13), 2003.

[15] A. Ross, K. Nandakumar, et A. Jain. *Handbook of multibiometrics*, volume 6. Springer, 2006.

[16] Romain Giot, Baptiste Hemery, Estelle Cherrier, et Christophe Rosenberger. La multibiométrie. Dans *Traitement du signal et de l'image pour la biométrie*, page Chapitre 9. Hermès, 2012.

L'analyse de la voix comme outil de diagnostic précoce de la maladie de Parkinson : état de l'art

L. Jeancolas[1] D. Petrovska-Delacrétaz[1] S. Lehéricy[3] H. Benali[2] B.-E. Benkelfat[1]

[1] SAMOVAR, UMR 5157, Télécom SudParis, CNRS, Université Paris-Saclay
{laetitia.jeancolas, badr-eddine.benkelfat, dijana.petrovska}@telecom-sudparis.eu

[2] Laboratoire LIB, UMR 1146 INSERM-CNRS-Université Pierre et Marie Curie
habib.benali@lib.upmc.fr

[3] Institut du Cerveau et de la Moelle (ICM) et Centre de Neuroimagerie de Recherche (CENIR)
Sorbonne Universités, UPMC Univ Paris 06, UMR S 1127, Inserm U 1127, CNRS UMR 7225

Résumé

Parmi les manifestations cliniques de la maladie de Parkinson, les troubles de la voix surviennent particulièrement tôt dans la maladie. Ils prennent principalement la forme d'une insuffisance prosodique, articulatoire et phonatoire. La voix des patients devient monotone, à débit variable, éraillée ou soufflée, et les articulations deviennent imprécises. Notre objectif est de dresser un état de l'art concernant les études réalisées sur l'analyse de la voix comme outil de diagnostic précoce de la maladie de Parkinson. Nous nous intéresserons aux études qui utilisent des paramètres acoustiques vocaux pour détecter la maladie de Parkinson chez des parkinsoniens récents, et chez des parkinsoniens en phase préclinique. L'ensemble des résultats nous porte à croire que la voix pourrait à terme être utilisée en clinique pour établir un diagnostic précoce de la maladie de Parkinson. Cependant le faible nombre d'études acoustiques au stade préclinique, et le faible nombre de patients dans chaque étude nous incitent à être prudents quant aux conclusions. Des études longitudinales, avec des groupes plus importants de patients, semblent nécessaires pour valider ces résultats.

Mots clefs

Parkinson, voix, diagnostic précoce.

1 Introduction

La maladie de Parkinson est une maladie neurodégénérative, la 2^e plus courante après la maladie d'Alzheimer, qui se manifeste essentiellement par des troubles moteurs s'aggravant au cours du temps. Sa prévalence augmente avec l'âge et est de 1 à 2 % chez les plus de 65 ans. Etant donné que notre espérance de vie est en constante progression, il en va de même avec le nombre de personnes atteintes de cette maladie. On connait mal la cause de la maladie de Parkinson mais on sait qu'elle s'accompagne d'une raréfaction des neurones dopaminergiques dans les ganglions de la base (situés dans les couches profondes du cerveau). Cette raréfaction a plus précisément lieu dans la substance noire compacte, aboutissant à un défaut de libération de dopamine dans le striatum. A ce jour, le diagnostic repose principalement sur un examen moteur effectué par un neurologue. Habituellement le diagnostic est posé si l'examinateur observe au moins deux des trois symptômes suivants : akynésie (lenteur d'initiation des mouvements), rigidité et tremblements au repos. Malheureusement ces symptômes moteurs ne se manifestent qu'après la perte de 70% des neurones dopaminergiques [1]. Un enjeu majeur de la recherche consiste donc à trouver des moyens de détecter plus précocement la maladie, afin de pouvoir ralentir, voire stopper, sa progression dès le début, quand on aura de tels traitements.

Parmi les manifestations cliniques diverses de cette maladie, la modification de la voix des malades semble être un élément d'intérêt à plusieurs égards. Un grand nombre de publications existent sur l'étude de la voix chez des patients parkinsoniens ; elles ont mis en évidence des perturbations telles que la dysarthrie hypokinétique, qui se manifeste par une diminution de la prosodie (de l'intonation), des irrégularités dans la phonation et des difficultés d'articulation. Actuellement on sait identifier de manière automatique la maladie de Parkinson au stade clinique avec une précision de 90 % en analysant seulement la voix [2]. De plus certaines perturbations de la voix caractéristiques de le maladie de Parkinson sont déjà visibles plusieurs années avant le diagnostic clinique [3, 4]. Cependant il n'y a eu que très peu d'études visant à identifier des marqueurs pronostiques de cette maladie dans la voix au stade préclinique [5]. Cet article a pour objectif de dresser un état de l'art sur ce que l'on sait des perturbations vocales (via des analyses perceptives et acoustiques) dans la maladie de Parkinson, en se focalisant surtout sur le début de la maladie. Nous nous intéresserons particulièrement aux marqueurs acoustiques les plus discriminants qui pourraient être utilisés pour éta-

blir un diagnostic précoce. Dans un premier temps nous présenterons un état de l'art des études qui se sont intéressées à la détection de la maladie de Parkinson par l'analyse de la voix chez des parkinsoniens récents (section 2). On appelle parkinsoniens récents les patients qui ont été enregistrés moins de 4 ans après leur diagnostic. Puis nous nous intéresserons aux études qui ont détecté des troubles de la voix avant même le diagnostic clinique de la maladie, donc pendant la phase prodromique (phase pendant laquelle un certain nombre de symptômes avant coureurs, souvent bénins, annoncent la venue de la phase principale, on peut aussi parler de phase préclinique) de la maladie de Parkinson (section 3). Nous finirons par une synthèse des performances obtenues lors des classifications (section 4). En Annexe le tableau 1 dresse l'inventaire des études qui se sont intéressées à la détection précoce de la maladie de Parkinson par l'analyse de la voix : chez les parkinsoniens récents et durant la phase préclinique. Le nombre de sujets, le type de tâches vocales et les paramètres acoustiques retenus comme étant les plus discriminants, et les performances des classifications y sont répertoriés. Le tableau 2 décrit, quant à lui, les différents paramètres acoustiques qui ressortent de ces études, et les meilleurs tâches vocales pour les extraire.

2 Détection de la maladie de Parkinson par l'analyse de la voix chez des parkinsoniens récents

2.1 Bases de données

La grande majorité des analyses acoustiques sur la voix dans la maladie de Parkinson, dont les résultats ont été publiés, ont été faites sur des groupes d'au plus 50 parkinsoniens et l'équivalent en sujets sains. De plus les enregistrements vocaux constituants les bases de données sont très peu disponibles publiquement. Ceci est en parti dû au fait que les données et informations médicales sont des données sensibles.

2.2 Enregistrements

Lors des études acoustiques sur la voix des sujets parkinsoniens, diverses sortes de microphones ont été utilisées pour enregistrer la voix. Certaines équipes ont choisi d'utiliser des microphones dynamiques [6, 7], d'autres ont choisi des électrostatiques (de meilleur qualité mais plus couteux et nécessitant une alimentation externe) [2, 4]. Concernant la directivité des microphones, certains ont préféré les cardioides (moins sensibles aux bruits extérieurs mais déformant un peu le timbre de la voix) [2, 7], et d'autres ont préféré les omnidirectionnels [4, 6]. Dans l'ensemble les microphones serre tête sont plus utilisés car ils permettent de maintenir une distance constante entre la bouche et le microphone, ce qui est important pour pouvoir analyser les variations d'intensité. Concernant la gamme fréquentielle, la plupart des microphones utilisés dans ces études vont de 20Hz à 20kHz, ce qui couvre largement le spectre de la voix qui va de 50 Hz à 12 kHz. La fréquence d'échantillonnage couramment utilisée est de 44,1 kHz. Les enregistrements se passent la plupart du temps dans des salles sourdes ou silencieuses, car les paramètres mesurant les variations d'intensité sont très sensibles aux bruits extérieurs. Depuis peu, certaines équipes s'intéressent aux enregistrements de la voix des parkinsoniens par téléphone [8] mais elles ne ciblent pas pour l'instant les parkinsoniens récents.

2.3 Troubles de la voix chez les parkinsoniens récents

Les études sur la voix dans la maladie de Parkinson parlent souvent de dysarthrie hypokinétique (qui signifie réduction de l'amplitude des mouvements des muscles responsables de l'articulation) pour catégoriser les troubles de la voix des parkinsoniens. Les différentes composantes de la paroles affectées par la dysarthrie parkinsonienne sont :
- la prosodie : une perte des modulations d'intensité et de hauteur donne à la voix un caractère monotone, le débit est altéré et on constate aussi des troubles de la fluence (palilalies, bredouillements ...) ;
- l'articulation : la précision articulatoire des voyelles et des consonnes est altérée ;
- la phonation : l'intensité de la voix diminue (le patient devient hypophone), la hauteur moyenne s'abaisse ou s'élève, la hauteur et l'intensité deviennent instables, et le timbre devient soufflé, voilé, éraillé) [9] ;
- le rythme : la capacité à maintenir un rythme de parole constant s'altère [10].

Dans la suite nous détaillerons les dysfonctionnements de la voix que l'on trouve dès le début de la MP.

