Détection et extraction de texte de la vidéo[1] Christian Wolf

Jean-Michel Jolion

Laboratoire Reconnaissance de Formes et Vision, INSA de Lyon
Bât Jules Verne 20, Avenue Albert Einstein
Villeurbanne, 69621 cedex, France
{wolf,jolion}@rfv.insa-lyon.fr

Résumé

Les systèmes d'indexation ou de recherche par le contenu disponibles actuellement travaillent sans connaissance (systèmes pré-attentifs). Malheureusement les requêtes construites ne correspondent pas toujours aux résultats obtenus par un humain qui interprète le contenu du document. Le texte présent dans les vidéos représente une caractéristique à la fois riche en information et cependant simple, cela permet de compléter les requêtes classiques par des mots clefs.

Nous présentons dans cet article un projet visant à la détection et la reconnaissance du texte présent dans des images ou des séquences vidéo. Nous proposons un schéma de détection s'appuyant sur la mesure du gradient directionnel cumulé. Dans le cas des séquences vidéo, nous introduisons un processus de fiabilisation des détections et l'amélioration des textes détectés par un suivi et une intégration temporelle.

Mots Clef

Vidéo OCR, détection de texte, reconnaissance de texte.

1. Introduction

L'extraction du texte des flux de vidéo est un domaine de recherche très récent mais cependant très fécond. Les débuts de notre travail sont inspirés par des travaux dans le domaine du traitement de documents. La recherche du texte dans les journaux a résulté dans les premiers travaux sur la vidéo. Par contre, le domaine de la vidéo rencontre de nouveaux problèmes dus aux données très différentes (pour une description détaillée du problème, consulter [8]). Les algorithmes évolués de traitement de texte ont donc été remplacés par des algorithmes spécialement conçus pour les données de la vidéo [2, 6, 11].

La plupart des travaux actuels traitent le problème de la détection et de la localisation du texte (nommé "détection"). En revanche, il existe très peu de recherche sur la reconnaissance. Par ailleurs, même si la plupart des problèmes sont identifiés, tout n'est pas encore résolu. Cela concerne, entre autres, l'amélioration du contenu. Bien que ce sujet soit abordé par plusieurs auteurs (e.g. par Li et Doermann[2]), les résultats ne sont pas encore tout à fait satisfaisants. La motivation de notre système étant l'indexation de la vidéo, nous avons concentré notre travail sur le texte statique horizontal (sous-titrage, infos, etc.) mais nous

1. Cette étude est soutenue par France Télécom Recherche et Développement dans le cadre du projet ECAV. La première partie de notre étude a donné lieu à une brevet [10].

avons abordé non seulement la détection det la localisation mais aussi la reconnaissance afin d'aboutir à un système complet.

Dans le chapitre 2, nous décrivons les détails de notre système de détection, suivi et binarisation. Le chapitre 3 montre les expériences poursuivies et les résultats obtenus/. Le chapitre 4 donne la conclusion et les perspectives de ce travail.

2. Notre système d'extraction

Le but principal d'un système d'extraction de texte est le suivant : 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.

La détection du texte est réalisée dans chaque frame de la vidéo [2]. Les rectangles figurant la localisation du texte sont suivis pendant leur période d'apparition pour associer 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 pré-requis 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.

2.1. *Détection*

Notre approche s'appuie sur le fait que les caractères du texte forment une texture régulière contenant des contours verticaux allongés horizontalement. L'algorithme de détection reprend la méthode de LeBourgeois [1], qui utilise l'accumulation des gradients horizontaux pour détecter le texte :

$$A(x,y) = \left[\sum_{i=-t}^{t} \left(\frac{\partial I}{\partial x}(x+t,y) \right)^2 \right]^{\frac{1}{2}}$$

Les paramètres de ce filtre sont l'implémentation de l'opération dérivatif (nous avons choisi la version horizontale du filtre de Sobel, qui a obtenu les meilleurs résultats) et la taille de la fenêtre de l'accumulation, qui correspond à la taille des caractères. Comme les résultats ne dépendent pas trop de ce paramètre, nous l'avons fixé à $2t + 1 = 13$ pixels. La réponse est une image contenant une mesure de la probabilité de chaque pixel d'être un pixel de texte.

