Applications de la détection de texte pour l'indexation de vidéos de télévision

Xavier Naturel*
IRISA - INRIA Rennes
Campus de Beaulieu
Rennes - France
{xnaturel}@irisa.fr

Résumé

La détection et la reconnaissance de texte dans une vidéo sont souvent reconnus comme étant des outils importants pour aider l'indexation de documents vidéos. Toutefois, il existe assez peu de travaux donnant des exemples concrets d'utilisation. Dans cet article, nous passons en revue les différentes applications de la détection de texte qui ont étés proposées, puis deux nouvelles applications sont ensuite définies dans un contexte d'indexation de flux télévisés. La principale application étudiée consiste à détecter les génériques de fin des programmes de télévision, afin d'aider ou de compléter la structuration du flux.

Mots clefs

Détection de texte, vidéo, structuration, télévision.

1 Introduction

De nombreux travaux existent en détection de texte dans des images ou dans des vidéos. Tous ces travaux font l'hypothèse que le texte est d'une grande importance pour les méthodes d'analyse automatique de la vidéo. En particulier, la détection de texte est l'une des rares techniques sensée permettre d'obtenir un résultat sémantique, uniquement à partir de traitements bas niveau. Toutefois, la manière d'utiliser les résultats de la détection de texte n'est pas évidente, et assez peu de travaux sont menés sur l'exploitation de ces résultats.

Nous nous proposons de rappeler brièvement, en section 2.1, le principe de la détection de texte, puis d'identifier, dans la section 2.2, les diverses applications de la détection de texte pour l'indexation vidéo. La section 3 décrit rapidement la méthode employée pour la détection. Nous nous plaçons ensuite, pour la suite de l'article, dans un contexte de structuration par le contenu de flux télévisés [1]. Une première application, utilisant les résultats de la reconnaissance de texte, est donnée dans la section 4. Finalement, dans la section 5, une étude plus détaillée est faite sur la détection de génériques de programmes de télévision à partir des résultats de la détection de texte.

*Xavier Naturel est désormais à l'IDIAP, Centre du Parc, Av. des Prés-Beudin 20, 1920 Martigny, Suisse

2 Travaux antérieurs

Cette section présente de manière non exhaustive les principales méthodes de détection et de suivi de texte développées lors de travaux antérieurs. un panel d'applications possibles de la détection de texte est aussi donné.

2.1 Méthodes d'extraction de texte

Nous rappelons succintement le processus d'extraction de texte, en parcourant les principaux travaux dans ce domaine.

Pour extraire le texte présent dans une vidéo sous la forme de chaine de caractères, on distingue, en général, quatre étapes. Dans l'ordre de traitement : la détection, le suivi, l'amélioration et la reconnaissance.

La détection est l'étape la plus délicate, qui consiste à déterminer quelles sont les zones de l'image qui contiennent du texte. De nombreuses méthodes ont été développées à cet effet. Le texte peut être détecté comme étant une zone de fort gradient [2], ou par détection de contours [3]. Cette détection est associée à des contraintes morphologiques et géométriques spécifiques au texte. Des approches basées classification sont aussi proposées [4, 5]. Par exemple, Li *et al.* [4] utilisent des statistiques sur les coefficients d'ondelettes pour entrainer un réseau de neurones. Chen *et al.* [6] localisent le texte par des méthodes de morphologie mathématique, le normalisent, puis en extraient plusieurs caractéristiques (dérivées spatiales, DCT...) pour entrainer deux types de classifieurs (perceptron et SVM). Des méthodes dans le domaine compressé existent aussi [7, 8].