Prosodie. L'insuffisance prosodique constituerait la marque la plus spécifique des troubles de la parole dans la maladie de Parkinson [11]. Elle se caractérise chez les parkinsoniens récents par une monotonie de la mélodie (diminution de la variation de la fréquence fondamentale F0), par une monotonie de l'intensité, et par une diminution du nombre de pauses de plus de 60 ms. La dysprosodie serait le résultat d'une diminution de l'amplitude du mouvement du larynx et des muscles respiratoires causés par une rigidité excessive [12]. Une étude acoustique en 2011 a montré que les problèmes de prosodie étaient présents chez plus de 60% des 23 parkinsoniens récents non traités testés [13]. Cette même étude montre que le monologue et la lecture de phrases émotionnelles mettent plus en avant la diminution de la variation de la fréquence fondamentale que la lecture de texte. Une étude plus récente sur 24 parkinsoniens récents non traités a même obtenu un taux de réussite de 81,3 % de classification parkinsonien récent vs sujets contrôles, en analysant juste la variation de la fréquence fondamentale pendant un monologue [14].

Articulation. Un déficit d'articulation a été mis en évidence dans la maladie de Parkinson par de nombreuses études [13, 14, 15], et son analyse acoustique permettrait

à elle seule de discriminer des parkinsoniens récents de sujets contrôle avec un taux de réussite de plus de 88% (d'après une étude acoustique qui a porté sur 24 parkinsoniens récents non traités et 22 sujets sains [16]). Les problèmes d'articulation se voient à la fois dans l'articulation des voyelles et dans l'articulation des consonnes, et se manifestent par une diminution des contrastes acoustiques.

Articulation des voyelles :
On remarque chez les parkinsoniens une tendance à la dédifférenciation des voyelles. Les formants ayant une fréquence naturellement élevée voient leur fréquence diminuer, ce qui est le cas du 2^e formant de la voyelle "i" et du 1er formant de la voyelle "a". Les formants ayant normalement une fréquence basse subissent une augmentation de leur fréquence, ce qui est le cas du 2^e formant de la voyelle phonétique "u". Ceci a pour conséquence une diminution de la surface du triangle vocalique (VSA, pour "Vowel Space Area") et de l'index d'articulation vocalique (VAI) [15]. Cette centralisation des formants traduirait une diminution de l'amplitude des mouvements de la langue et des lèvres ; ce serait le corolaire vocal de la bradykinésie (réduction de la vitesse et de l'amplitude des mouvements) [12]. Elle apparaitrait dès le début de la maladie (chez des patients récents et non traités) [17]. D'après ces auteurs, la centralisation des formants a lieu surtout lors de la parole spontanée. On la trouve de façon plus atténuée pendant la lecture d'un texte, et n'a pas du tout lieu lors de prononciation de voyelles soutenues. Les auteurs de cette étude l'expliquent par le fait que pendant la lecture de texte, le patient peut se concentrer sur l'articulation, alors que pendant le discours spontané il doit d'abord se concentrer sur le sens de ce qu'il est en train de dire et donc prête moins d'attention à l'articulation.

Articulation des consonnes :
L'articulation des consonnes s'effectue de manière imprécise chez les patients parkinsoniens, et ce également chez des patients récents non traités. Les patients ont tendance à ne pas fermer complètement leur conduit vocal lors de la prononciation de consonnes occlusives orales (p, t, k, b, d, g). Cela crée une fuite d'air turbulent qui peut se détecter à la place du silence qui est censé avoir lieu pendant l'occlusion. Les consonnes occlusives orales ressemblent alors un peu plus à des fricatives (f, s). Le rapport signal sur bruit permet de mesurer cet effet. Il correspond à $10log(Is/In)$ avec Is l'intensité acoustique qu'on peut mesurer pendant le son voisé et In celle qui correspond à la partie censée être silencieuse pendant l'occlusion [16].
Un autre défaut de l'articulation provient d'une mauvaise coordination entre les muscles laryngaux (cordes vocales) et supralaryngaux (langue, lèvres, mandibules), ce qui entraîne une articulation imprécise. Plusieurs études [4, 16] ont montré que la durée des consonnes, mesurée grâce au paramètre "Voice Onset Time" (VOT), chez des parkinsoniens récents était alors augmentée.
Ces études ont aussi montré des anomalies formantiques chez les patients parkinsoniens récents, expliquées par des perturbations dans les mouvements de la langue lors de tâches de diadococinésie (DDK) [16].
Les différents problèmes d'articulation rencontrés ont également comme impact une diminution du débit de parole lors des tâches DDK. On l'observe notamment chez les parkinsoniens récents non traités [16].

Phonation. Les problème de phonation dans la maladie de Parkinson concernent la hauteur, l'intensité et le timbre. La tâche vocale qui les met le mieux en évidence est la prononciation de voyelles soutenues : on demande aux sujets de prononcer la voyelle "a" le plus longtemps possible sans respirer. Lors de ce type de tâche, les patients parkinsoniens, même les plus récents, ont du mal à maintenir la hauteur de la voix constante et cela ne fait que s'accentuer quand la maladie progresse. L'instabilité à moyen terme de la hauteur se mesure par l'écart type de la fréquence fondamentale (F0 SD). L'instabilité à court terme de la fréquence fondamentale se traduit par des variations de fréquence entre chaque cycle d'oscillation (appelées "jitter"). Les patients atteints de la maladie de Parkinson souffrent également d'une réduction de l'intensité moyenne et de sa stabilité. L'instabilité de l'intensité sur le moyen terme a été mise en évidence chez des parkinsoniens récents non traités lors de tâche de diadococinésie, où on a observé une augmentation du "Relative Intensity Range Variation" (RIRV) [13]. L'instabilité de l'intensité sur le court terme, apparaît surtout lors de voyelles soutenues et se traduit par une variation de l'amplitude entre chaque cycle d'oscillation (dénommée "shimmer") [13].
Le timbre de la voix des parkinsoniens est aussi altéré et apparaît comme légèrement soufflé et éraillé. Il serait dû à un accolement incomplet des cordes vocales, qui a été mis en évidence par des analyses laryngoscopiques [18]. Il peut être mesuré par le paramètre "Harmonic-to-Noise ratio" (HNR). Ce paramètre indique l'amplitude du bruit par rapport aux composantes tonales. Il est plus élevé chez les parkinsoniens que chez les sujets sains et ce dès les premières années après le diagnostic [13].
Une autre composante de la parole qui peut influencer la phonation est la respiration. Les parkinsoniens ont des problèmes de respiration (ils prennent des inspirations moins profondes et ont du mal à coordonner respiration et parole) qui font que l'intensité de leur voix est plus faible, surtout quand la maladie est un peu plus avancée [19]. Cela a aussi comme conséquence de diminuer la durée maximale de phonation (MPT, pour "Maximum Phonation Time") des femmes parkinsoniennes récentes, lorsqu'elles doivent dire des voyelles soutenues le plus longtemps possible [7]. Il est intéressant de noter que cela concerne les femmes mais pas les hommes. En effet certains paramètres acoustiques, et leur évolution au cours de la maladie, sont très sensibles au genre [15, 20] (F0, MPT, les coefficients cepstraux, VSA...). Donc pour ce type de paramètres, il vaut mieux faire des analyses séparées pour les hommes et les

femmes.

Rythme. Effectuer des mouvements automatiques à un rythme stable est quelque chose qui est connu pour être difficile dans la maladie de Parkinson. Cette instabilité motrice serait la conséquence d'un dysfonctionnement des ganglions de la base qui ne pourraient plus assurer correctement la préparation et le maintien de séquences motrices simples qui s'effectuent normalement de manière quasi automatique [21]. Cette difficulté apparaît notamment dans la parole : les parkinsoniens ont du mal à répéter une série de syllabes ("pa" par exemple) à un rythme régulier. Cette difficulté étant accrue quand le rythme leur est imposé et encore plus quand il s'agit d'alterner entre deux syllabes différentes ("pa","ti") [10]. On retrouve cette difficulté dès les premières années après le diagnostic. En effet une étude sur 50 parkinsoniens récents traités et 32 sujets sains a montré que le coefficient de variation relative du rythme est significativement plus élevé chez les parkinsoniens que chez les sujets sains, lors de la répétition de syllabes à un rythme choisi et imposé [22]. L'auteur de cette étude a aussi montré une corrélation entre le score UPDRS ("Unified Parkinson's Disease Rating Scale") qui est une mesure de l'avancement de la maladie, et le nombre maximal de syllabes que les sujets pouvaient dire par seconde, quand on leur demandait de répéter la syllabe "pa" le plus rapidement possible. Cette corrélation n'a cependant pas été trouvée avec la variation relative du rythme. L'auteur en a conclu que la vitesse de répétition et sa régularité correspondaient à des domaines différents de performance motrices basiques, avec possiblement des physiopathologies différentes.

2.4 Effet des traitements pour la maladie de Parkinson sur la voix

Certains troubles de la voix dus à la maladie de Parkinson s'améliorent avec des traitements, et ce même chez les parkinsoniens récents. Nous allons d'abord nous intéresser à un traitement comportemental : le "Lee Silverman Voice Training" (LSVT) dont le but est exclusivement d'améliorer les problèmes de voix dus à la maladie de Parkinson. Ensuite nous verrons quels sont les conséquences des traitements pharmaceutiques dopaminergiques, que l'on donne pour améliorer les dysfonctionnements moteurs de la maladie de Parkinson, sur la voix.

