Détection et extraction de texte de la vidéo*

Christian Wolf et Jean-Michel Jolion

Laboratoire Reconnaissance de Formes et Vision
Bât. Jules Verne
20, Avenue Albert Einstein
Villeurbanne, 69621 Cedex

tél : 04.72.43.60.54
fax : 04.72.43.80.97

email : {wolf,jolion}@rfv.insa-lyon.fr
web : http ://telesun.insa-lyon.fr/∼wolf

Résumé :
Nous présentons dans cet article les premières étapes d'un projet visant à la détection et la reconnaissance du texte présent dans des images ou des séquences vidéo. Nous insisterons ici sur la caractérisation de ce type de texte en regard des textes présents dans les documents classiques. Nous proposons un schéma de détection s'appuyant sur la mesure du gradient directionnel cumulé. Dans le cas des séquences, nous introduisons un processus de "fiabilisation" des détections et l'amélioration des textes détectés par un suivi et une intégration temporelle.

Mots-clés : indexation de vidéo, OCR, détection de texte

1 Introduction

Depuis quelques années, les documents audiovisuels numérisés sont de plus en plus fréquents. De grandes bases de données audiovisuelles ont été créées par des entreprises, des organisations et aussi par des personnes privées. Cependant, l'utilisation de ces bases reste encore problématique. Tout particulièrement, ces nouveaux types de données (images et vidéos) ont conduit à de nouveaux systèmes d'indexation où la recherche par le contenu se fait à partir d'une image exemple.

Les systèmes disponibles actuellement travaillent sans connaissance (systèmes pré-attentifs). Ils utilisent des méthodes de traitement d'images pour extraire des caractéristiques de bas niveau (couleur, texture, forme, etc.). Malheureusement les requêtes construites à partir de ces caractéristiques ne correspondent pas toujours aux résultats obtenus par un humain qui interprète le contenu du document.

*Cette étude est soutenue par France Télécom Recherche & Développement dans le cadre du projet ECAV.

La cause de cet échec est le manque de sémantique dans l'information extraite. Que représente la sémantique ? En cherchant dans le dictionnaire Larousse on trouve : *"Sémantique : les études de sens des mots"*. Par analogie, dans le contexte de l'image, la sémantique représente donc le sens de l'image.

Considérons l'image de la figure 1a. Un utilisateur humain qui choisit cette image comme image requête est probablement intéressé par des images de cyclistes (notons que ne nous pouvons pas deviner le vrai désir de l'utilisateur et par conséquent, la similarité entre deux images n'est donc pas accessible). D'autre part les systèmes d'indexation trouvent des images qui sont similaires par rapport aux caractéristiques de bas niveau. Les images contenant des cyclistes qui ne sont pas similaires par rapport à ces caractéristiques (voir figure 1b) ne seront pas trouvées.

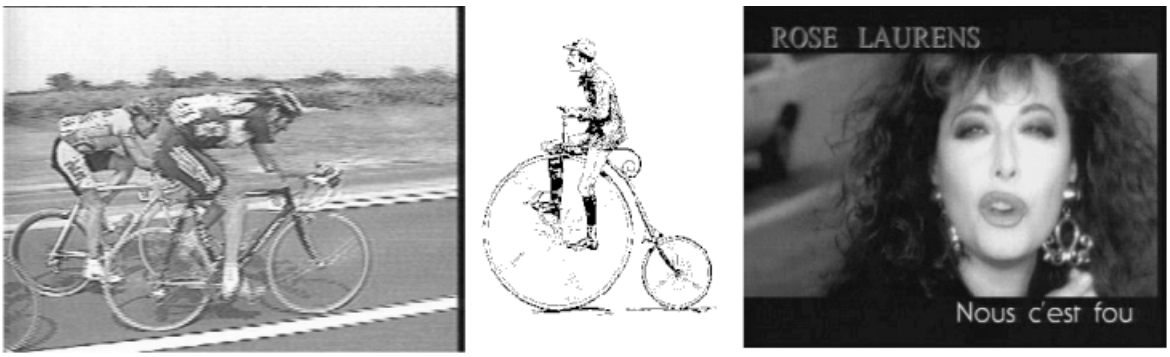

FIG. 1 – Images d'exemples.