L'étape de suivi consiste à suivre les zones de texte au cours du temps, ce qui permet de pouvoir suivre les zones de texte mobiles, mais aussi de supprimer un grand nombre de fausses alarmes. Le suivi peut simplement consister en une étape d'intégration temporelle, afin de vérifier la stabilité de la détection du texte, et de filtrer ainsi les fausses alarmes [2]. Des approches permettant de suivre le texte en mouvement ont aussi été developpées, en définissant une mesure de similarité entre zones de texte, par block-matching [4], ou en définissant une signature à partir des profils horizontaux et verticaux [9]. Le suivi est alors effectué en maximisant la mesure de similarité entre deux zones de textes détectées. Une autre approche propose

d'estimer le mouvement de translation par le spectre croisé [10].

Avant l'étape de reconnaissance, il est en général nécessaire d'améliorer la qualité du texte détecté, pour que les OCR, non adaptés à la reconnaissance de textes bruités et de petite taille, donnent des résultats satisfaisants. L'amélioration consiste à utiliser une intégration temporelle pour améliorer la qualité et la taille du texte [10].

La reconnaissance consiste à transformer les images binarisées en chaîne de caractères, en utilisant un OCR, en général un logiciel commercial.

2.2 Applications de la détection de texte

Dans cette section, nous cherchons à identifier les diverses applications de la détection de texte.

McGee *at al.* [11] cherchent à identifier les inter-programmes à la télévision, et utilisent la présence du texte comme un indice supplémentaire à la présence de publicités et autres inter-programmes.

Dans des domaines spécifiques, comme les vidéos de sport, ou les journaux télévisés, il existe une connaissance a priori sur l'endroit où apparait le texte, et sa sémantique. Par exemple, Delakis *et al.* [12] utilisent le score affiché à l'écran dans des vidéos de tennis pour guider un processus de structuration. Une application originale est réalisée dans le projet FERIA[13]. Elle consiste à aligner les résultats de la détection des sous-titres d'un opéra avec son livret, ce qui permet de retrouver une scène à partir du livret.

Une expérience menée par Lienhart *et al.* [14] consiste à construire un système de recherche vidéo par mot-clé, en étiquetant chaque image par la sortie d'un OCR, auquel on a fourni les zones de texte binarisées. Les nombreuses erreurs de l'OCR et de la détection rendent intéressante l'utilisation de techniques de recherche approchées. Les auteurs montrent qu'il est alors possible, à partir d'un mot-clé requête, de retrouver les séquences vidéos dans lesquelles ce mot-clé apparait en tant que texte. Toutefois, les auteurs n'ont pas évalué la pertinence d'un tel système pour l'indexation : le texte à l'écran est il vraiment pertinent par rapport à la séquence ? Les résultats retournés par une requête textuelle sur le texte à l'écran ont ils un sens ? Ceci n'est pas évident.

Li *et al.* [15] développent un système plus réaliste. Ils construisent un corpus vidéo classé en quatre catégories : Sport, Jounal, météo, et publicité. Le titre de chaque catégorie est utilisé comme requête, le but étant alors, à partir d'un mot-clé requête caractérisant une catégorie, de retrouver toutes les séquences vidéos classées dans cette catégorie. Deux techniques sont utilisées : l'expansion de requête, et l'utilisation d'une technique de recherche approchée incluant un dictionnaire. De très bons résultats sont obtenus.

3 Extraction de texte et contexte applicatif

3.1 Méthode

Nous présentons rapidement la méthode utilisée pour détecter et reconnaitre le texte à l'écran, sans la détailler, la méthode est standard par rapport à l'état de l'art. La détection est réalisée à partir des travaux de Wolf [2] : accumulation du gradient, contraintes morphologiques et géométriques. Le suivi est réalisé en reprenant l'idée générale de Pitié *et al.* [16]. Le principe est de générer un ensemble de candidats, indépendamment pour chaque image, et de prendre une décision hors-ligne, en déterminant la séquence de candidats de probabilité a posteriori maximale, par l'algorithme de Viterbi.
La figure 1 donne un exemple de suivi de texte sur un générique de fin.
Pour reconnaitre le texte détecté, la binarisation des zones de texte est effectuée par l'algorithme d'Otsu[17], et les OCR libres gOCR et OCRAD sont utilisés. Une tentative d'amélioration des résultats a posteriori a été effectuée, d'après les travaux de Neuhoff [18], mais n'a pas donné d'amélioration satisfaisante des résultats, certainement à cause des très faibles taux de reconnaissance des OCR. Cette technique n'est en conséquence pas incluse dans notre méthode générale.