LSVT ("Lee Silverman Voice Training"). La LSVT est une technique d'orthophonie utilisée depuis 2004 dont le but est de limiter la diminution d'intensité vocale et la perte de prosodie, en améliorant l'accolement des cordes vocales et en renforçant de façon générale l'activation des muscles laryngés et leur contrôle. L'entrainement se déroule pendant 16 sessions d'1h réparties de façon homogène sur 4 semaines. Durant ces sessions le patient est invité à prononcer avec une voix forte des voyelles soutenues, en faisant varier ou pas la hauteur de la voix, et à parler d'une voix forte en ce concentrant sur l'intensité de sa voix. D'une manière générale on conseille au patient de "penser fort" ("think loud"), pour améliorer le traitement de l'information sensorielle auditive d'origine proprioceptive. En effet le patient parkinsonien hypophonique a tendance à ne pas se rendre compte qu'il ne parle pas assez fort [11]. Les effets bénéfiques sur l'intensité de la voix et la prosodie apparaissent généralement au bout d'un mois et sont encore visibles 2 ans après [23].

Traitements dopaminergiques. L'effet des traitements pharmaceutiques dopaminergiques sur la voix des patients a été mis en évidence récemment sur un groupe de 19 patients parkinsoniens récents [12]. Les traitements dopaminergiques ont induit des améliorations, classées de la plus à la moins importante, dans les domaines suivants (cf Tableau 2 en annexe pour la signification des paramètres acoustiques) :
- intensité de la voix (Int SD pour monologue et lecture) ;
- qualité de la voix ("jitter", "shimmer", HNR, RPDE, PPE) ;
- intonation (F0 SD pour monologue et lecture) ;
- articulation des voyelles (VAI, F2i/F2u).

Pour ces patients récents, les améliorations sont visibles dans les analyses acoustiques mais n'apparaissent pas dans les analyses perceptives (item 18 de l'UPDRS [24] inchangé). La dopamine semble donc avoir un impact sur l'intensité, la qualité et l'intonation de la voix, mais d'après une autre étude, elle n'aurait pas d'influence sur la régularité du débit de la parole, le nombre de pauses et le rythme [22].

Les traitements orthophoniques (de type LSVT) et dopaminergiques ont un impact positif sur certains troubles vocaux rencontrés dans la maladie de Parkinson. L'influence de ces traitements sur la voix doit donc être prise en compte lors de l'interprétation d'analyses vocales chez des patients parkinsoniens traités.

3 Particularités de la voix au stade préclinique de la maladie de Parkinson

Nous avons pu voir que plusieurs études avaient montré qu'il était possible de détecter la maladie de Parkinson chez des parkinsoniens récents en analysant simplement la voix. Mais l'enjeu réside surtout dans le fait de pouvoir diagnostiquer plus tôt la maladie qu'il n'est possible à l'heure actuelle avec l'examen moteur. Quelques équipes ont donc cherché à savoir si certains troubles de la voix n'apparaitraient pas avant les symptômes moteurs qui servent aux diagnostic actuel, et pourraient ainsi par la suite servir de marqueurs de diagnostic très précoce.

3.1 Etude longitudinale à partir d'extraits télévisés

En 2004 une étude a pour la première fois mis en évidence des changements mesurables dans la voix durant le stade préclinique d'un individu atteint de la maladie de Parkin-

son. Cet individu donnait régulièrement des interviews et des conférences à la télévision. En analysant les enregistrements vidéos qui dataient de 7 ans avant le diagnostic jusqu'à 3 ans après celui-ci, et en les comparant avec des enregistrements d'un sujet sain apparié, les auteurs ont montré que les variations de la fréquence fondamentale commençaient à diminuer significativement à partir de 5 ans avant le diagnostic [3]. Cette étude a le mérite d'être la première étude longitudinale à effectuer une analyse acoustique de la voix d'un patient parkinsonien pendant sa phase prodromique. Néanmoins il faudrait refaire cette analyse sur un nombre plus important de sujets pour pouvoir valider ces résultats.

3.2 Etudes sur les RBD ("REM sleep Behaviour Disorder")

Pendant la phase de sommeil paradoxale, on a normalement une atonie : nos mouvements sont inhibés. Certaines personnes n'ont pas cette atonie, on nomme ce dysfonctionnement RBD pour "REM (Rapide Eye Mouvement) sleep Behaviour Disorder". Deux tiers des individus atteints de la maladie de Parkinson souffrent aussi de RBD. Inversement quasiment toutes les personnes ayant un RBD vont développer un syndrome parkinsonien. En effet au bout de 14 ans, 91% des patients RBD ont développé un syndrome parkinsonien [25]). Parmi les syndromes parkinsoniens développés par les RBD on trouve la maladie de Parkinson et d'autre maladies proches qui, en plus des symptômes parkinsoniens courants, comprennent d'autres troubles (le plus courant étant la démence) : il y a notamment la démence à corps de Lévy (DCL), et plus rarement l'atrophie multisystématisée (MSA). Les RBD qui n'ont pas encore développé de syndrome parkinsonien peuvent donc être considérés comme étant dans le phase prodromique d'un syndrome parkinsonien. L'analyse de leur voix peut alors donner des indications sur les marqueurs vocaux prédictifs de la maladie de Parkinson.
Une étude longitudinale a montré qu'en effectuant une analyse perceptive de la voix de 78 RBD tous les ans jusqu'au diagnostic d'un syndrome parkinsonien, on pouvait estimer le début des perturbations vocales à 7 ans avant le diagnostic pour ceux qui ont finalement développé la maladie de Parkinson, et à 15 ans avant le diagnostic pour ceux qui ont développé une DCL [26].
Une étude acoustique a quant à elle analysé quels paramètres acoustiques différaient significativement chez 16 RBD en phase prodromique par rapport à des sujets sains [4]. Les auteurs ont testé des paramètres en rapport avec la phonation, l'articulation, et la prosodie. Ils ont trouvé que l'articulation était le domaine le plus affecté, suivi de la phonation, puis de la prosodie. Parmi les paramètres acoustiques discriminants on note une irrégularité du débit (DDK reg) lors des tâches de diadococinésie, une diminution de l'énergie spectrale (RFA) lors du monologue ainsi qu'une augmentation de disfluences, et une apériodicité phonatoire (DUV) lors des voyelles soutenues (cf Tableau 1 en annexe). Il faut faire attention à ne pas interpréter ces paramètres comme étant forcément des marqueurs de prédiction de la maladie de Parkinson car les RBD pourront développer un autre syndrome parkinsonien, comme la démence à corps de Lévy ou la MSA. Or ces maladies sont accompagnées de perturbations vocales qui peuvent être légèrement différentes de celles que l'on trouve dans la maladie de Parkinson [7, 27]. Une étude complémentaire s'est focalisée sur la comparaison entre ces RBD en phase prodromique et des parkinsoniens récents [4]. Les auteurs ont noté qu'en moyenne les RBD étaient moins affectés vocalement que les parkinsoniens récents, surtout en ce qui concerne la prosodie. Les paramètres F0 SD (variation de la fréquence fondamentale) et NoP (nombre de pauses) pour le monologue et le VOT pour la DDK tâche sont les paramètres les plus discriminants quand on compare les RBD avec les parkinsoniens récents.

4 Analyses statistiques et classifications

Une fois les paramètres acoustiques vocaux extraits, des tests de significativité sont effectués (comme le test de Student pour les paramètres avec distribution gaussienne, et le test non paramétrique de Wilcoxon-Mann-Whitney pour les distributions non gaussiennes [4]) pour évaluer la différence entre les groupes (patients parkinsoniens vs sujets sains par exemple). Dans la plupart des cas, seuls les paramètres significatifs (souvent définis comme tels pour $p < 0{,}05$) sont gardés. Ensuite des mesures de correlations (coefficient de correlation de Bravais-Pearson pour les distributions gaussienns et corrélation de Spearman pour les données non normalement distribuées [4]) sont effectuées pour éliminer les paramètres redondants. Dans certaines études une analyse séquentielle de Wald est utilisée pour trouver les paramètres vocaux les plus souvent affectés dans le maladie de Parkinson et pour évaluer l'étendue des perturbations vocales pour chaque patient [4, 13].
Les paramètres ainsi préselectionnés sont ensuite utilisés par des algorithmes de classification qui vont tester les différentes combinaisons possibles afin de trouver la combinaison de paramètres qui permette de classer au mieux, de façon automatique, les sujets (patients parkinsoniens vs sujets sains par exemple). Les algorithmes de classifications les plus souvent utilisés dans ces études sont des algorithmes à apprentissage supervisé de type machines à vecteurs de support (SVM) associés à des méthodes de réseaux à fonctions de base radiales [4, 16].
Pour évaluer la performance de chaque système de classification, plusieurs mesures existent ; la plus simple et la plus répandue est le taux de réussites : "Accuracy" (Acc). Il se définit comme étant le nombre de bonnes réponses sur le nombre de personnes testées. Pour affiner l'évaluation de la performance des classification, deux autres mesures sont souvent rajoutées : la sensibilité et la spécificité. La sensibilité correspond au pourcentage de patients parkinsoniens classés comme tels par rapport au nombre total de sujets

parkinsoniens. Un test très sensible est un test qui a très peu de faux négatifs. La spécificité correspond quant à elle au pourcentage de vrais patients parkinsoniens parmi ceux qui ont été classés comme tel. Un test très spécifique est un test qui a très peu de faux positifs. Un bon test est un test qui est à la fois sensible et spécifique.
Comme on peut le voir dans le tableau 1 en annexe, les études sur le diagnostic précoce de la maladie de Parkinson par l'analyse acoustique de la voix obtiennent des taux de réussite (Acc) de 80% quand elles ne prennent en compte qu'un paramètre spécialement pertinent, par exemple VOT dans la tâche DDK [16], F0 SD dans le monologue [14], VSA ou F2i/F2u toujours dans le monologue [17]. Ces mêmes études améliorent leur taux de réussite jusqu'à 85% [14] voire 88% [16] quand elles prennent en compte plusieurs paramètres acoustiques. Dans une autre étude [4], les auteurs ont obtenu une sensibilité de 99,1% et une spécificité de 87,5% pour un classement parkinsonien récent vs sujet sain. Les mêmes auteurs ont proposé une classification RBD vs sujet sain avec une sensibilité de 96% et une spécificité de 79%. Il est aussi possible de séparer les patients parkinsoniens récents des MSA récents, mais les sensibilités obtenues sont un peu moins bonnes (de l'ordre de 60%) [7].
Les taux de réussite et les mesures de sensibilité et spécificité publiés sont à prendre avec précaution car les études en question n'ont concerné que des groupes de petite taille (une vingtaine de sujets maximum par groupe), et les bases de données ne sont pas publiques.