2. Un sous échantillonnage de la vidéo peut accroître la rapidité du système mais cela réduit la fiabilité du suivi.

La binarisation des gradients cumulés est poursuivie par une version deux-seuils de la méthode d'Otsu [4]. Cette méthode calcule un seuil global à partir de l'histogramme de niveau de gris. Nous avons changé la décision de binarisation pour chaque pixel comme suit :

$$\begin{array}{lll} I_{x,y} < k_{bas} & \Rightarrow & B_{x,y} = 0 \\ I_{x,y} > k_{haut} & \Rightarrow & B_{x,y} = 255 \\ k_{bas} > I_{x,y} > k_{haut} & \Rightarrow & B_{x,y} = \end{array}$$

$$\begin{cases} 255 & \text{s'il existe un chemin} \\ & \text{à un pixel } I_{u,v} > k_{haut} \\ 0 & \text{sinon} \end{cases}$$

où le seuil k_{haut} correspond au seuil calculé par la méthode d'Otsu, et le seuil k_{bas} est calculé à partir du seuil k_{haut} et le premier mode m_0 de l'histogramme : $k_{bas} = m_0 + 0.87 \cdot (k_{haut} - m_0)$.

Avant de passer des pixels à des zones compactes, nous utilisons un post-traitement morphologique afin d'extraire les rectangles englobant 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

La phase morphologique est composée des étapes suivantes :

– fermeture (1 itération) ;

– suppression des ponts (suppression de tous les pixels qui font partie d'une colonne de hauteur inférieur à 2 pixels) ;

– dilatation conditionnelle (16 itérations) suivie par une érosion conditionnelle (16 itérations) ;

– érosion horizontale (12 itérations) ;

– dilatation horizontale (6 itérations).

Nous avons conçu un algorithme de dilatation conditionnelle pour connecter les caractères d'un mot ou d'une phrase, qui forment différentes composantes connexes dans l'image binarisée. Cela s'avére nécessaire si la taille de la police ou les distances entre les caractères sont relativement grandes. L'algorithme est basé sur une dilatation "standard" avec l'élément structurant $B_H = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$ et des conditions suivantes vérifiées pour chaque pixel : Un pixel P est dilaté si et seulement, si

– la différence de hauteur de la composante C_a incluant le pixel P et la composante C_b voisine à droite ne dépasse pas un seuil t_1.

– la différence de positions verticale de ces deux composantes ne dépasse pas un seuil t_2.

– la hauteur de la boite englobante incluant ces deux composantes ne dépasse pas un seuil t_3.

Les pixels dilatés sont marqués. L'érosion conditionnelle suivant la dilatation conditionnelle utilise le même élément structurant B_H. Les pixels supprimés par cette opération sont restreint aux pixels marqués préablement. Après les deux opérations tous les pixels encore marqués sont classés comme texte.

La figure 1 montre les résultats pendant la détection. Les figures 1a - 1c affichent l'image originale, les gradients et l'image des gradients accumulés. Les régions de texte sont visiblement marquées en blanc. La figure 1d montre l'image binarisée, portant encore du bruit, qui est supprimé dans la figure 1e. Les rectangles détectes sont superposés à l'image initiale sur la figure 1f.

Après le traitement morphologique, chaque région (composante connexe) est vérifiée en imposant des contraintes sur sa géométrie de façon à encore réduire le nombre des fausses alarmes. Enfin, une recherche des cas particuliers est appliquée pour considérer aussi les traits détachés de la zone de chaque composante (les hauts de casse et les bas de casse).

2.2. *Suivi*

Le but de ce module est le suivi et l'association des rectangles se correspondant afin de produire des apparitions de texte pendant plusieurs frames de la vidéo. Pour ceci, nous utilisons des mesures du recouvrement calculées entre chaque rectangle du frame et chaque rectangle de la liste des apparitions. 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.à.d le nombre de frames où elle apparaît, nous sert comme mesure de stabilité du texte. Nous considérons les apparitions de longueur plus courte qu'un seuil fixe comme fausses alarmes.