Entre le niveau du pixel et celui de la sémantique, il existe des caractéristiques à la fois riches en information et cependant simples. Le texte présent dans les images et les vidéos fait partie de cette catégorie. La figure 1c montre une image extraite d'une publicité fournie en requête par un utilisateur. On peut imaginer que celui-ci cherche l'image de la chanteuse. La sémantique de cette image exemple peut être résumée par : *"Une publicité pour le nouvel album de la chanteuse Rose Laurens avec une photo de son visage"*. Il est impossible de déduire cette information à partir des caractéristiques basses. Pourtant dans ce cas le nom de la chanteuse est présent dans l'image. Par conséquent, en extrayant le texte on obtient une information supplémentaire très valable. Les requêtes classiques peuvent donc être complétées par des mots-cléfs. Le texte automatiquement extrait peut être stocké. Le processus que nous étudions dans ce travail concourt donc à une meilleure automatisation de la phase d'indexation d'images et de vidéos.

2 Formulation du problème

Les recherches en détection et l'extraction du texte à partir des séquences vidéo sont encore confrontées à de sérieux problèmes. Pourtant la recherche dans le domaine de l'OCR des documents classiques (c'est-à-dire les documents écrits, les journaux etc.) a développé de bonnes méthodes et des outils commerciaux produisent de bons résultats pour des images de documents. Le problème principal peut être expliqué par la différence entre l'information présente dans un document et celle donnée par une séquence vidéo ainsi que les méthodes de stockage de chaque type. Nous allons décrire ces différences dans les paragraphes suivants.

Les images de documents sont créées en vue de les numériser pour les passer ensuite à une phase "OCR" pour reconnaître la structure et le texte. Pour améliorer la qualité et le taux de la reconnaissance, les images sont numérisées à très haute résolution (200-400 dpi), ce qui donne des fichiers de taille très élevée (un fichier de 100 Mo résultant d'une page A4 numérisée à 400 dpi). Les fichiers sont comprimés sans perte pour garder la qualité et empêcher des artefacts de compression. Ces grandes tailles ne sont pas une limite puisque ni leur transport ni leur stockage ne sont prévus. La plupart du temps les images de documents sont bien structurées et contiennent un fond uniforme et la couleur du texte est également uniforme. Ceci permet de séparer les caractères du fond avec un seuillage fixe ou adaptatif. La majorité de la page contient des caractères structurés dans différents paragraphes, quelques images étant incluses dans la page.

Par contre les images des séquences vidéo contiennent de l'information plus difficile à traiter. Le texte n'est pas séparé du fond. Il est soit superposé (le "texte artificiel" comme les sous-titres, les résultats de sport, etc.) soit inclus dans la scène de l'image (le "texte de scène", par exemple le texte sur le tee-shirt d'un acteur). Le fond de l'image peut être très complexe ce qui empêche une séparation facile des caractères. De plus, contrairement aux documents écrits, les séquences vidéo contiennent de l'information très riche en couleurs. Enfin, le texte n'est pas structuré en lignes, et souvent quelques mots courts et déconnectés flottent dans l'image.

Le but principal de la conception des formats de fichiers vidéo est de garder une qualité suffisante pour l'affichage, pour le stockage et le transport par des réseaux informatiques. Pour cette raison et pour une quantité de données vidéo plus importante, il est nécessaire de réduire fortement l'information avant son stockage dans le fichier vidéo. Pour limiter la taille de l'information deux méthodes sont souvent appliquées : la forte réduction de la résolution et le codage avec perte, c'est-à-dire avec élimination des données redondantes et perte d'une partie de l'information originale, en utilisant les algorithmes de compression *JPEG* et *MPEG*.

En conséquence, il est clair que les méthodes et les résultats développés pour les documents traditionnels ne peuvent être utilisés pour les documents audiovisuels. Nous allons discuter les problèmes principaux dans les paragraphes suivants.