5 Conclusion

De ces études on peut conclure que les composantes de la parole chez les parkinsoniens récents, dans lesquelles on trouve le plus de perturbations acoustiques, semblent être la prosodie et l'articulation, même si on peut noter quelques anomalies dans la phonation et le rythme. Les tâches vocales qui les mettent le mieux en valeur semblent être le monologue (discours spontané pendant au moins 1 min), et les tâches de diadococinésie (répétition rapide de syllabes). Dans l'analyse acoustique il faut faire spécialement attention aux paramètres qui varient significativement en fonction du genre (F0, durée maximale de phonation...), le mieux étant de séparer les analyses pour les femmes et les hommes, ou de prendre des paramètres acoustiques qui ne dépendent pas du genre. Il faut également prendre des précautions par rapport aux interprétation des résultats quand les patients suivent un traitement, car les traitements pour la maladie de Parkinson peuvent avoir un impact sur certains paramètres acoustiques.

L'ensemble des résultats présentés dans cette revue de la littérature montre qu'il y a bien des marqueurs spécifiques de la maladie de Parkinson dans la voix et ce dès le début de la maladie. Certains marqueurs apparaissent même avant les symptômes moteurs qui servent au diagnostic clinique actuel. On peut donc envisager, qu'une fois ces marqueurs validés par des études longitudinales sur un nombre plus important de patients, on puisse établir aisément un diagnostic plus précoce de cette maladie. Ceci permettrait à terme à des sujets prédisposés (génétiquement par ex.) de savoir s'ils sont en train de développer la maladie de Parkinson assez tôt pour pouvoir commencer des traitements qui ralentiraient voire stopperaient l'évolution de la maladie avant que le cerveau ne soit trop endommagé. Ces marqueurs acoustiques pourraient aussi permettre aux RBD de savoir quel type de syndrome parkinsonien ils vont développer pour adapter les traitements au plus tôt.

Remerciements

Nous tenons à remercier le programme Futur & Ruptures (programme financé par l'Institut Mines-Télécom, la Fondation Télécom et l'Institut Carnot Télécom & Société Numérique) pour son soutien financier.

Références

[1] Maria C Rodriguez-Oroz, Marjan Jahanshahi, Paul Krack, Irene Litvan, Raúl Macias, Erwan Bezard, et José A Obeso. Initial clinical manifestations of Parkinson's disease : features and pathophysiological mechanisms. *The Lancet Neurology*, 8(12) :1128–1139, Décembre 2009.

[2] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, et L. O. Ramig. Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease. *IEEE Transactions on Biomedical Engineering*, 59(5) :1264–1271, Mai 2012.

[3] Brian Harel, Michael Cannizzaro, et Peter J. Snyder. Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease : A longitudinal case study. *Brain and Cognition*, 56(1) :24–29, Octobre 2004.

[4] Jan Rusz, Jan Hlavnička, Tereza Tykalová, Jitka Bušková, Olga Ulmanová, Evžen Růžička, et Karel Šonka. Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder. *Sleep Medicine*, Septembre 2015.

[5] Ronald B. Postuma. Voice changes in prodromal Parkinson's disease - is a new biomarker within earshot ? *Sleep Medicine*, Septembre 2015.

[6] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva, et Elmar Nöth. New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease. Dans Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, et Stelios Piperidis, éditeurs, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik,

Iceland, 2014. European Language Resources Association (ELRA).

[7] Young Eun Huh, Jongkyu Park, Mee Kyung Suh, Sang Eun Lee, Jumin Kim, Yuri Jeong, Hee-Tae Kim, et Jin Whan Cho. Differences in early speech patterns between Parkinson variant of multiple system atrophy and Parkinson's disease. *Brain and Language*, 147 :14–20, Août 2015.

[8] Parkinson's Voice Initiative.

[9] Véronique Rolland-Monnoury. Les troubles de la parole dans la maladie de Parkinson. *L'échos*, (104) :17–18, Septembre 2010.

[10] Sabine Skodda, Julia Lorenz, et Uwe Schlegel. Instability of syllable repetition in Parkinson's disease—Impairment of automated speech performance ? *Basal Ganglia*, 3(1) :33–37, Mars 2013.

[11] François Viallet et Bernard Teston. La dysarthrie dans la maladie de Parkinson. Dans P. Auzou, éditeur, *Les Dysarthries*, pages 169–174. SOLAL, 2007.

[12] Jan Rusz, Roman Čmejla, Hana Růžičková, Jiří Klempíř, Veronika Majerová, Jana Picmausová, Jan Roth, et Evžen Růžička. Evaluation of speech impairment in early stages of Parkinson's disease : a prospective study with the role of pharmacotherapy. *Journal of Neural Transmission*, 120(2) :319–329, Février 2013.

[13] Jan. Rusz, Roman Čmejla, Hana Růžičková, et Evžen Růžička. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The Journal of the Acoustical Society of America*, 129(1) :350, 2011.

[14] Jan Rusz, Roman Čmejla, Hana Růžičková, Jiří Klempíř, Veronika Majerová, Jana Picmausová, Jan Roth, et Evžen Růžička. Acoustic assessment of voice and speech disorders in Parkinson's disease through quick vocal test. *Movement Disorders*, 26(10) :1951–1952, Août 2011.

[15] Sabine Skodda, Wenke Grönheit, et Uwe Schlegel. Impairment of Vowel Articulation as a Possible Marker of Disease Progression in Parkinson's Disease. *PLoS ONE*, 7(2) :e32132, Février 2012.

[16] M. Novotny, J. Rusz, R. Cmejla, et E. Ruzicka. Automatic Evaluation of Articulatory Disorders in Parkinson's Disease. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(9) :1366–1378, Septembre 2014.

[17] Jan Rusz, Roman Cmejla, Tereza Tykalova, Hana Ruzickova, Jiri Klempir, Veronika Majerova, Jana Picmausova, Jan Roth, et Evzen Ruzicka. Imprecise vowel articulation as a potential early marker of Parkinson's disease : Effect of speaking task. *The Journal of the Acoustical Society of America*, 134(3) :2171–2181, Septembre 2013.

[18] Jack Jiang, Emily Lin, Jian Wang, et David G. Hanson. Glottographic Measures Before and After Levodopa Treatment in Parkinson's Disease. *The Laryngoscope*, 109(8) :1287–1294, Août 1999.

[19] Stefanie Countryman, Jennifer Camburn, et Janet Schwantz. *Parkinson's Disease : Speaking Out*. National Parkinson Foundation, 6 édition, 2003.

[20] A. Tsanas, M. A. Little, P. E. McSharry, et L. O. Ramig. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of The Royal Society Interface*, 8(59) :842–855, Juin 2011.

[21] R. Iansek, J. L. Bradshaw, J. G. Phillips, R. Cunnington, et M. E. Morris. Interaction of the basal ganglia and supplementary motor area in the elaboration of movement. Dans Denis J. Glencross and Jan P. Piek, éditeur, *Advances in Psychology*, volume 111 de *Motor Control and Sensory Motor IntegrationIssues and Directions*, pages 37–59. North-Holland, 1995.

[22] Sabine Skodda. Steadiness of syllable repetition in early motor stages of Parkinson's disease. *Biomedical Signal Processing and Control*, 17 :55–59, Mars 2015.

[23] L. O. Ramig, S. Sapir, S. Countryman, A. A. Pawlas, C. O'brien, M. Hoehn, et L. L. Thompson. Intensive voice treatment (LSVT®) for patients with Parkinson's disease : a 2 year follow up. *Journal of Neurology, Neurosurgery & Psychiatry*, 71(4) :493–498, 2001.

[24] Guide d'évaluation de l'UPDRS.

[25] Alex Iranzo, Ana Fernández-Arcos, Eduard Tolosa, Mónica Serradell, José Luis Molinuevo, Francesc Valldeoriola, Ellen Gelpi, Isabel Vilaseca, Raquel Sánchez-Valle, Albert Lladó, Carles Gaig, et Joan Santamaría. Neurodegenerative Disorder Risk in Idiopathic REM Sleep Behavior Disorder : Study in 174 Patients. *PLoS ONE*, 9(2), Février 2014.

[26] R. B. Postuma, A. E. Lang, J. F. Gagnon, A. Pelletier, et J. Y. Montplaisir. How does parkinsonism start ? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder. *Brain*, 135(6) :1860–1870, Juin 2012.

[27] Müller J, Wenning GK, Verny M, et et al. Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders. *Archives of Neurology*, 58(2) :259–264, Février 2001.

[28] Brian T. Harel, Michael S. Cannizzaro, Henrí Cohen, Nicole Reilly, et Peter J. Snyder. Acoustic characteristics of Parkinsonian speech : a potential biomarker of early disease progression and treatment. *Journal of Neurolinguistics*, 17(6) :439–453, Novembre 2004.

