Suivi et amélioration de textes issus de génériques vidéos

D. Marquis[1] S. Bres[1]

[1] LIRIS-RFV (Laboratoire de Reconnaissance de Formes et Vision)

INSA de Lyon, Bât. Jules Verne,
20, avenue Albert Einstein, 69621 Villeurbanne cedex – France

{dmarquis, sbres}@rfv.insa-lyon.fr

Résumé

Cet article traite de l'extraction automatique d'information dans les documents audio visuel. La quantité de documents de ce type qui sont produits et archivés chaque jour est en effet considérable. Pour pouvoir exploiter les archives colossales ainsi créées, il faut être capable de savoir ce qui est stocké, et où le retrouver. Pour cela, il faut créer des indexes les plus précis possibles. Les volumes de données à traiter montrent clairement l'intérêt d'une indexation automatique et donc d'une extraction automatique d'informations.

L'information particulière traitée ici est le texte contenu dans les images et plus particulièrement les textes en mouvements de translation des génériques de films ou vidéos. Ces textes sont localisés dans les images successives, puis intégrés pour en améliorer la qualité, avant de les transmettre à un OCR chargé de les reconnaître. Cela permet d'obtenir des informations très difficiles à obtenir autrement.

Mots clefs

Détection de Texte, Suivi de Mouvement, Génériques de vidéos, Recalage d'Images, Super-résolution.

1 Introduction

La quantité de documents audio visuels créés chaque jour est sans cesse grandissante. L'exploitation de ces documents nécessite de savoir ce qu'ils contiennent, qui y apparaît, qui les a fait … De plus, tous ces documents, ou au moins la plus grande partie d'entre eux, sont archivés. Pourtant, cet archivage n'a d'intérêt que si l'on peut savoir ce que contient l'archive et où trouver le ou les documents contenant une information particulière. Il est donc impératif de créer des indexes, contenant le maximum d'informations possible. Le volume de données à traiter montre que l'extraction automatique d'informations est ici indispensable.

Dans ce contexte, l'extraction d'informations dans les images et les vidéos est devenue une activité de recherche de plus en plus importante. Parmi toutes les informations, de nature très différentes, que l'on peut retirer de ces documents audio-visuels, nous nous sommes plus particulièrement intéressés aux textes qu'ils contiennent. Parmi ces textes, on distingue généralement les *textes artificiels*, ajoutés a posteriori dans la vidéo, et les *textes de scène* qui correspondent à tous les textes que l'on trouve dans la vidéo d'origine (panneaux de signalisation routière, enseignes de magasins, mots sur des affiches …).

L'étude présentée ici fait suite à'une coopération[1] avec France Télécom, et concerne plus particulièrement les textes artificiels. Ces textes sont le plus souvent là pour exprimer une information importante de façon très synthétique. L'intérêt de cette extraction est de fournir des informations complémentaires sur l'image, la scène ou la séquence vidéo, informations qui permettront d'affiner la "compréhension" et donc l'indexation du document. En effet, les génériques de documents vidéos tels que des documentaires, films ou émissions diverses, sont souvent très riches d'une information permettant d'aider l'indexation du document en entier. Il en va bien sûr de même pour la plupart des textes que l'on trouve dans un document vidéo comme par exemple les sous-titres nommant les personnes interrogées lors d'un journal télévisé. Ce texte donne accès à de l'information qui serait très complexe à extraire d'une autre façon (par reconnaissance de visage par exemple pour le nom d'une personne).

Une première étude, déjà mené par notre laboratoire dans le cadre de la collaboration avec France Télécom (projet ECAV), traitait de la localisation, de l'extraction et de la reconnaissance des textes artificiels statiques dans les vidéos. Cependant, les images issues de séquences vidéos standards sont d'une qualité telle qu'il n'est pas possible de les traiter directement, au moyen d'un OCR, pour en reconnaître le texte, même si celui-ci à été correctement localisé par une boite englobante, par exemple. En effet, ces images ont une résolution assez faible (liée aux standards PAL ou NTSC) et ont subit une étape de compression. Il faudra donc en améliorer la qualité au

[1] Cette étude est financée par France Télécom R&D dans le cadre des contrats ECAV et ECAV2 (Enrichissement de Contenus Audio-Visuels)

préalable. La solution retenue pour cela est généralement l'intégration temporelle.
Les problèmes sont les mêmes dans le cas de textes en mouvement (générique déroulant de fin de film ou de fin d'émission) avec en plus une difficulté supplémentaire pour l'intégration puisqu'il faut superposer les boites englobantes. Nous aborderons donc pour commencer, le problème de la localisation des textes dans les images, puis le problème du suivi des textes dynamiques.