La résolution basse

La résolution spatiale de la vidéo dépend de l'application et prend des valeurs entre 160×100 pixels (diffusion par Internet) et 720×480 pixels (DVD vidéo codée en MPEG 2). Un format typique est le format CIF avec 384×288 pixels. Notons que la taille du texte affiché avec cette résolution est beaucoup moins grande que le texte résultant d'un document numérisé. Les *frames* d'une vidéo de cette résolution contiennent des caractères d'une hauteur de moins de 10 pixels, alors que dans les documents numérisés, les caractères ont une hauteur comprise entre 50 et 70 pixels.

Ce sont également les différences de taille de la police qui rendent le processus plus difficile. La figure 2a montre une image contenant des polices entre 8 pixels et 80 pixels. Cette variété de tailles pose des problèmes pendant la phase de détection, induisant la nécessité d'une approche multi-résolution. D'un autre côté les polices de petites tailles rendent difficile la phase de reconnaissance des caractères.

Les artefacts d'*anti-aliasing* et de la compression

Pendant la production de la vidéo, le signal original est sous-échantillonné plusieurs fois. Pour empêcher des effets d'*aliasing*, des filtres passe-bas sont appliqués. Le signal résultant a alors une qualité suffisante pour la lecture mais il est complètement lissé. La figure 2b

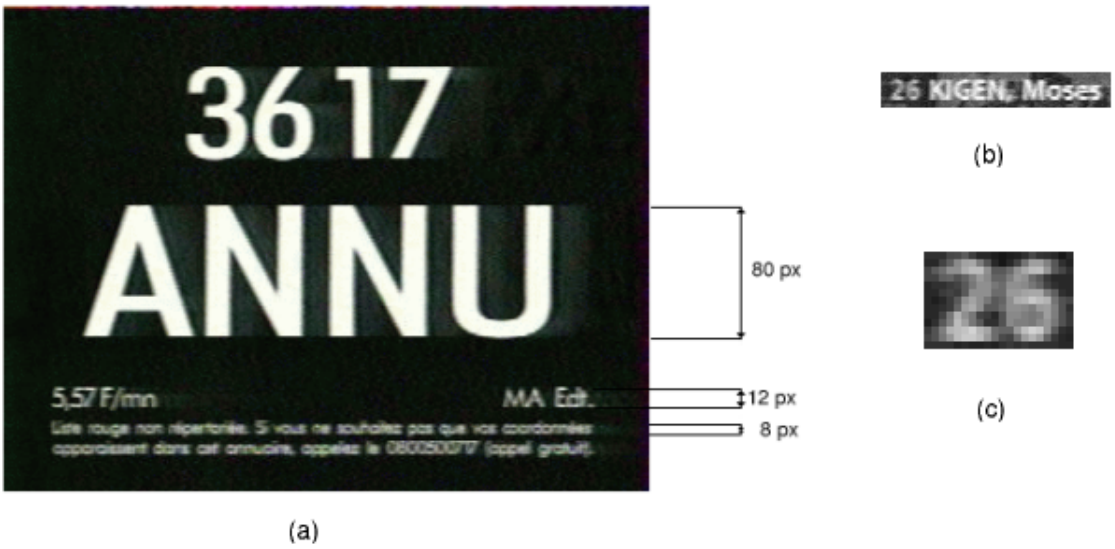

FIG. 2 – Les petites polices et la grande différence de tailles (a). Artefacts d'*anti-aliasing* (b,c).

montre un exemple d'image extraite d'une émission de sport. La hauteur de la police du texte affiché est seulement de 11 pixels, pourtant les caractères sont suffisamment lisibles. La figure 2c montre deux caractères agrandis. Le lissage du signal rend la segmentation des caractères très difficile.

Après la réduction de la résolution, la compression du signal ajoute également des artefacts. L'information supprimée par le schéma MPEG est considérée comme redondante pour le système visuel humain. Cependant, les artefacts de l'encodage cause des problèmes pendant la reconnaissance (par exemple en perturbant l'uniformité des couleurs). Même si les nouvelles normes comme JPEG 2000 apportent un plus en regard de ce type de problèmes, l'existence de très nombreuses données codées selon le format JPEG classique justifie nos recherches dans cette direction.