ANNEXE 1

auteurs	nb sujets	tâches	paramètres les + discriminants	
Harel & al. 2004 [28]	4 MP 4 SC	- DDK (papapa) - lecture phrase - monologue	- F0 SD (monologue) - durée pause (lecture)	
Rusz & al. 2011 [13]	23 MP 23 SC	- voyelles soutenues "a", "i", "u" - DDK(pataka) - lecture texte - lecture texte avec mots en MAJ - lecture phrases émotionnelles - monologue 90 sec	- F0 SD (monologue et phrase émo) - RIRV(DDK)	
Rusz & al. 2011 [14]	24 MP 22 SC	- voyelle soutenue "a" - DDK(pataka) - monologue 80 sec	- F0 SD (monologue), SPLD (DDK), RFPC (DDK),NHR("aa") -> Acc :85% - F0 SD (monologue) seul -> Acc :81,3%	
Rusz & al. 2013 [17]	20 MP 15 SC	- voyelles soutenues "a" "i" "u" - lecture de phrases - répétitions d'une même phrase - monologue	- VSA (monologue) -> Acc :80,4% - F2i/F2u (monologue) -> Acc :80,0%	
Novotny & al. 2014 [16]	24 MP 22 SC	DDK(pataka)	- VOT (pa) SNR, CST (ka), VSQ30, 2FT (ta), DDK rate -> Acc :88% - VOT seul -> Acc :80%	
Rusz & al. 2015 [4]	19 MP 19 SC	- voyelle soutenue "a" - DDK(pataka) - monologue	- VOT (DDK), F0 SD (monol.), NoP (monol.), HNR("aa") sont les plus discriminants - F0 SD (monol.), Int SD (monol.), NoP (monol.) -> Se :99,1 % , Sp :87,5%	
Skodda 2015 [22]	50 MPt 32 SC	répétition de "pa" ou "pa-ti" à un rythme choisi ou imposé	- COV - pa-ti ratio	
Huh & al. 2015 [7]	29 MP 26 MSA 37 SC	- voyelle soutenue "a" - lecture de phrases	"aa"	- F0 (hom.) (MSA/SC et MSA/MP-> Se :57, Sp :87) - MPT (fem.)(MSA/SC et MP/SC)
			lecture (hom.)	- F0 SD et PRww (MSA/SC et MP/SC) - TSR et TPT (MSA/SC et MSA/MP-> Se :64, Sp :73)
Harel & al. 2004 [3]	1 pré-MP 1 SC	- monologue *étude longitudinale*	- F0 SD (à partir de 5 ans avant le diagnostic)	
Postuma & al. 2012 [26]	78 RBD 39 SC	- monologue (UPDRS item18) *étude longitudinale*	analyse perceptive -> différences à partir de 7 ans avant diagnostic pour MP et 15 ans pour DCL	
Rusz & al. 2015 [4]	16 RBD 16 SC	- voyelle soutenue "a" - DDK(pataka) - monologue	- DDK reg, RFA, DUV, PDW -> Se :96 % , Sp :79% - DDK rate et DUV encore + discriminants quand UPDRS>4	
Rusz & al. 2015 [4]	19 MP 16 RBD	- voyelle soutenue "a" - DDK(pataka) - monologue	MP/RBD : - VOT (DDK), F0 SD (monol.), NoP (monol.)	

Tableau 1 – *Inventaire des études sur les marqueurs vocaux qui pourraient contribuer à un diagnostic précoce de la maladie de Parkinson. La 1ere partie fait l'inventaire des études sur des parkinsoniens récemment diagnostiqués (< 4 ans après diagnostic). La 2^{e} partie concerne la phase préclinique de la maladie de Parkinson. MP : Sujet avec Maladie de Parkinson sans traitement, MPt : Sujet parkinsonien avec traitement, SC : sujet contrôle, pré-MP : parkinsonien dans la phase préclinique, MSA : patient avec atrophie multisystématisée, RBD : sujet atteint de Rapid eye movement sleep Behaviour Disorder, DCL : patient avec Démence à Corps de Lévy, DDK :tâche de diadococinésie, UPDRS : Unified Parkinson's Disease Rating Scale, Acc : accuracy, Se : sensibilité, Sp : spécificité. Pour les abréviations des paramètres acoustiques cf Tableau 2.*

ANNEXE 2

paramètre	description	tâche
Phonation		
HNR	Harmonics-to-Noise Ratio : amplitude du bruit par rapport aux composantes tonales	voyelle soutenue
MPT	Maximum Phonation Time : durée maximale de phonation	voyelle soutenue
F0 SD	Standard Deviation (écart type) de la fréquence fondamentale (F0)	voyelle soutenue
DUV	Degree of Unvoiced Segment : fraction des segments silencieux (<0,45)	voyelle soutenue
jitter	Variabilité de F0 d'un cycle à l'autre	voyelle soutenue
shimmer	Variation relative de l'amplitude entre 2 cycles consécutifs	voyelle soutenue
Articulation		
VOT	Voice Onset Time : durée d'une consonne occlusive	DDK
DDK rate	Taux diadococinésie (DDK) : nombre de syllabes par seconde	DDK
DDK reg	Régularité DDK : écart type des distances entre 2 centres syllabiques consécutifs	DDK
RFA	Resonant Frequency Attenuation	DDK
SNR	Signal-to-Noise Ratio	DDK
1FT	First Formant Trend : régression de la fréquence du 1er formant	DDK
2FT	Second Formant Trend : régression de la fréquence du 2^e formant	DDK
VSQ	Vowel Similarity Quotient : autocorrélation de la voyelle sur sa durée totale -> régularité	DDK
VSQ30	VSQ sur une durée de 30ms à partir du début de la voyelle	DDK
VVQ	Vowel Variability Quotient : variabilité dans les longueurs des voyelles	DDK
CST	Consonent Spectral Trend : régression du spectre de la consonne	DDK
RIRV	Relative Intensity Range Variation : variation relative de l'intensité	DDK
RRIS	Robust Relative Intensity Slope	DDK
SDCV	Spectral Distance Change Variation	DDK
RFPC	Robust Formant Periodicity Correlation	DDK
VSA	Vowel Space Area : aire du triangle vocalique	monologue
VAI	Vowel Articulation Index = (F1a+F2i)/(F1i+F1u+F2a+F2u)	monologue
F2i/F2u	Rapport des 2^e formants des voyelles "i" et "u"	monologue
Prosodie		
F0 SD	Standart Deviation de F0 : variation de la hauteur (intonation)	monologue
Int SD	Standart Deviation de l'Intensité après avoir enlevé les silences de plus de 60ms	monologue
NoP	Nb de Pauses par rapport au temps parlé total après avoir enlevé les silences < 60 ms	monologue
PDW	Percentage of Disfluent Words : nombre de disfluences sur nombre total de mots	monologue
rythm	Mesure de la capacité à reproduire les rythmes d'une lecture après écoute	lecture texte rythmé
Rythme		
COV	Coefficient of Variation : variation relative du rythme	répét. rythmée de syll.
pa-ti ratio	Intervalle pa-ti par rapport à intervalle pa-pa	répét. rythmée de syll.

Tableau 2 – *Description des paramètres acoustiques discriminants pour le diagnostic précoce de la maladie de Parkinson d'après la littérature, et les meilleurs tâches vocales qui permettent de les extraire. DDK : tâche de diadococinésie.*

Performances de compression HEVC et H.264 dans le contexte du temps réel : Application en Télémédecine

Amine Chaabouni[1], Yann Gaudeau[2,3] et Jean-Marie Moureaux[1]
[1]Université de Lorraine, CRAN, UMR 7039, 9 Avenue de la Foret de Haye Vandoeuvre les Nancy, 54500, France
e-mail: amine.chaabouni@univ-lorraine.fr,
jeanmarie.moureaux@univ-lorraine.fr
[2]Université de Strasbourg, 30 Rue du Maire Andre Traband, Haguenau, 67500, France
e-mail: yann.gaudeau@unistra.fr
3CRAN, UMR 7039, France

Résumé

La télémédecine apparait aujourd'hui comme une solution indispensable pour garantir la portabilité des données médicales de haute qualité. Afin d'offrir ces différents traitements en temps réel, une compression avec perte est nécessaire, réduisant ainsi la taille des données numériques tout en conservant l'information utile. Malgré l'efficacité de la compression H264, cette norme reste toujours limitée surtout pour les réseaux mobiles et bas débits. Nous essayons alors d'apporter une nouvelle solution à travers cette étude, en se basant sur des tests de performances du nouveau standard HEVC dans le contexte du temps réel pour la télémédecine. Ces tests permettent d'évaluer la qualité objective d'une vidéo médicale full HD après transmission, utilisant différentes vitesses de calculs CPU à des débits de compression inférieurs à 4 Mbits/s. Les résultats montrent la possibilité d'utiliser l'encodeur x265 dans un contexte temps réel avec un gain en qualité par rapport à x264, permettant ainsi de réduire le débit de transmission de 4 à 1 Mbits/s pour un PSNR identique.

Mots clefs

Standard HEVC, Compression H.264, Qualité objective, Temps réel, Streaming.

1 Introduction

La télémédecine est aujourd'hui un enjeu majeur en matière de santé. D'une part, elle représente une solution permettant l'accès aux soins à tous les citoyens à travers le territoire, en particulier ceux qui ont des difficultés de déplacement ou une offre de soins limitée (du fait des problèmes de démographie médicale). Elle offre alors la garantie d'une prise en charge de qualité par les grands centres hospitaliers de référence (CHU, CHR), à travers autant d'outils que sont la téléconsultation, la télé-radiologie, ou encore le suivi à distance. D'autre part, elle représente un outil collaboratif précieux dans les scénarios professionnel à professionnel, c'est à dire impliquant plusieurs médecins ou professionnels de santé sur des sites distants. Ainsi par exemple, les réunions de concertation pluridisciplinaires sont particulièrement importantes, en particulier pour le diagnostic et la prise en charge de patients atteints de cancers. Le partage d'expérience et d'expertise est reconnu aujourd'hui comme un gage d'amélioration du diagnostic et de la prise en charge thérapeutique, notamment pour les cancers les plus difficiles à traiter.