2.3. *Amélioration du contenu et binarisation*

Pour qu'un OCR soit en mesure d'extraire le texte contenu dans un rectangle, il est nécessaire d'améliorer la qualité de l'image. Nous utilisons les contenus de tous les frames F_i d'une apparition de texte pour produire une seule image améliorée. Cela est fait d'une façon robuste,

Figure 1: Les résultats intermédiaires pendant la détection : l'image d'entrée (a), le gradient (b), le gradient accumulé (c), l'image binarisée (d), l'image après le post-traitement (e), le résultat final (f).

basée sur des statistiques calculées sur le niveau de gris de chaque pixel pendant le temps d'apparition.

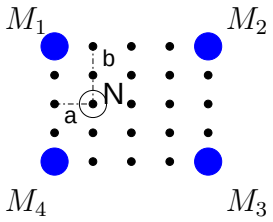


Figure 2: L'interpolation bi-linéaire.

De plus, le processus d'amélioration comprend une phase d'augmentation de résolution. Cela n'ajoute pas d'information, mais adapte l'image au format nécessaire pour un traitement avec un logiciel commercial de reconnaissance. Nous avons choisi l'interpolation bi-linéaire, qui calcule le niveau de gris d'un pixel comme moyenne des niveaux de gris de ses voisins. Le poids de chaque voisin correspond à la distance au pixel calculé : $(1-a)\cdot(1-b)$ (voir figure 2). Cette augmentation est appliquée à chaque frame de l'apparition, l'image finale étant la moyenne de tous les frames.

L'amélioration du contenu est accomplie par un poids additionnel ajouté au schéma d'interpolation agrandissant chaque frame. Le but de cette amélioration est une meilleur robustesse vis-à-vis du bruit en diminuant le poids pour les pixels voisins aberrants en regard de la moyenne des niveaux de gris. Le facteur g_k^i correspondant au frame F_i et au voisin M_k est calculé comme suit :

$$g_k^i = \frac{1}{1 + \frac{|F_i(M_k) - M(M_k)|}{1 + S(M_k)}}$$

où M est l'image moyenne et S est l'image d'écart type de l'appartion. Le poids final pour le voisin M_k est $(1-a)\cdot(1-b)\cdot g_k^i$.

L'image améliorée doit être binarisée avant le traitement par un OCR afin de supprimer le bruit de fond. Nous avons choisi la version améliorée par Sauvola et al. [5] de la méthode de Niblack [3]. Cette approche calcule une surface de seuillage en glissant une fenêtre sur l'image. Pour chaque pixel un seuil T est calculé à partir de quelques statistiques sur les niveaux de gris dans la fenêtre :

$$T = m \cdot (1 - k \cdot (1 - \frac{s}{R}))$$

où m est la moyenne, s est l'écart type, k est un paramètre fixé à $k = 0.5$ et R est la dynamique de l'écart type, fixé à $R = 128$. Cet algorithme a prouvé sa capacité à binariser des documents numérisés. Par contre, face aux documents multimédias caractérisés par des propriétés différentes (faible contraste, écart de niveaux de gris etc.), les résultats sont insuffisants. Nous avons changé le calcul du seuil T pour améliorer l'efficacité face aux vidéos :

$$T = m - k\,\alpha\,(m - M) \quad , \quad \alpha = 1 - \frac{s}{R} \quad , \quad R = max(s)$$

où M est le minimum des niveaux de gris de toute l'image et la dynamique d'écart type R est fixée au maximum de l'écart type s de toutes les fenêtres. Plus de détails sur cet algorithme sont proposès dans [9].

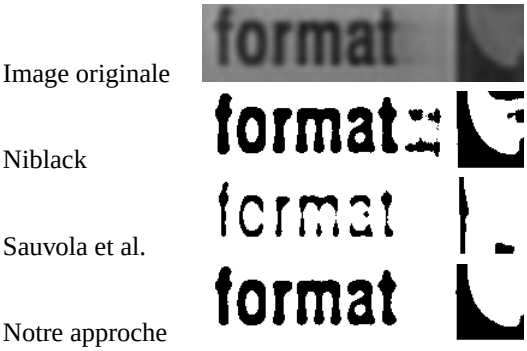


Figure 3: Différentes méthodes de binarisation.