Le fond complexe et détaillé
Le fond sur lequel est inscrit le texte ne peut être considéré comme constant. Le plus souvent, un seuillage fixe, même determiné localement, n'est pas suffisant pour clairement mettre en évidence le texte comme le montre l'exemple des figures 3a et 3b.

Le rehaussement artificiel du contraste
Afin de faciliter la lecture du texte, des artifices d'inscription du texte sont utilisés. Ces techniques ne sont malheureusement pas un avantage pour la détection comme le montrent les figures 3 où le texte est artificiellement rehaussé par l'ajout d'un soulignement.

3 État de l'art

L'extraction du texte des flux de vidéo est un domaine de recherche très récent mais cependant très fécond. La plupart des travaux traite le problème de la détection et de la localisation du texte (nommé "détection"). Ainsi, il existe très peu de recherches sur la reconnaissance, surtout à cause des problèmes présentés dans la section 2.

Les premières méthodes pour la détection ont été proposées comme des généralisations

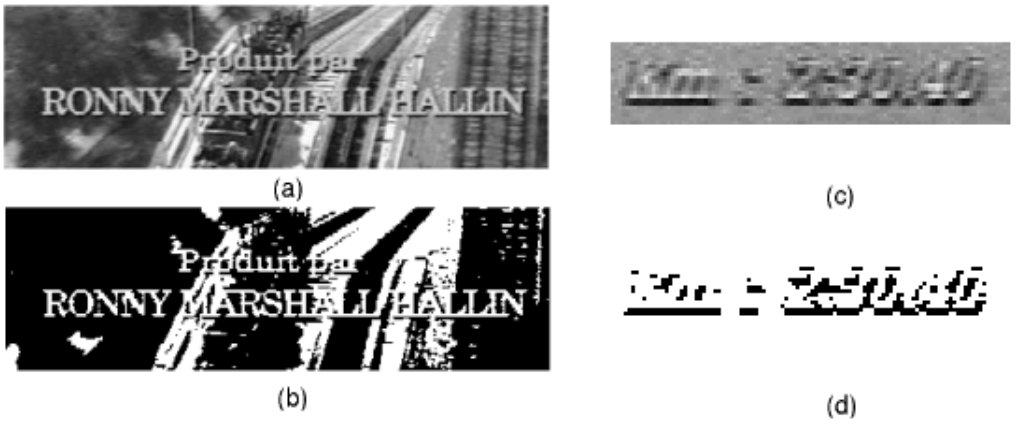

FIG. 3 – Texte sur un fond complexe (a). Un essai de seuillage fixe (b). Rehaussement artificiel du contraste (c). Un essai de seuillage fixe (d).

de techniques du domaine de l'OCR dans le cadre de l'analyse automatique des documents. La problématique de la détection est en effet proche de celle de l'analyse de la structure d'un document contenant du graphisme (journaux, page web etc, ...). A titre d'exemple, on peut citer les travaux de Jain et al. [2]. En supposant que la couleur des caractères est uniforme, Jain et al. utilisent une réduction des couleurs suivie d'une segmentation et d'une phase de regroupement spatial pour trouver les caractères du texte. Bien que la présence des caractères joints est prévue par les auteurs, la phase de segmentation pose des problèmes dans le cas de documents de mauvaise qualité et notamment les séquences vidéo de basse résolution.

Une approche similaire, qui donne des résultats impressionnants sur le texte de grande police, est présentée par Lienhart et al. [5]. Ils sont d'ailleurs parmi les premiers qui ont présenté une étude aussi complète partant de la détection jusqu'à la reconnaissance et l'indexation. Étant mieux adapté aux propriétés de la vidéo, l'algorithme proposé combine une phase de détection par segmentation des caractères et la détection par une recherche de contraste local élevé. Ensuite une analyse de texture supprime les fausses alarmes. Le suivi du texte est effectué au niveau des caractères. Pour la reconnaissance les auteurs utilisent un produit commercialisé. Malheureusement ils ne pouvaient pas montrer l'aptitude de l'algorithme de segmentation appliqué aux textes de petite taille.