Toutes ces raisons font de la télémédecine au sens large un enjeu très important du présent mais également de l'avenir. Cependant, si la télémédecine est amenée à se développer de façon significative dans les années à venir, elle pose encore certains problèmes scientifiques et technologiques, en particulier ceux liés aux grandes masses de données mises en jeu. Ainsi par exemple, l'imagerie radiologique qu'elle soit de type scanner, IRM ou encore pet-scan met en jeu des volumes de données de plusieurs centaines de Mégaoctets pour un seul examen. Les flux vidéo stéréoscopiques d'un robot de chirurgie haute définition représentent quant à eux aujourd'hui 2 x 1,5 Gbits/s à transmettre en temps réel. De même, le flux vidéo haute définition d'une caméra endoscopique utilisée en chirurgie ORL pour l'ablation de tumeurs est codé à environ 2 Gbits/s. Ces chiffres sont aujourd'hui un frein au développement de la télémédecine en routine. Afin de compenser ce défaut, la compression avec perte des données médicales devient une solution incontournable, à condition qu'elle n'affecte pas la qualité médicale des données pour une utilisation régulière par les professionnels de la santé.
En effet, la compression dans le contexte médical était toujours effectuée sans perte quand elle existait car elle était le seul type de compression toléré par les médecins. Cette compression sans perte garantit ainsi l'intégrité des données et permet d'éviter les erreurs de diagnostic. Cependant, ce type de compression ne permet pas de réduire significativement le volume des données. Hier encore inenvisageable, la compression avec pertes semble aujourd'hui de mieux en mieux acceptée par les médecins [1,2,3], comme l'exemple de la Canadian Association of Radiologists (CAR) [2] qui estime que les techniques de

compression avec pertes peuvent être utilisées à des taux raisonnables, sous la direction d'un praticien qualifié, sans aucune réduction significative de la qualité de l'image pour le diagnostic clinique.
Afin de trouver un compromis entre l'efficacité de la compression et la perception des experts de la qualité médicale de la vidéo après compression, il est important d'effectuer des tests subjectifs dédiés à l'évaluation de la qualité, en d'autres termes, estimer l'impact de la compression sur la vidéo par rapport à l'utilisation. Ces tests ont été programmés et validés pendant les travaux faits dans le cadre du projet européen HIPERMED (HIgh PERformance teleMEDicine platform) [3], élu meilleur projet européen à l'événement EUREKA INNOVATION en Novembre 2014. Durant ce projet, les médecins ORL (oto-rhino-laryngologie) participants aux différents scénarios, ont exprimé leurs satisfactions de la qualité des vidéos, compressées avec le standard H.264 et transférées à distance et en temps réels.
Pour améliorer les résultats démontrés lors du projet HIPERMED, nous essayons, à travers ce papier, de tester les performances de la compression avec le nouveau standard HEVC dans le contexte du streaming en temps réel. Ainsi, cet article est organisé comme suit : tous les outils et les méthodes utilisés durant les tests de streaming sont présentés au paragraphe 2. Une présentation de l'évaluation objective de la qualité mise en œuvre est donnée au paragraphe 3. Le paragraphe 4 est consacré à l'analyse des résultats de performance. Enfin, nous concluons et présentons les perspectives de ce travail au paragraphe 5.

2 Outils de tests de streaming

Afin de mettre en œuvre les tests de streaming, nous avons eu recours à la plateforme FFMPEG qui supporte les standards de compression H264 et HEVC.

2.1 FFMPEG

Pour simuler un environnement de streaming, il nous faut 3 éléments clés pour transférer le flux vidéo d'un point à un autre : Un encodeur afin de compresser le flux avant le transfert, un serveur qui permet de stocker le flux envoyé et un lecteur vidéo qui permet de décoder ce flux après réception. Ainsi, nous avons choisi d'utiliser la plateforme FFMPEG (www.ffmpeg.org), contenant les 3 éléments :
- *ffmpeg* : un outil qui permet de convertir les fichiers multimédia (vidéo dans notre cas) entre différents formats. Cet outil sera utilisé comme encodeur.
- *ffserver* : un server de streaming multimédia pour le temps réel.
- *ffplay* : un lecteur multimédia basé sur la bibliothèque SDL (Simple Directmedia Layer).

2.2 H.264

Durant les scénarios HIPERMED, les flux vidéo ont été compressés en utilisant le standard de compression H264. Le choix de la norme H.264 est dû à sa performance améliorée dans l'encodage vidéo, fournissant plus d'efficacité et de flexibilité, que les normes précédentes, pour différentes applications dans une très large variété de supports de réseau, des systèmes et des protocoles de transport. En fait, H.264 est conjointement développé par l'UIT-T et l'ISO / CEI et il est le produit d'un effort de partenariat connu sous le nom Joint Video Team (JVT). Il offre plus de flexibilité de partitionnement de bloc pour l'estimation de mouvement (16 × 16, 16 x 8, 8 x 16, 8 x 8, 8 x 4, 4 x 8, 4 x 4), un plus grand nombre d'images de référence dans la structure du GOP (Group Of Picture). Il utilise la compensation de mouvement avec une précision allant jusqu'à un quart de pixel de résolution et deux codeurs entropiques (CABAC et CAVLC) [4].
La plateforme FFMPEG supporte plusieurs encodeurs et décodeurs. Parmi ces codecs, on utilise l'encodeur/ décodeur *x264*.

2.3 HEVC

Afin d'améliorer les résultats validés durant les tests HIPERMED, nous allons utiliser la nouvelle norme de codage HEVC, considérée comme le successeur de la norme H.264 élaborée et finalisée le 25 Janvier 2013, par l'équipe JVT.
Cette nouvelle norme de compression vidéo peut supporter l'ultra haute résolution 4K (3840 x 2160) et 8K (7680 x 4320). Elle permet le traitement parallèle de l'architecture multi-core. Par rapport à H.264, elle est basée sur l'unité codage CTU (Coding Tree Unit), offrant un partitionnement en macro-blocs avec de grandes tailles (16, 32 ou 64), offrant ainsi, plus d'efficacité et de flexibilité. En outre, HEVC utilise 3 types de filtres: filtre de déblocage adaptatif, filtre de décalage d'échantillon (SAO) et le filtre de filtrage adaptif en boucle (ALF) contrairement à H.264 qui utilise juste le filtrage de déblocage. Enfin, son codeur entropique utilise seulement le codage arithmétique (CABAC) [5].
Parmi les codecs supportés par FFMPEG, nous allons utiliser l'encodeur *x265* et son décodeur associé *hevc*.
Nous essayons alors, à travers cette étude, de comparer l'efficacité du standard HEVC par rapport à H.264 dans un contexte de télémédecine. Cette comparaison de performance sera basée sur 3 critères essentiels dans le contexte des tests de streaming temps réel :
- La qualité de la vidéo, en calculant les métriques objectives.
- Les ressources de traitement CPU utilisées durant le codage et décodage.
- La latence produite par la compression dans le contexte du temps réel.

3 Evaluation de la qualité objective

Après compression, nous avons deux types de tests possibles à réaliser pour évaluer la qualité de la vidéo. Comme première solution, on cite les tests subjectifs. Malgré leur importance, ils sont considérés comme des tests couteux, très longs et lourds à mettre en œuvre. Ainsi, cette solution est souvent difficile à réaliser humainement. Pour éviter un tel inconvénient, les métriques objectives, qui représentent la deuxième alternative, permettent de définir des mesures de qualité qui soient fortement corrélées aux notes de qualité, données par les experts observateurs.

Les premières métriques, comme le MSE (Mean Squared Error) et le PSNR (Peak Signal to Noise Ratio), sont utilisées pour mesurer la similitude de l'image en utilisant une comparaison mathématique simple entre l'image de référence et l'image compressée. Basés sur la comparaison pixel par pixel, ces critères sont simples et leurs performances restaient limitées. En outre, parmi les métriques les plus connues et utilisées pour l'évaluation objective de la qualité, on trouve SSIM (Structutal SIMilarity index) [6] et son extension MSSSIM [7]. Ces critères sont basés sur l'information structurelle, extraite du stimulus, qui contient la distorsion perçue. D'autres métriques objectives récentes ont vu le jour ces dernières années, comme le PSNR-HVS [8] et HDR-VDP [9], des outils psychophysiques qui permettent à mieux comprendre le comportement du Système Visuelle Humain (SVH) et affiner les modèles associés. De plus, nous citions le critère UQI (Universal Quality Index) qui estime la distorsion comme une combinaison de changements de la luminance, le contraste et une perte de corrélation entre l'image originale et l'image compressée. Par ailleurs, d'autres critères permettent de calculer la similitude de deux images plutôt que la qualité perçue de l'image, sur la base des statistiques des scènes naturelles (NSS). Par exemple, IFC (Information Fidelity Criterion) [10], VIF (Visual Information Fidelity) et VIFP [11]. Le lecteur pourra se reporter à la référence [12] pour plus d'information sur ces métriques.

Dans ce contexte, nous utilisons la librairie Matlab Metrix Mux [13].