Figure 1 – *Exemple de suivi de texte sur un générique de série télévisée, sur 6 images consécutives.*

3.2 Contexte applicatif

Nous nous plaçons dorénavant dans un contexte de structuration par le contenu de vidéos de télévision [1]. Dans ce contexte, l'objectif est double : identifier les débuts et fins de programmes, et caractériser relativement grossièrement ces programmes, en général en les étiquetant par leur titre. Un certain nombre de travaux ont étés effectués sur cette problématique [1, 19, 20]. Il n'est bien entendu pas possible de réaliser la structuration à partir de la seule détection de texte, nous présentons ici simplement des

améliorations possibles des méthodes existantes grâce à la détection de texte.

4 Étiquetage par reconnaissance de texte

La première application concerne l'utilisation de la reconnaissance de texte. Cette dernière peut être très utile pour confirmer ou compléter les résultats de l'étiquetage des programmes. Rappelons que dans ce contexte, le flux télévisé est automatiquement segmenté en programmes/inter-programmes, et les programmes sont étiquetés par alignement avec le guide des programmes [19]. Deux intérêts principaux peuvent être dégagés.
Le premier est l'étiquetage de programmes. Il existe un intérêt à vouloir confirmer les résultats d'étiquetage automatique, afin de fournir un indice de confiance pour chaque étiquette. Les résultats de reconnaissance de texte peuvent donc être comparés, par une distance de Levenshtein, à l'étiquette fournie par la structuration. Un autre point de vue, est que, dans le cas où une étiquette ne peut être obtenue par la structuration, la reconnaissance de texte peut fournir des hypothèses. La mauvaise qualité de la reconnaissance impose toutefois de filtrer ces résultats, par exemple en les comparant avec une base de connaissance de titres de programmes.
Le deuxième intérêt est l'étiquetage des bandes-annonces, qui comportent toujours le titre du programme annoncé, afin que le télespectateur puisse identifier la bande-annonce en tant que tel. L'intérêt de l'étiquetage des bandes-annonces est expliqué plus en détail dans [21]. Nous donnons ici quelques exemples de texte reconnu sur trois bandes-annonces. Ces résultats sont limités mais montrent le potentiel de la reconnaissance de texte dans une application réelle.

Boulevard du palais Bande annonce de 41 secondes d'un téléfilm. Le texte « Boulevard du palais » est présent durant toute la bande annonce, et est correctement suivi. Le meilleur résultat fourni par la reconnaissance de texte est « Boulevard _ », mais la plupart des résultats sont de très faible qualité : « h,'_l », « _WA_ DU_ ».

Tout vu, tout lu Pré-annonce de 18 secondes d'un jeu. Le texte « Tout vu tout lu » n'est présent qu'en fin de bande annonce, sur peu d'images, et n'est pas détecté. En revanche, le texte « Dans un instant » est bien suivi mais les résultats de reconnaissance sont de très faible qualité : « dansq », « _n4 rm », ou encore « _ns un ins&_ _ ».

C'est au programme Bande-annonce de 33 secondes d'une émission de plateau. Le texte « C'est au programme » n'est présent qu'en tout début et fin de bande annonce, soit environ 4 secondes, mais le texte est correctement suivi. Les résultats sont corrects, avec, par exemple « _estau _ Drogramme », « c'est au _or_ogramme ».