La méthode de Sato, Kanade et al. [6] repose sur le fait que le texte est composé de traits de contraste élevé. L'algorithme de détection cherche des contours groupés en rectangles horizontaux. Les auteurs reconnaissent la nécessité d'augmenter la qualité de l'image avant la phase d'OCR. Par conséquent, ils effectuent une interpolation pour augmenter la résolution de chaque boîte de texte suivie par une intégration de plusieurs *frames* (prenant les valeurs minimale/maximale des niveaux de gris). La reconnaissance est réalisée par une mesure de corrélation. Wu, Manmatha et Riseman [8] combinent la recherche des contours verticaux et l'application d'un filtre de texture.

Le travail de LeBourgeois [3] est basé sur l'accumulation des gradients horizontaux. La reconnaissance dans le système proposé est effectuée par des règles statistiques sur les caractéristiques de projections des niveaux de gris.

Li et Doermann présentent une approche d'apprentissage [4]. En laissant glisser une sous-fenêtre sur l'image, ils utilisent un réseau de neurones de type MLP pour classifier chaque

pixel comme “texte” ou “non texte”. Les caractéristiques sont extraites des échelles de hautes fréquences d'une ondelette de Haar. Clark et Mirmehdi utilisent aussi un réseau de neurones pour la détection du texte [1]. Ils extraient des caractéristiques diverses comme la variance de l'histogramme, la densité des contours etc.

Une méthode qui travaille directement dans le domaine de la vidéo comprimée est proposée par Zhong, Zhang et Jain [9]. Leurs caractéristiques sont calculées directement à partir des coefficients de DCT des données MPEG. La détection est basée sur la recherche des variations horizontales d'intensité, suivie par une étape de morphologie mathématique.

Les différentes approches reflètent l'état de l'art dans le domaine de *vidéo OCR*. La plupart des problèmes sont identifiés, mais pas encore résolus. Surtout, l'aspect d'amélioration du contenu, bien qu'abordé par Sato, Kanade et al. [6] et par Li et Doermann [4], ne donne pas encore des résultats tout à fait satisfaisants. Nous considérons aussi la segmentation des caractères du fond complexe comme un problème de grande importance, qui demande encore une recherche intensive.

4 Un système d'extraction

Nous pouvons décrire le but principal d'un système d'extraction de texte par quelques mots : accepter des fichiers d'images et de vidéo, détecter le texte, l'extraire et produire un fichier ASCII incluant le texte dans un format utilisable pour l'indexation. En tenant compte de la problématique expliquée précédemment, la structure d'un système proposée dans la figure 4 s'impose naturellement.

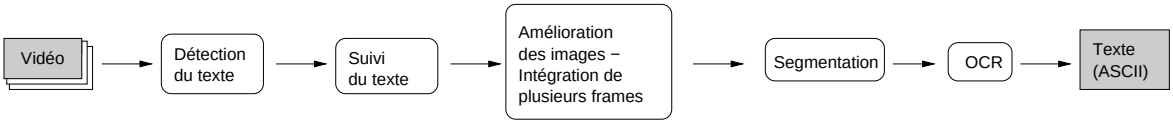

Fig. 4 – Le fonctionnement général du système.

La détection du texte est réalisée dans chaque *frame* de la vidéo[1]. Les rectangles figurant la localisation du texte sont suivis pendant leur période d'apparition pour associr les rectangles se correspondant dans les différents *frames*. Cette information est nécessaire pour améliorer le contenu de l'image, ce qui peut être atteint par l'intégration de plusieurs rectangles contenant le même texte. Cette phase doit produire des sous-images d'une qualité en accord avec les prérequis d'un processus OCR. Par conséquent il est aussi nécessaire d'augmenter la résolution, en utilisant l'information supplémentaire prise dans la séquence d'images. La phase finale avant la reconnaissance par un algorithme d'OCR s'occupe de la segmentation des caractères en tenant compte de la complexité du fond de l'image.