La figure 2.3 montre une image exemple et les résultats obtenus par différentes méthodes. La méthode de Niblack segmente bien les caractères de texte, par contre du bruit est crée dans les zones sans texte. Les résultats obtenus avec la méthode de Sauvola et al. montrent moins de bruit dus aux hypothèses sur les données. Par contre, ces hypothèses ne sont pas toujours justifiées dans le cas de données vidéo, provoquant des trous dans les caractères. Notre méthode résout ce problème en gardant les bonnes performances au niveau de bruit.

3. Résultats

Pour évaluer notre système nous avons utilisé 60.000 frames de 4 vidéos différentes du corpus de l'INA[3](voir figure 4) en format MPEG 1 (384×288 pixels). Ces vidéos contiennent 371 occurrences de texte avec 3519 caractères. Pour la reconnaissance nous avons utilisé le produit commercial Abbyy Finereader 5.0.

3. L'institut national de l'audiovisuel, voir http ://www.ina.fr

Les résultats de détection sont regroupés dans la table 1a. Nous obtenons un taux de détection de 93.5% de rectangles. La précision de la détection est située à 34.4%, c.à.d. qu'il y a malheureusement un grand nombre de fausses alarmes. Cela est due au fait qu'il y a beaucoup de structures avec des caractéristiques similaires à celles du texte. Ce nombre de fausses alarmes peut être diminué en changeant les paramètres du système, bien sûr en entraînant aussi une baisse du rappel. Les nombreux paramètres permettent une grande flexibilité du système et une bonne adaptation aux caractéristiques intrinsèques des vidéos auxquelles ils sont directement liés.

Les résultats de la reconnaissance sont affichés dans la table 1b (les rectangles extraits du fichier #3 - la chaîne franco-allemande "Arte" - étaient séparés dans 2 groupes afin de pouvoir utiliser deux dictionnaires différents pour la reconnaissance). Nous avons comparé notre méthode de binarisation avec les méthodes de Niblack, Sauvola, et la méthode globale introduit par Otsu [4], qui utilise l'analyse discriminante pour déterminer un seuil global à partir de l'histogramme de l'image. Les résultats d'OCR confirment les meilleures performances de notre méthode de binarisation dans le cas de la vidéo. 85.4% de caractères détectés étaient reconnus correctement par l'OCR.

4. 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. Les taux de détection sur le texte horizontal sont prometteurs.

La continuité de ce travail est consacrée au texte ayant des propriétés moins spéciales (texte animé, texte avec des orientations générales etc.) et sur la phase de reconnaissance. Nous travaillons également sur la binarisation de boîtes de texte en utilisant des connaissances a priori [7].

5. Bibliographie

F. LeBourgeois. Robust Multifont OCR System from Gray Level Images. In *Proceedings of the 4th Int. Conference on Document Analysis and Recognition*, pages 1–5, Août 1997.

H. Li and D. Doerman. A Video Text Detection System based on Automated Training. In IEEE Computer Society, editor, *Proceedings of the ICPR 2000*, pages 223–226, 3 Septembre 2000.

W. Niblack. *An Introduction to Digital Image Processing*, pages 115–116. Englewood Cliffs, N.J. : Prentice Hall, 1986.

N.Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1) :62–66, 1979.

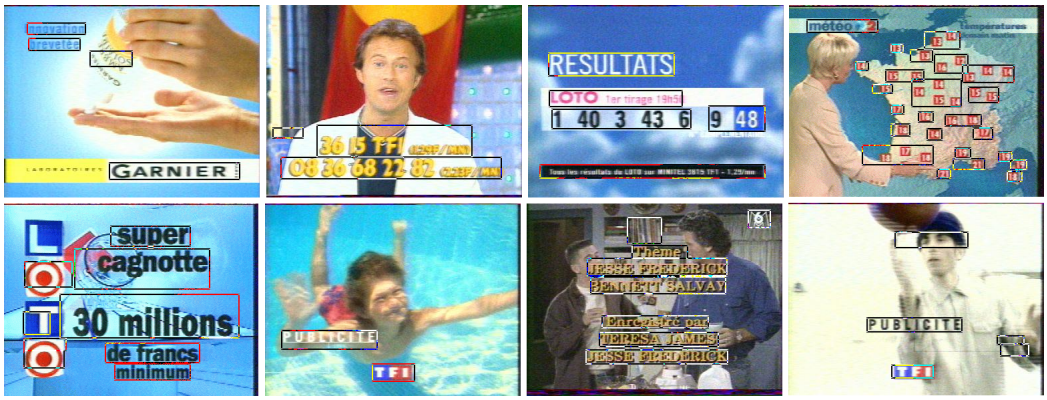
4. T=Texte, NT=Non-Texte, VT=Vérité terrain, FA=Fausses alarmes