Les faibles résultats montrent qu'il est difficile d'utiliser ces résultats directement, et qu'il faut donc une étape supplémentaire. Par exemple, il est envisageable de comparer ces résultats par une distance de Levenshtein avec les étiquettes du guide des programmes, afin de corriger les erreurs et d'identifier les résultats pertinents. Dans ce cas, les bandes annonces « Boulevard du palais » et « C'est au programme » seraient correctement étiquetées, avec des distances de Levenshtein normalisées de respectivement 0,53 et 0,2.
Ces résultats sont à interpréter comme étants préliminaires. Nous avons toutefois présenté un cadre où de bons résultats seraient directement utilisables.

5 Détection de génériques

Nous nous plaçons ici de la même façon dans un contexte applicatif de structuration de flux de télévision. L'idée proposée est d'utiliser le texte à l'écran pour détecter les génériques de fin des programmes. Ces derniers comportent souvent, en effet, une liste défilante des personnes impliquées dans le programme en question. La détection du générique de fin permettrait donc d'obtenir la borne de fin d'un programme avec plus de précision, ou permettre une segmentation, lorsque deux programmes ne sont pas séparés par un inter-programme (cas de certains programmes la nuit).
Le corpus utilisé est constitué de 12 heures continues de télévision française, et a été enregistré le 10 mai 2005 sur la chaîne France2. Le corpus est découpé en 12 segments (un par heure), sur lesquels on effectue la détection et le suivi de texte. Ceci produit, pour chaque image, un ensemble de boites englobantes. Trois caractéristiques sont calculées pour chaque image :
– Le nombre de boites
– L'aire totale que recouvrent les boites
– La longueur moyenne du suivi par boite
Sur ces 12 segments, 12 génériques sont présents, non uniformément répartis, 3 segments ne comportent aucun générique.
Deux types de classifieurs sont proposés, l'un utilisant un réseau de neurones de type perceptron multi-couches, le deuxième un modèle de Markov caché (HMM). Afin de comparer les résultats de ces deux classifieurs à une méthode de référence, une méthode heuristique très simple est aussi définie. Elle consiste à considérer qu'un générique se caractérise comme étant une séquence qui comporte un suivi de texte continu, d'une certaine durée minimale. La méthode réalise donc un simple seuillage de la caractéristique de durée de suivi. Une durée minimale de suivi de 30s a été déterminée empiriquement comme pertinente. Les informations sur l'aire et le nombre de boites ne sont pas utilisées.

5.1 Perceptron multi-couches (MLP)

La première méthode proposée est d'utiliser un réseau de neurones de type perceptron multi-couches pour l'appren-

tissage des caractéristiques.
Les caractéristiques définies en section 5 sont utilisées pour former un vecteur à 3 dimensions, qui sert d'entrée à un perceptron multi-couches. L'architecture du réseau a été déterminée par essais successifs, qui ont montré une stabilité des résultats. L'architecture retenue comprend deux couches cachées composée de 4 neurones, soit une architecture 3-4-4-1. La procédure d'optimisation qui permet la recherche des paramètres du réseau est de type gradient conjugué.
L'apprentissage est délicat car il existe très peu d'exemples positifs (1.5%), et beaucoup de ces exemples sont bruités. L'utilisation d'une technique classique de validation croisée sur les 12 segments définis n'a pas permis de réaliser un apprentissage correct. Seul l'apprentissage réalisé sur un segment comportant un seul générique avec des caractéristiques claires, a donné des résultats.
La sortie du réseau étant une valeur numérique, cette sortie est filtrée, en la seuillant à 0.7 (la fonction cible vaut 1 lorsque générique, -1 sinon), et ne ne conservant que les segments supérieurs à trente secondes, de la même manière que pour la méthode heuristique.