Dans cette section nous proposons un système de détection et de suivi du texte. Notre système est capable de détecter et de localiser le texte dans une image ou dans un *frame*, et de suivre les occurrences du texte dans une séquence vidéo. Nous ne présenterons que peu de détails dans cette section (uniquement les principes généraux) car cette recherche est effectuée dans le cadre d'un contrat et les premiers résultats donnent lieu à dépôt de brevets[2].

[1] Un sous-échantillonnage de la vidéo peut accroître la rapidité du système mais cela réduit la facilité du suivi des zones de texte.

[2] Une démonstration en ligne pour les images statiques est accessible à l'adresse :
http ://telesun.insa-lyon.fr/∼wolf/demos/textdetect.html

Les perspectives que nous pouvons envisager pour notre projet sont donc fondées sur le développement d'un système complet qui sera capable d'extraire le texte, de le reconnaître et enfin de l'utiliser pour l'indexation.

4.1 Détection

Notre algorithme de détection repose sur une succession de traitements. Dans un premier temps, on utilise le fait que les caractères du texte forment une texture régulière contenant des contours verticaux allongés horizontalement. Une première mesure met en évidence les pixels de l'image conformes à cette hypothèse. Afin de passer des pixels à des zones compactes, nous utilisons ensuite une binarisation suivie d'un post-traitement afin d'extraire les rectangles englobants des zones de texte. Cette dernière étape permet d'atteindre plusieurs buts :

- Réduire le bruit.
- Corriger des erreurs de classification à partir de l'information du voisinage.
- Connecter des caractères afin de former des mots complets.

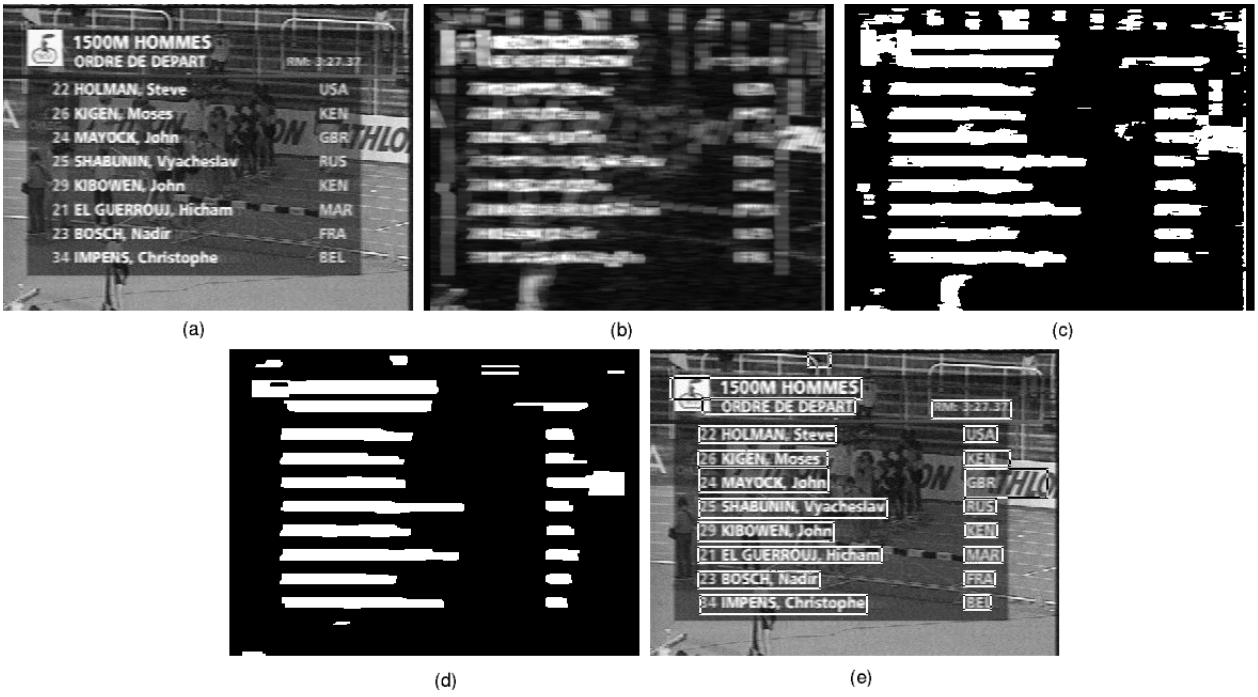

FIG. 5 – Les résultats intermédiaires pendant la détection : l'image d'entrée (a), les mesures de probabilité de présence d'un texte (b), l'image binarisée (c), l'image après le post-traitement (d), les rectangles finaux (e).