2 Les textes artificiels statiques

2.1 Localisation

Nous commencerons par parler de la localisation des textes artificiels statiques. La figure 1 en donne un exemple. L'extraction de ce type de texte dans la vidéo repose généralement sur la détection dans l'image de zones de hautes fréquences ou hautes énergies, (par ondelettes par exemple [1,2] ou directement à partir du codage MPEG [3]), qui sont caractéristiques du texte. Le plus souvent, c'est une information monochrome qui est utilisée mais la couleur peut aussi avoir son importance dans la ségmentation [4]. Une autre approche possible consiste à calculer des gradients dans l'image, dont on peut déduire une image de force de contours dans une fenêtre, par exemple [5]. L'utilisation de la multi-résolution permet de s'affranchir dans une certaine mesure des contraintes d'échelles.

Figure 1 - *Un exemple d'image de video contenant un texte artificiel statique.*

Nous utiliserons, quant à nous, une approche différente, développée par C. Wolf et JM. Jolion dans le cadre du projet ECAV [6]. Son principe consiste à calculer les gradients cumulés sur l'image noir et blanc sur laquelle on veut détecter le texte et à appliquer diverses opérations de dilatations et érosions, conditionnelles ou non, avec des masques favorisant les blocs de pixels horizontaux, direction privilégiée pour le texte (artificiel).

2.2 Quelques améliorations

Une fois les zones candidates détectées, il est intéressant de filtrer un certain nombre de fausses alarmes qui souvent correspondent à des parties de l'image ne contenant pas de texte mais dont les caractéristiques sont similaires à une zone de texte. Il se peut également qu'il s'agisse de textes très petits. Dans ce cas, on peut se demander s'il est réellement pertinent d'extraire ces zones que le spectateur de la scène ne peut percevoir et qui s'avéreront probablement bien trop petites pour pouvoir être utilisées dans un programme de reconnaissance de caractères.
Les méthodes utilisées reposent généralement sur un principe qui veut que le texte apparaissant à l'écran ait une taille ou une forme minimale. C. Wolf et JM. Jolion [6] font ainsi intervenir les contraintes géométriques suivantes : les textes sont plus allongés que hauts et il ne faut donc conserver que les boîtes dont le rapport largeur sur hauteur est supérieur à un certain seuil ; de même ne seront conservées que les boîtes de texte dont le rapport du nombre de pixels de texte sur l'aire de la boîte englobante est supérieur à un autre seuil.

3 Les textes artificiels dynamiques

Maintenant que nous pouvons localiser les textes statiques dans les images, nous allons nous intéresser aux textes artificiels dynamiques, comme les génériques défilants, ou de façon générale les génériques dans lesquels les mouvements du texte se limitent à des translations, à l'exclusion de rotations ou effets de zoom. Remarquons que cette catégorie regroupe plus de 95% des génériques.

3.1 Le suivi du texte

La plus grande partie des méthodes de suivi utilise le « block matching » [2,5] et teste différentes translations possibles autour des boites englobantes. On peut aussi calculer des signatures particulières sur les blocs pour les caractériser et les retrouver dans les images suivantes. Parmi ces signatures possibles, les projections horizontales et verticales [5], l'analyse des composantes connexes [3] ou l'utilisation de points d'intérêt [7,8].
Nous avons, quant à nous, utilisé le calcul du *spectre croisé* pour déduire la translation intervenue entre deux images. Ce calcul de translation est réalisé de la manière suivante :

- Sur chaque image de la vidéo, on extrait un carré de 256x256 pixels au centre de l'image. Ce carré est la fenêtre utilisée pour le suivi. Nous y reviendrons.
- On calcule le spectre de ce carré par FFT ; on procède de même avec une image de référence (la première image sur laquelle nous détectons du texte). Cette image est recalculée périodiquement, dès qu'un mouvement de plus de 64 pixels est détecté.