(a) vidéo #1 : Publicité (France 3, TF1)

(b) vidéo #2 : Infos (M6, Canal+)

(c) vidéo #3 : Dessins animés et infos (Arte)

(d) vidéo #4 : Infos (France 3)



(e)

Figure 4: Exemples de la base de vidéo utilisée (a-d) quelques résultats de la détection statique, c.à.d. sans prendre en compte l'information de plusieurs frames (e).

Catégorie[4]	Fichier Vidéo #1	#2	#3	#4	Total	Précision, Rappel
Class. T	102	80	59	60	301	**93.5%**
Class. NT	7	5	4	5	21	
Total VT	109	85	63	65	322	
Positives	114	78	72	86	350	
FA	138	185	374	250	947	
Logos	12	0	34	29	75	
Texte de scène	22	5	28	17	72	
Total - FA	148	83	134	132	497	**34.4%**
Total det.	286	268	508	382	1444	

(a)

Entrée	méthode bin.	Rappel	Précision	Coût
#1	Otsu	39.6%	87.5%	791.7
	Niblack	79.0%	78.3%	528.4
	Sauvola	66.5%	75.8%	625.8
	Notre méthode	**81.5%**	**88.3%**	**361.1**
#2	Otsu	54.8%	**93.2%**	371.6
	Niblack	92.4%	79.6%	257.2
	Sauvola	82.8%	88.5%	203.8
	Notre méthode	**94.8%**	91.5%	**116.9**
#3-fr	Otsu	51.1%	**96.1%**	300.1
	Niblack	85.1%	93.4%	129.8
	Sauvola	61.5%	68.2%	367.2
	Notre méthode	**88.1**%	92.9%	**102.9**
#3-all	Otsu	57.5%	**98.2%**	121.8
	Niblack	**99.0%**	93.4%	25.4
	Sauvola	80.3%	98.1%	73.7
	Notre méthode	98.9%	97.4%	**11.2**
#4	Otsu	45.7%	84.8%	500.9
	Niblack	62.8%	70.5%	527.9
	Sauvola	76.0%	85.5%	281.4
	Notre méthode	**76.2%**	**89.3%**	**252.7**
Total	Otsu	47.3%	90.5%	2086.1
	Niblack	80.5%	80.4%	1468.7
	Sauvola	72.4%	81.2%	1551.9
	Notre méthode	**85.4%**	**90.7%**	**844.8**

(b)

Tableau 1: Les résultats de détection (a) Les résultats d'OCR (b).

J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen. Adaptive Document Binarization. In *International Conference on Document Analysis and Recognition*, volume 1, pages 147–152, 1997.

A. Wernike and R. Lienhart. On the segmentation of text in videos. In *Proc. of the IEEE Int. Conference on Multimedia and Expo (ICME) 2000*, pages 1511–1514, Juillet 2000.

C. Wolf and D. Doermann. Binarization of low quality text using a markov random field model. In IEEE Computer Society, editor, *Proceedings of the ICPR 2002*, volume 3, pages 160–163, Août 2002.

C. Wolf and J.-M. Jolion. Détection et Extraction du texte de la vidéo. In *ORASIS 2001, Congrès francophone de vision, Cahors, France*, pages 415–424, 5-8 Juin 2001.

C. Wolf and J.M. Jolion. Extraction de texte dans des vidéos : le cas de la binarisation. In *13ème congrès francophone de reconnaissance des formes et intelligence artificielle*, volume 1, pages 145–152, Janvier 2002.

C. Wolf, J.M. Jolion, and F. Chassaing. Procédé de détection de zones de texte dans une image vidéo. Patent France Télécom, Ref.No. FR 01 06776, Juin 2001.

Y. Zhong, H. Zhang, and A.K. Jain. Automatic Caption Localization in Compressed Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4) :385–392, Avril 2000.