L'effet principal de cette phase réside dans la séparation du texte de la plupart des fausses alarmes. Ensuite, nous effectuons une analyse de chaque composante connexe pour trouver celles qui correspondent aux zones de texte. A ce stade, nous passons du traitement global de l'image au traitement local par composante. Chaque région est vérifiée en imposant des contraintes sur sa géométrie de façon à encore réduire le nombre de fausses alarmes. A ce stade, une suite de tests est effectuée afin de prendre en compte les spécificités de polices particulières ou d'effets d'incrustation.

A partir de ce point le traitement des composantes de texte s'effectue au niveau de leurs boîtes englobantes seulement. Pour cela, nous avons choisi le terme "*rectangle*". Le résultat de la détection est donc une liste de rectangles. Dans le cadre global du système, nous construisons cette liste des rectangles détectés pour chaque *frame* de la séquence. Les listes sont passées au module de suivi pour détecter les occurrences de texte pendant plusieurs *frames* de la vidéo.

La figure 5 montre des images de résultats intermédiaires pendant les différentes phases de la détection. L'image d'entrée est prise dans une émission de sport. L'image de probabilité de présence d'un texte (figure 5b) montre la forte réponse du filtre sur les zones de texte, mais également sur les endroits possédant des textures régulières (la grille). Bien que ce bruit reste présent dans l'image binarisée (figure 5c), la phase de post-traitement a réussi à le supprimer (figure 5d). Le résultat final est affiché sur la figure 5e.

4.2 Suivi du texte

L'algorithme de détection produit une liste de rectangles détectés par *frame*. Le but de notre module de suivi est l'association des rectangles correspondants afin de produire des apparitions de texte pendant plusieurs *frames* de la vidéo. Ceci est réalisé en gardant une liste L des apparitions actuelles dans la mémoire, initialisée avec les rectangles du premier *frame*. Pour chaque *frame* i, nous comparons la liste des rectangles L_i détectés avec cette liste pour vérifier si les rectangles détectés font partie d'une apparition déjà existante.

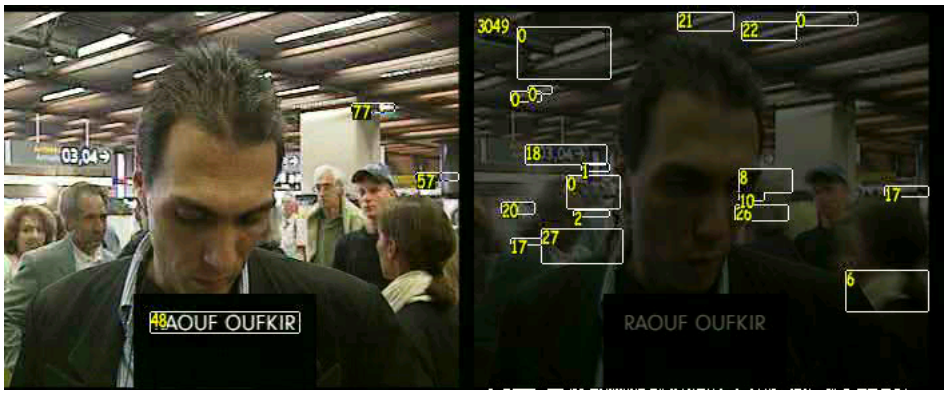

FIG. 6 – Suppression des fausses alarmes.