5.2 Modèle de Markov caché (HMM)

Dans le modèle précédent, les images sont classées indépendamment les unes des autres, sans prendre en compte l'aspect temporel. L'aspect temporel est toutefois présent grâce à la caractéristique de longueur du suivi, mais la dépendance temporelle n'est pas explicite. Afin de prendre en compte cet aspect, nous avons étudié la possibilité d'une classification par modèle de Markov caché.
Les observations sont continues. Nous avons choisi de les discrétiser afin de se ramener à un modèle à observations discrètes. Une quantification vectorielle est appliquée sur le signal vectoriel d'apprentissage. Il a été déterminé expérimentalement que 8 classes donnaient le meilleur résultat. Le modèle est de type ergodique, à deux états.
L'apprentissage est réalisé de la même manière que dans la section précédente, sur le même segment. Cet apprentissage est réalisé de façon classique par l'algorithme EM, en initialisant les valeurs de départ des matrices de transition et d'émissions par comptage sur le segment d'apprentissage.

5.3 Résultats

La méthode est évaluée sur les 11 segments de test, soit 11 heures de télévision. Afin d'évaluer les performance, nous définissons trois métriques : une métrique par image, une métrique par évènement, et un taux de classification global. La métrique image mesure le nombre d'images correctement classées en tant que générique, et la métrique par évènement mesure le nombre de génériques correctement détectés. Ces deux mesures sont exprimées en terme de précision et rappel. Le taux de classification mesure le pourcentage d'images bien classées (génériques et non-génériques). De ces trois métriques, c'est la métrique par évènement qui est la plus pertinente, puisqu'elle donne une information intuitive en terme de génériques correctement détectés.

	Heuristique		MLP		HMM	
Métrique	P	R	P	R	P	R
Image	15.1	55.9	12.6	27.1	7	51
Générique	10.5	40	10.3	60	1.4	60
Globale	94.3		95.8		83.8	

Tableau 1 – *Résultats des trois méthodes proposées*

Il est important de remarquer qu'en terme de classification globale, c'est à dire en générique/non-générique, les résultats paraissent très bons. En revanche, les résultats avec les deux autres métriques sont très mauvais. Ceci tient à la taille du corpus, et surtout au fait que les génériques sont des évènements rares, le nombre d'images correctement classées en tant que non-générique est très élevé, dóù un taux de classification élevé.
La modélisation par modèles de Markov cachés ne semble pas adaptée, avec des résultats bien inférieurs à la méthode de référence, et une précision extrêmement faible, indication que les résultats ont peu de sens. C'est la méthode par perceptron multicouche qui semble donner les meilleurs résultats, qui sont exploitables, bien que le nombre de fausses alarmes reste élevé. Il semble indispensable de réfléchir à un mécanisme permettant d'augmenter la précision.
De façon curieuse, les fausses alarmes ne proviennent pas des publicités. Un couplage des résultats précédents avec le résultat d'une structuration de flux [19], qui permet de filtrer les instants de d'inter-programmes, et donc de ne garder que les instants de programmes comportant des détections, ne donne que très peu voir pas du tout d'amélioration.

6 Conclusion

Il est proposé dans cet article d'utiliser les résultats de détection et de suivi de texte pour deux applications. Nous nous intéressons en particulier à la détection des génériques de fin de programmes de télévision. Ceci est utile dans une problématique de structuration de flux de télévision, où il est parfois difficile de détecter la fin d'un programme et le début d'un inter-programme.
Deux types de classifieurs sont étudiés, un perceptron multicouche, et un modèle de Markov caché. Les résultats montrent de très nombreuses fausses alarmes pour les deux méthodes. Il semble difficile de pouvoir généraliser les caractéristiques des génériques à partir des observations choisies. La diversité des génériques (défilant sur tout l'écran, sur une moitié d'écran, sur seulement un bandeau, générique statique...) explique sans doute cette difficulté. Les nombreuses fausses alarmes résiduelles de la détection de texte rendent aussi la tâche plus ardue.
Il semble donc que le problème soit relativement difficile et mérite un travail plus approfondi. Il est probable que l'uti-

lisation d'un détecteur de texte plus sophistiqué et des prétraitements sur les détections soient nécessaires afin d'obtenir des résultats corrects. Enfin, il peut être intéressant d'envisager un couplage plus étroit entre le résultat de la structuration et la détection de génériques.