- On réalise la division terme à terme des deux spectres qui donne le *spectre croisé*.
- La transformée de Fourier de ce spectre croisé donne une image dont le point d'intensité maximale correspond à la translation entre les deux images. Le calcul de translation envisagé ici est donc un calcul *global* et non boîte (de texte) par boîte.

Bien évidemment, utiliser un carré directement extrait de la vidéo est impossible dès lors que le fond du générique n'est plus uni. En effet, le recalage se ferait alors principalement sur les pixels du fond qui sont majoritaires. Nous fabriquons donc une image à partir des boîtes de texte détectées dans l'image originale de la vidéo en les recopiant à leur position d'origine dans un carré au fond noir uni (voir figure 2). L'image de référence est construite de la même façon.

Figure 2 - *Génération du carré utilisé pour le calcul de recalage entre deux images. On extrait les boîtes de texte puis on les recopie sur fond noir.*

On obtient ainsi des images dans lesquelles d'une part la majeure partie du fond a été supprimée et d'autre part dont les positions des boîtes de texte sont conservées et accentuées. Ces deux facteurs contribuent à augmenter de manière assez significative le rapport signal/bruit lors du calcul de spectre croisé. En outre, ce calcul de translation, de manière globale sur toutes les boîtes de texte plutôt que boîte par boîte contribue encore à améliorer ce rapport signal/bruit. L'intérieur des boîtes a son importance pour le recalage, contrairement à ce qu'il pourrait paraître : les variations légères des positions des boîtes sont compensées par leur contenu.

3.2 L'amélioration du texte

Passons maintenant à l'étape suivante, motivée par le fait que le texte que l'on extrait des différentes images d'une vidéo est, sauf cas près particulier, de très mauvaise qualité. Même si il peut être lu par le spectateur, il ne peut pas être reconnu de façon automatique par un OCR (voir figure 3). Il faut remédier à cela, par intégration des images extraites dont les mouvements ont été compensés. Le calcul de l'image intégrée se fait selon une méthode d'intégration robuste, déjà développée et utilisée par C. Wolf et JM. Jolion dans [6] pour le texte statique. Cette méthode consiste à calculer deux images statistiques. La première est une image « moyenne », dont la luminance de chaque pixel est obtenu en calculant la moyenne des luminances des pixels des images d'origine. La seconde est une image d'écarts types dont les valeurs des pixels correspondent aux écarts-types constatés en chaque pixel de l'image moyenne. A partir de ces deux images, on va calculer une image interpolée d'un facteur 4 dans les deux directions. Cette interpolation est comparable à une interpolation bilinéaire dans laquelle la couleur d'un point est fonction de la couleur des quatre points voisins pondérés par l'inverse de la distance à ces voisins. Les pondérations utilisées ici prennent en compte l'écart type calculé précédemment. Cela permet de filtrer les points du fond ainsi que le bruit (points à forts écarts-types) alors que les points du texte (écarts-types plus faibles) sont renforcés.

Figure 3 - *Exemple d'un mot issu d'une vidéo, ainsi qu'une version zommée qui met en évidence les problèmes courants de qualité.*

Les résultats présentés dans le paragraphe suivant, montrent que l'amélioration ainsi réalisée, par intégration de 16 images de la vidéo d'origine, est tout à fait satisfaisante. Cette étape d'intégration est suivie d'une binarisation afin de pouvoir envoyer les résultats vers un logiciel d'OCR commercial. Cette binarisation est faite par seuillage adaptatif sur des portions de l'image.