Les résultats de la phase de suivi sont des occurrences du texte contenant de l'information sur la localisation du texte dans chaque *frame*. La longueur de l'apparition, c'est-à-dire le nombre de *frames* où elle apparaît, nous sert comme mesure de stabilité du texte. En supposant que chaque occurrence de texte doit rester à l'écran un minimum de temps, nous considérons les apparitions de longueur plus courte qu'un seuil fixe comme des fausses alarmes. La figure 6 montre dans l'image gauche les rectangles considérés comme stables d'un *frame* exemple, et dans l'image droite les rectangles associés aux apparitions trop courtes, donc considérés comme des fausses alarmes.

4.3 Résultats

Pour l'évaluation de la détection statique dans chaque *frame* nous avons créé une base de test. La table 1 présente le taux de détection atteint. Nous avons imposé des critères d'évaluation

stricts, ce qui entraîne que les rectangles qui n'ont pas été détectés entièrement, n'ont pas été comptabilisés dans le taux de reconnaissance. La figure 7 montre quelques exemples de texte que nous considérons comme détectés. La suite de notre travail impliquera également la création d'une base de vérité terrain pour l'évaluation du processus temporel.

Nombre de rectangles dans la vérité terrain	339
Nombre de rectangles détectés	306
Pourcentage	90.48%

TAB. 1 – Les résultats.

5 Temps d'exécution

Le temps d'exécution de notre algorithme est présenté dans la table 2. Le système est implementé en C++ (non optimisé) sur Linux et le temps est spécifié pour un processeur Intel Pentium III de 700 Mhz. Les chiffres donnés correspondent à l'exécution d'un *frame*.

Décodage MPEG + conversion d'image	0.05 s
Intégration initiale + stockage interne	0.12 s
Détection	0.46 s
Suivi	0.002 s
Total	0.64 s

TAB. 2 – Temps d'exécution par *frame*.

6 Conclusion

Le travail présenté dans cet article constitue la première phase de notre projet. Les différents éléments de l'état actuel de notre système doivent être optimisés mais la structure générale est maintenant stable. La continuité de ce travail se situe dans la phase de reconnaissance. Nous avons déjà procédé à des essais sur les principaux outils commerciaux de reconnaissance de caractères [7]. De cette étude, nous pouvons déduire les caractéristiques minimales que doivent valider les textes inclus dans la vidéo pour être reconnus. Si les caractéristiques sont trop contraignantes, il sera alors nécessaire de développer un outil spécifique de reconnaissance.

Références

[1] P. Clark et M. Mirmehdi. Combining Statistical Measures to Find Image Text Regions. In IEEE Computer Society, editor, *Proc. of the ICPR*, pages 450–453, 3 septembre 2000.

[2] A.K. Jain et B. Yu. Automatic Text Location in Images and Video Frames. Technical Report MSU-CPS-97-33, PRIP Lab., Department of Computer Science, 1997.

[3] F. LeBourgeois. Robust Multifont OCR System from Gray Level Images. In *Proc. of the 4th Int. Conf. on Document Analysis and Recognition*, pages 1–5, août 1997.

FIG. 7 – Quelques résultats de détection statique.

[4] H. Li et D. Doerman. A Video Text Detection System based on Automated Training. In IEEE Computer Society, editor, *Proc. of the ICPR*, pages 223–226, 3 septembre 2000.

[5] R. Lienhart et W. Effelsberg. Automatic Text Segmentation and Text Recognition for Video Indexing. Technical report, Univ. of Mannheim, Prakt. Informatik IV, 1998.

[6] T. Sato, Takeo Kanade, E.K. Hughes, M.A. Smith, et S. Satoh. Video OCR : Indexing digtal news libraries by recognition of superimposed captions. Manuscrit Multimedia Systems.

[7] C. Wolf et J.M. Jolion. ECAV - Enrichissement de Contenu Audio-Visuels. Technical Report ECAV-1, Laboratoire Reconnaissance de Formes et Vision, 26 octobre 2000.

[8] V. Wu, R. Manmatha, et E.M. Riseman. Finding Text In Images. In ACM, editor, *Proc. 2nd ACM Int. Conference on Digital Libraries*, juillet 1997.

[9] Y. Zhong, H. Zhang, et A.K. Jain. Automatic Caption Localizatio in Compressed Video. *IEEE PAMI*, 22(4) :385–392, avril 2000.