4 Tests et résultats

Après avoir défini les différents outils et les méthodes utilisés lors de ces tests de streaming, nous allons présenter, dans la suite, le résumé des résultats.

4.1 Environnement de tests

Afin de faire les tests de streaming dans les meilleures conditions, nous avons choisi de travailler sur deux machines performantes en termes de calculateurs de processus et de mémoire RAM. Les configurations de ces machines sont présentées dans le tableau suivant.

	Système d'exploitation	RAM	Processeur	Nombre de processeurs
Machine 1	Linux mint 17.2	16 Go	I7, 2.7Ghz	8 intel cores
Machine 2	Linux mint 17.2	8 Go	I7, 4 Ghz	8 intel cores

Tableau 1. Configurations des machines de tests

Comme vidéo de test, nous avons utilisé une séquence originale d'une vidéo endoscopique acquise, d'une opération chirurgicale ORL, au Centre Hospitalier Régional Universitaire CHRU de Nancy. Cette vidéo dure 40s, d'une résolution full HD (1920x1080 - 1080p60 - 4:2:2 - 8 bits) avec un débit de transmission originale égal à 1.99 Gbits/s. Elle a été choisie comme une des séquences de référence de l'imagerie médicale pour le développement de la norme HEVC par l'équipe de collaboration sur les Video Coding (JCT-VC) [14] (voir figure 1).

La transmission de cette vidéo a été faite à travers un réseau privé LAN 1Gbit/s dédié pour ces tests de streaming. Pour se mettre dans le contexte des réseaux bas débits, nous avons décidé de compresser la séquence médicale sur 7 différents débits allant de 0.5 à 4 Mbits/s. Afin de satisfaire la contrainte du temps réel, nous avons utilisé le mode de transmission « zerolatency ».

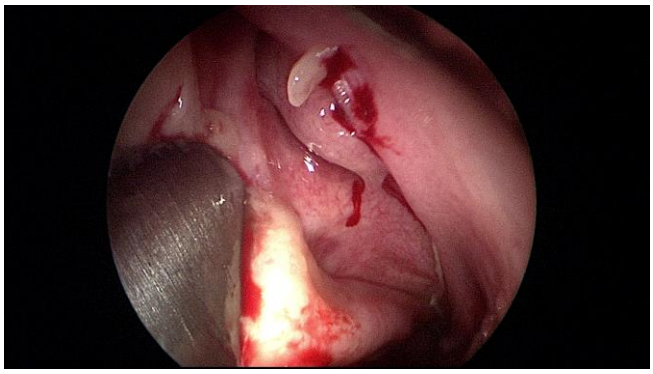


Figure 1 – *Séquence de test (1920x1080 - 1080p60 - 4:2:2 - 8 bits)*

4.2 Résultats

L'objectif de ces tests est de calculer les performances des deux encodeurs *x264* et *x265*, dans le contexte du streaming temps réel. Pour cela, nous devons déterminer 3 critères essentiels : la qualité de la vidéo après compression, les ressources des calculateurs CPU consommées durant la transmission et la latence (le délai) produite par la compression.

4.2.1 Ressources CPU consommées

Afin de quantifier les ressources des calculateurs CPU consommées pour la compression et la lecture après transmission, nous avons déterminé 2 critères (le pourcentage de CPU dédié et le temps de l'utilisation système), associés à l'encodeur et au décodeur.

Dans ce contexte, nous avons compressé la séquence vidéo avec l'option « ultrafast ». Ce paramètre, appelé « preset », permet de définir une certaine vitesse d'encodage par rapport aux taux de compression. Dans notre cas, le mode « ultrafast » fourni la vitesse maximale, satisfaisante à la contrainte de la transmission HEVC en temps réel (http://x265.readthedocs.org). Nous calculons alors les 2 critères des deux encodeurs *x264* et *x265* en se basant sur ce mode.

Comme nous montrent les figures 2 et 3, l'encodeur *x265* est très gourmand en termes de ressources CPU par rapport à *x264*. Pour un débit de 1 Mbits/s, *x265* consomme plus que 450% du CPU, c'est-à-dire 7 fois plus de ressources que x264 qui n'utilise, à peu près, que 60%. La même chose pour le décodeur hevc, où le lecteur consomme 16% de CPU, le double par rapport au décodeur *x264*.

Afin de réduire cette consommation, nous avons testé le mode « fastdecode » développé par *x265*. Cette option nous permet de réduire la quantité de ressources utilisée pour la compression *x265*, en désactivant les fonctions de l'encodeur, comme par exemple désactiver les filtres adaptatifs en boucle ALF, qui tendent à produire le blocage pour le décodeur (http://x265.readthedocs.org). Cette réduction de consommation intervient plus dans la partie haut débit comme on le voit sur les figures 2 et 3.

Concernant l'encodeur *x264*, nous avons fait des tests supplémentaires pour déterminer le mode qui consomme une quantité de CPU équivalente à celle de *x265*. Ce mode de processus est le preset « veryslow », testant plusieurs options d'encodage, comme par exemple la taille des blocs CTU, la taille minimale des blocs CU et le nombre des trames bidirectionnelles au sein du GOP, en utilisant plus de calculs pour obtenir la meilleure qualité par rapport au débit de compression sélectionné.

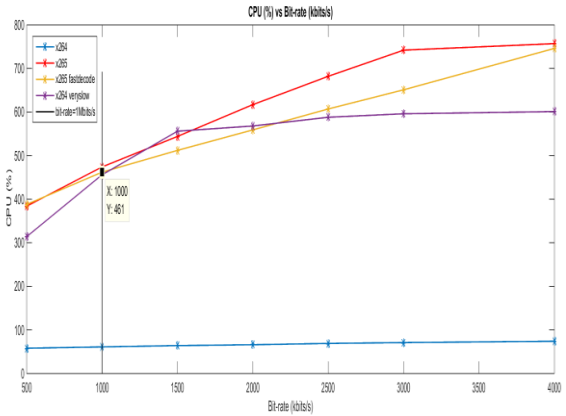


Figure 2 – *Ressources CPU consommées par l'encodeur*

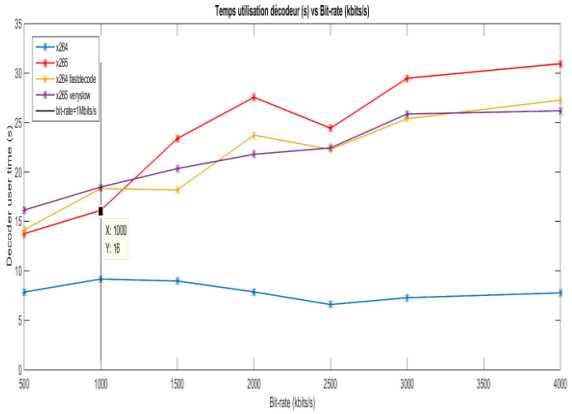


Figure 3 – *Temps d'utilisation par le décodeur*

4.2.2 Qualité objective

A travers la figure 4a et 4b, nous remarquons que la qualité de la vidéo compressée avec l'encodeur *x265* est meilleure que celle compressée avec l'encodeur *x264*. Ce résultat est validé par les métriques objectives.

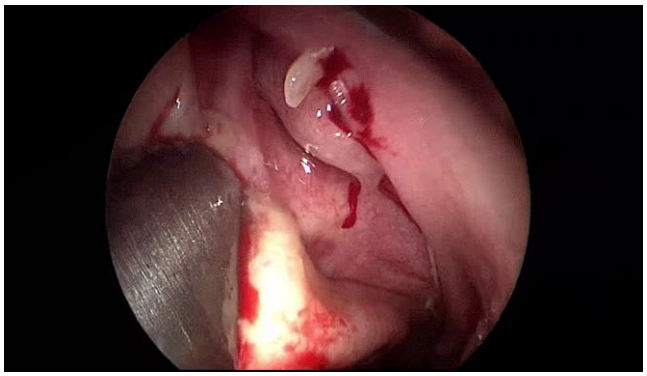


Figure 4.a – *Vidéo médicale compressée avec x265 à un débit de 1 Mbits/s*

Figure 4.b - *Vidéo médicale compressée avec x264 à un débit de 1 Mbits/s*

En se basant sur les résultats des tests de corrélation entre les tests subjectifs et objectifs, faits durant le projet HIPERMED [3], nous avons sélectionnés 4 critères parmi

les critères les plus performants et les plus corrélés avec les notes d'observation des experts ORL. Ces 4 métriques sont : PSNR, MSE, NQM et MSSSIM. Les valeurs associées sont affichées dans les tableaux 2,3,4 et 5, correspondants à *x265*, *x265* mode « fastdecode », *x264* et *x264* « veryslow ». Ces tableaux confirment la supériorité de x265 par rapport à *x264* et *x264* « veryslow » pour toutes les métriques à un débit donné.

Comme nous montre la figure 5, pour une même qualité de PSNR (PSNR=34.12), correspondante à la compression *x264* à 4 Mbits/s, nous pouvons par exemple réduire le débit de compression 4 fois moins, c'est-à-dire à 1 Mbits/s comme débit de compression *x265*.