Références

[1] Sid-Ahmed Berrani, Patrick Lechat, et Gaël Manson. Tv broadcast macro-segmentation : metadata-based vs. content-based approaches. Dans *CIVR '07 : Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 325–332, Amsterdam, The Netherlands, 2007. ACM Press.

[2] C. Wolf et J.M. Jolion. Extraction and recognition of artificial text in multimedia documents. Dans *Pattern Analysis and Applications*, 2003.

[3] L. Agnihotri et N. Dimitrova. Text detection for video analysis. Dans *CBAIVL '99*, page 109, Washington, DC, USA, 1999.

[4] H. Li, D. Doermann, et O. Kia. Automatic text detection and tracking in digital videos. Dans *IEEE Transactions on Image Processing*, volume 9, pages 147–156, January 2000.

[5] Qixiang Ye, Qingming Huang, Wen Gao, et Debin Zhao. Fast and robust text detection in images and video frames. *Image and Vision Computing*, 23(6) :565–576, March 2005.

[6] Datong Chen, Jean-Marc Odobez, et Jean-Philippe Thiran. A localization/verification scheme for finding text in images and video frames based on contrast independent features and machine learning methods. *Signal Processing : Image Communication*, 19(3) :205–217, Mars 2004.

[7] D. Crandall et R. Kasturi. Robust detection of stylized text events in digital video. Dans *ICDAR*, pages 865–, 2001.

[8] Julinda Gllavata, Ralph Ewerth, et Bernd Freisleben. Tracking text in mpeg videos. Dans *MULTIMEDIA '04 : Proceedings of the 12th annual ACM international conference on Multimedia*, pages 240–243, New York, NY, USA, 2004. ACM Press.

[9] R. Wernicke, A. ; Lienhart. On the segmentation of text in videos. Dans *ICME*, volume 3, pages 1511–1514, 2000.

[10] D. Marquis et S. Bres. Suivi et amélioration de textes issus de génériques vidéos. Dans *CORESA'03*, pages 179–182, 2003.

[11] T. McGee et N. Dimitrova. Parsing tv program structures for identification and removal of non-story segments. Dans *in SPIE Conf. on Storage and Retrieval for Image and Video Databases*, 1999.

[12] Manolis Delakis, G. Gravier, et P. Gros. Score oriented viterbi search in sport video structuring using hmm and segment models. Dans *MMSP'06*, Victoria, Canada, October 2006.

[13] Jean Carrive. Feria : Framework pour l'expérimentation et la réalisation industrielle d'applications multimédias. Dans *RIAM*, page 10, Rennes, France, 2004.

[14] Rainer Lienhart et Wolfgang Effelsberg. Automatic text segmentation and text recognition for video indexing. *Multimedia Syst.*, 8(1) :69–81, 2000.

[15] H. Li et D. Doermann. Video indexing and retrieval based on recognized text. Dans *IEEE Workshop on Multimedia Signal Processing*, pages 245–248, Dec 2002.

[16] F. Pitié, S.-A. Berrani, A. Kokaram, et R. Dahyot. Off-line multiple object tracking using candidate selection and the viterbi algorithm. *ICIP*, September 2005.

[17] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9 :62–66, Mars 1979.

[18] D.L. Neuhoff. The viterbi algorithm as an aid in text recognition. *IEEE Trans. Inform. Theory*, 21(2) :222–226, March 1975.

[19] Xavier Naturel, Guillaume Gravier, et P. Gros. Fast structuring of large television streams using program guides. Dans *AMR'06*, volume 4398 de *Lecture Notes in Computer Science*, pages 223–232, Geneva.

[20] Jean-Philippe Poli. *Structuration automatique de flux télévisuels*. Thèse de doctorat, Université Paul Cézanne, mai 2007.

[21] Xavier Naturel. *Structuration automatique de flux vidéos de télévision*. Thèse de doctorat, Université de Rennes 1, avril 2007.