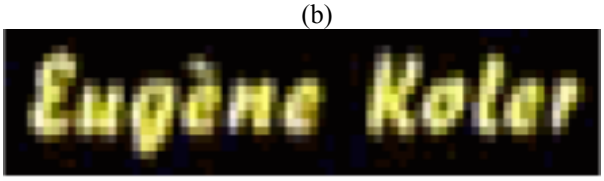
4 Quelques résultats

Nous présenterons maintenant quelques résultats obtenus sur des génériques vidéos de divers formats. Les génériques les plus faciles à traiter sont ceux qui comportent un fond uniforme. Un exemple de texte issu d'un tel générique est donné sur la figure 4. Dans ce cas, le texte est reconnu parfaitement part l'OCR utilisé (figure 4b), et après apprentissage de la police (figure 4d).
La figure 5 présente un exemple de résultat obtenu sur une vidéo à fond fixe. Là encore la reconnaissance ne pose pas de problème à l'OCR. Les résultats sont très similaires dans le cas de génériques à fond non-uniforme en mouvement. L'intégration permet d'améliorer sensiblement la qualité du texte. La situation la plus pénalisante serait un générique défilant sur un fond non-

uniforme qui aurait une translation identique à celle du générique. Nous n'avons pas rencontré cette situation dans nos expérimentations.

(a)

(b)



(c)

(d)

Figure 4 - *Exemple de mots issus d'une vidéo à fond uniforme. Le mot d'origine (a) et (c) et sa version après intégration et binarisation (b) et (d) .*

(a)

(b)

(c)

(d)

Figure 5 - *Exemple de mots issus d'une vidéo à fond fixe. Le mot d'origine (a) et sa version binarisée (b). Le mot après intégration (c) et binarisation (d) .*

5 Conclusion

L'approche proposée ici permet la reconnaissance automatique du texte des génériques défilants avec un taux de réussite assez important. Le principe même de la méthode oblige à se limiter aux mouvements de translation pure. Il n'est cependant pas nécessaire que cette translation soit rectiligne uniforme. Les résultats présentés ici se basent sur des génériques à scrolling vertical. La méthode s'applique aussi directement sur les génériques en ligne à scrolling horizontal : il faut juste adapter la forme des boites utilisées. En revanche, la prise en compte d'autres types de mouvements plus complexes (rotations zooms …) nécessiterait l'emploi de méthodes plus lourdes.

Les perspectives de ce travail concerne principalement l'extraction des textes de scène qui sont une source importante d'informations. Cependant, ce type de texte pose de nombreux problèmes, à la fois en ce qui concerne l'orientation du texte, le type de mouvement que l'on peut rencontrer, mais aussi le faible contraste possible puisque ce texte se présente tel qu'il est, sans les précautions et améliorations que l'on retrouve pour les textes artificiels.

Références

[1] H. Li, D. Doermann. Text Extraction and Recognition in Digital Video. Proc. Third IAPR Workshop on Document Analysis Systems. pp. 119–128. 1998.

[2] H. Li, D. Doermann. Automatic Text Detection and Tracking in Digital Video. IEEE Trans. on Image Processing, 9, pp. 147–156. 2000.

[3] D. Crandall, R. Kasturi. Robust detection of stylized text events in digital video. Proc. of the Sixth Int. Conference on Document Analysis and Recognition, Seattle, pp. 865-869. 2001.

[4] A.K. Jain, B. Yu. Automatic Text Location in Images and Video Frames. Pattern Recognition, 31(12), pp. 2055-2076. 1998.

[5] A. Wernicke, R. Lienhart. On the Segmentation of Text in Videos. IEEE Int. Conference on Multimedia and Expo (ICME2000), vol. 3, pp. 1511-1514. 2000.

[6] C. Wolf, JM. Jolion and F. Chassaing. Text Localization, Enhancement and Binarization in Multimedia Documents. In Proceedings of the International Conference on Pattern Recognition *(ICPR) 2002*, Quebec City, vol. 4, pp. 1037-1040, 2002.

[7] C. Schmid, R. Mohr. Local Grayvalue Invariants for Image Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5). pp. 530-535. 1997.

[8] E. Etiévent. Assistance à l'indexation vidéo par analyse du mouvement Th. Doct. INSA de Lyon. 2002. 150 pages.