Débit (kbits/s)	500	1000	1500	2000	2500	3000	4000
PSNR	33.4368	34.1357	34.5259	34.8724	35.1864	35.4912	36.0741
MSE	31.7052	27.1354	24.7344	22.6168	20.9733	19.4939	16.9928
NQM	29.5388	31.7134	32.7981	33.9755	34.2279	34.6076	35.1413
MSSSIM	0.8735	0.8828	0.8905	0.8981	0.9031	0.9085	0.9177

Tableau 2. Métriques objectives pour la transmission x265 (mode « zerolatency »)

Débit (kbits/s)	500	1000	1500	2000	2500	3000	4000
PSNR	33.2652	33.9885	34.3793	34.7323	35.0302	35.3511	35.9372
MSE	32.8819	28.1210	25.7322	23.7539	22.0676	20.4105	17.7940
NQM	29.5703	31.7787	32.7500	33.9853	34.1557	34.7345	35.8768
MSSSIM	0.8662	0.8768	0.8838	0.8910	0.8966	0.9028	0.9128

Tableau 3. Métriques objectives pour la transmission x265 (mode « fastdecode »)

Débit (kbits/s)	500	1000	1500	2000	2500	3000	4000
PSNR	28.8517	31.7533	32.7516	33.2065	33.4895	33.7466	34.1197
MSE	87.0518	45.2934	36.4014	32.9070	30.7466	29.0127	26.6321
NQM	19.1587	25.7236	28.7721	30.2445	31.2548	32.0896	33.0851
MSSSIM	0.7501	0.8207	0.8446	0.8543	0.8600	0.8649	0.8714

Tableau 4. Métriques objectives pour la transmission x264 (mode « ultrafast »)

Débit (kbits/s)	500	1000	1500	2000	2500	3000	4000
PSNR	30.6545	32.8900	33.6220	34.0027	34.2759	34.5235	34.9646
MSE	58.9569	35.8454	30.4237	27.9460	26.2567	24.8121	22.3974
NQM	23.3452	28.8141	31.1360	32.4214	33.3124	34.0135	35.1197
MSSSIM	0.8352	0.8678	0.8796	0.8863	0.8916	0.8960	0.9035

Tableau 5. Métriques objectives pour la transmission x264 (mode « veryslow »)

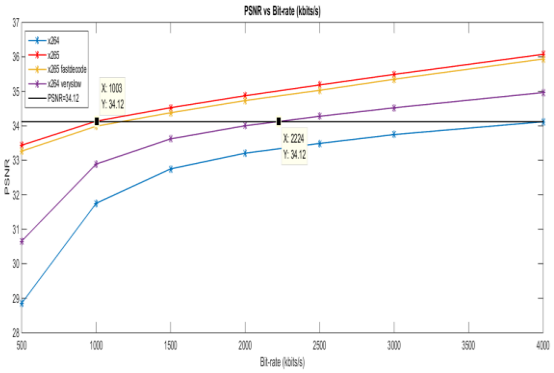


Figure 5. *Qualité PSNR par rapport au débit de compression*

Cette même figure nous montre que le mode « fastdecode » utilisé pour *x265*, génère une qualité comparable au mode « zerolatency ».
Nous pouvons aussi conclure que *x265* est toujours meilleur que *x264*, à une même consommation de ressources CPU.
Ces résultats sont validés par les autres métriques comme nous montrent les tableaux 2, 3,4 et 5.

4.2.3 Latence

Comme troisième critère de performance dans nos tests, nous calculons le délai produit par la compression *x264* et *x265*. Pour cela, nous avons calculé la durée de transmission associée à chaque encodeur. La figure 6 nous montre les résultats.
A 1 Mbits/s, pour transmettre 40s d'une vidéo médicale, les encodeurs *x264* et *x265* ne produisent presque pas de latence (40.05s). Cependant, à un débit supérieur à 3 Mbits/s, *x265* commence à générer un grand délai, inacceptable dans notre contexte de temps réel. Ce qui constitue une limite des performances de cet encodeur par rapport à x264.

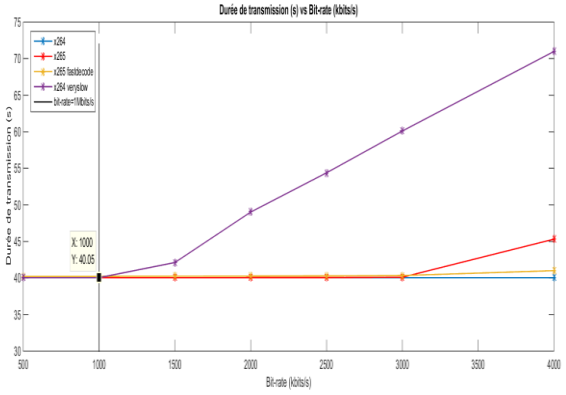


Figure 6. *Durée de transmission par rapport au débit de compression*

5 Conclusion

Dans ce papier, nous avons essayé de tester les performances de la compression du nouveau standard HEVC et de les comparer par rapport à celles de la norme H264. Ces différents tests ont pour objectif de valider l'utilisation de HEVC dans le contexte de la transmission des vidéos médicales en temps réel. Nous avons pu démontrer à travers les différents résultats, que *x265* est très efficace en termes de qualité pour les bas débits, permettant de réduire le débit de compression, jusqu'à 4 fois moins par rapport à *x264* pour ces vidéos de chirurgie ORL. En d'autres termes, à une qualité égale, on pourra alors passer de 4 Mbits/s comme débit de transmission, d'une vidéo médicale full HD, en *x264* à 1 Mbits/s en utilisant l'encodeur *x265*. Cependant, les tests nous ont démontré que l'encodeur *x265* est très gourmand en calculs CPU et peut consommer jusqu'à 7 fois plus de CPU que *x264*. Des tests supplémentaires nous ont montré que *x265* nous donne toujours une meilleure qualité que *x264*, à une consommation comparable de CPU. Finalement, ces tests de streaming nous ont confirmé la possibilité de l'utilisation de *x265* dans le contexte du temps réel pour les bas débits inférieurs à 3 Mbits/s.
Dans les prochains travaux, nous essayerons de transmettre d'autre type de vidéos médicales en utilisant l'encodage matériel HEVC afin de diminuer la consommation de CPU. Cette solution devra être validée par les tests subjectifs qui restent plus importants que les métriques objectives dans l'évaluation de la qualité des vidéos.

6 Remerciements

Cette étude a été menée dans le cadre du projet européen Celtic E3, la suite du projet HIPERMED. En conséquence, nous remercierons nos partenaires polonais de la compagnie PSNC, spécialement Piotr Pawalowski, Piotr Szymaniak et Wojciech Kapsa, pour leur aide à monter ces tests de streaming. Les résultats de ce travail reposent sur la plate-forme PROMETEE, dont les outils ont permis de réaliser les mesures objectifves. Ainsi, nous remerciement Julien Lambert pour son support durant ces tests. Finalement, nous voulons remercier Dr. Gallet, du CHRU de Nancy, qui nous a permis d'enregistrer la vidéo médicale utilisée durant ces derniers tests.

Références

[1] N.Nouri, D. Abraham, J-M. Moureaux, M. Dufaut, J.Hubert, M. Perez. "Subjective MPEG2 compressed video quality assessment: Application to Tele-surgery," 7th IEEE International Symposium on Biomedical Imaging, ISBI 2010, Rotterdam, 2010.

[2] Y.Gaudeau et J.M. Moureaux, "Lossy compression of volumetric medical images with 3d dead-zone lattice vector quantization", Annals of Telecommunication, vol.64, no 5-6, 2009.

[3] A. Chaabouni, Y. Gaudeau, J. Lambert, J.M. Moureaux, P. Gallet, "Subjective and objective quality assessment for H.264 compressed medical video sequences", International Conference on Image Processing Theory, Tools and Applications, pp. 18-22, 2014

[4] Heiko Schwarz, Detlev Marpe, et Thomas Wiegand, Overview of the Scalable Video Coding Extension of the H.264/AVC Standard , IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 17, NO. 9, Septembre 2007.

[5] G.J. Sullvian, J.R. Ohm, T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard", IEEE Transactions on circuits and systems for video technology, vol.22, no.12, Décembre 2012.

[6] Z.Wang, A.C.Bovik, H.R.Sheikh, E.P.Simoncelli. "Image quality assessment: From error visibility to structural similarity," IEEE Transactions of Image Processing, vol.13, no. 4, pp 600-612, Avril 2004.

[7] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multi-Scale Structural Similarity for Image Quality Assessment," in Proceedings of the 37th IEEE Asiloma Conference on Signal, Systems and Computers, Pacific Grove, CA, 2003

[8] K.Egiazarian, J.Astola, N.Ponomarenko, V.Lukin, F.Battisti, M.Carli. "Two new fullreference quality metricss based on HVS," Workshop on Video Processing and Quality Metricss, Scottsdale USA, 2006.

[9] R. Mantiuk, K. Kim, Allan G. Rempel et W. Heidrich. "HDR-VDP-2: A calibrated visual metrics for visibility and quality predictions in all luminance conditions," ACM Transactions on Graphics (Proc. of SIGGRAPH'11), vol.30, no. 4, 2011.

[10] H. R. Sheikh, A. C. Bovik et G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics," IEEE Transactions on Image Processing, vol. 14, no. 12, pp. 2117-2128, 2005

[11] H. R. Sheikh et A. C. Bovik, "Image Information and Visual Quality," IEEE Transactions on Image Processing, vol. 15, no. 2, pp. 430-444, 2006.

[12] M. Gaubatz, "Metrix Mux Visual Quality AssessmentPackage," disponible sur foulard.ece.cornell.edu.

[13] A.Moorthy, L. Choi, A. Bovik, Fellow, G. de Veciana. "Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies," IEEE Journal of selected topics in Signal Processing vol.6, no.6, Octobre 2012.

[14] D. Nicholson P. Pawałowski et J-M. Moureaux, "Selected medical imaging sequences for HEVC development," Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 - 15th Meeting: Geneva, CH, 23 Oct. – 1 Nov. 2013