Détection et extraction automatique de texte en vidéo:une approche par morphologie mathématique

Sophie Schüpp, Youssef Chahir, Abderrahim Elmoataz

GREYC UMR CNRS 6072, 6 Bd Maréchal JUIN, 14050 Caen, France

Résumé

Dans cet article, nous présentons une méthode générique de détection et de segmentation automatique de texte dans des images fixes ou issues de séquences vidéo, afin d'aboutir à une
meilleure automatisation de la phase d'indexation des scènes audiovisuelles. Cette méthode est fondée sur une approche combinée basée sur les opérations de la morphologie mathématique et des critères de cohérence spatiale tels que l'élongation, la rectangularité et la périodicité pour bien identifier le texte. L'algorithme de détection que nous proposons possède la particularité de n'être dépendent d'aucun seuil ni du contraste de l'image. Ce travail s'inscrit dans le cadre de l'indexation par le contenu des images animées où le besoin de séparer une scène en objets sémantiques distincts, est de plus en plus nécessaire.

Abstract

Text is very powerful index in content-based image and video indexing. Text extraction is also one of key tasks in document image analysis. Many methods proposed for this purpuse are based on classification or contour detection. In this paper we propose a morphological approach for text detection in video images. This approach can be applied to color image with complicated backround.

1. Introduction

Ces dernières années, avec les avancements réalisés en technologie numérique, il y a eu une révolution dans le domaine des communications visuelles. Mais jusqu'à présent le manque de progrès en recherche numérique l'a fortement freinée. La vidéo en est un sujet d'importance qui jusqu'à présent continue à être manipulé directement comme un objet de base (non décomposable) dans les documents multimédias. Son contenu reste rarement explicité et il est souvent très difficile de le classer ou d'en extraire des connaissances. Dans de nombreuses applications, telles que l'indexation et la recherche par le contenu, on demande à accéder à la structure interne de la vidéo, et à disposer ou à manipuler des informations de granularité plus fine, tels que le texte ou les objets visuels. Ainsi, l'utilisateur pourra faire des interprétations ou de la synchronisation avec d'autres objets de base comme des images ou des sons, dans différentes domaines d'applications tels que le filtrage dans un flot de données audiovisuelles, les conversions de média (texte vers parole, etc.) ou la détection et l'interprétation de messages incrustés (surveillance, caméra intelligente, etc.).

Habituellement la classification et l'annotation se font manuellement selon une liste de mots clefs choisie par un l'utilisateur. Cette technique est fastidieuse et l'automatisation du processus d'indexation présente un grand 'intérêt. L'extraction d'informations pertinentes telles que le texte qui peut en effet nous fournir des informations supplémentaires sur le contenu sémantique dans ces vidéos. Néanmoins, la détection et la reconnaissance de textes sont confrontées à plusieurs problèmes. S'il est souvent relativement bien contrasté par rapport à son environnement, le texte peut se retrouver superposé à un fond non homogène et complexe. De plus le texte peut lui-même être multicolore et inhomogène. Ces deux caractéristiques rendent son extraction difficile. Enfin, le texte lui-même peut être divisé selon deux classes : le texte graphique ou le texte scénique.

Le texte scénique apparaît dans une scène filmée. Il fait partie intégrale de l'image et peut être considéré comme un élément du monde réel. Des éléments comme un panneau de signalisation, un panneau publicitaire, une plaque d'immatriculation sont des considérés comme du texte scénique. Généralement ce type de texte est peu informatif sauf lorsqu'il s'agit d'identifier des personnes ou des véhicules.

Le texte graphique, quant à lui, est un élément qui est manuellement ajouté pour accompagner le support audiovisuel. Ainsi il est souvent plus structuré et étroitement relié avec le sujet global de la séquence. Des exemples de textes graphiques sont les titres des journaux télévisés, les indications de temps et de lieu, le

nom des personnes, etc.… Ces descripteurs sont souvent simples et prédéfinis : un style simple, généralement unicolore et placés dans l'intention d'être vu par le spectateur.

Dans cet article, nous proposons une méthode de détection et de segmentation automatique d'un texte dans des images avec des fonds complexes, afin d'aboutir à une meilleure automatisation de la phase d'indexation des images et de la vidéo. La méthode de détection est basée sur une approche morphologique qui repose essentiellement sur la reconstruction morphologique continue et sur des fermetures morphologique directionnelle dans une image de la séquence vidéo. Par la suite, par la suite pour localiser et reconnaître les zones susceptibles de contenir du texte on utilise des critères de cohérence spatiale tels que la proximité et la direction et des critères géométriques tels que l'élongation, la rectangularité et la périodicité, pour bien identifier la zone du texte à l'intérieur de l'image et débuter ensuite L'algorithme de détection possède la particularité de n'être dépendent d'aucun seuil ni du contraste de l'image.

Ce travail s'inscrit dans le cadre de l'indexation par le contenu des images animées où le besoin de séparer une scène en objets sémantiques distincts, tels que le texte, est de plus en plus nécessaire. Ce besoin est confirmé par le travail en cours du standard de compression vidéo MPEG 4 (MPEG-4 Industry Forum: http://www.m4if.org/), qui décompose d'abord la scène en ses différents objets sémantiques et code ceux-ci de manière indépendante.

2. Morphologie mathématique

La morphologie mathématique est habituellement décrite en termes de la théorie des ensembles [1] [2] [3]. Des travaux proposent de formuler ce domaine sous la forme d'une analyse multi-échelle hyperbolique [4] à l'aide d'équations aux dérivées partielles[5] [6]. Les opérateurs de base sont la dilatation et l'érosion. Ainsi la dilatation et l'érosion multi-échelle d'une image 2D $I(x, y)$ par un élément structurant plat B s'écrivent :

$$\delta(x, y, t = (I \oplus tB)(x, y) = \sup_{(a,b)\in tB}(I(x-a, y-b) \text{ et}$$

$$\varepsilon(x, y, t = (I \oplus tB)(x, y) = \inf_{(a,b)\in tB}(I(x-a, y-b) \ ,$$

où $tB = \{(ta, tb): \quad (a, b) \in tB\}$.

Si l'on prend comme élément structurant B un disque unitaire, alors ces deux opérations sont décrites par une équation aux dérivées partielles de la forme :

$$\begin{cases} \dfrac{\partial U}{\partial t} = \pm|\nabla U| \\ U(x, y, 0) = I(x, y) \end{cases}$$

Pour un élément structurant linéaire B_θ d'orientation θ la dilatation et l'érosion directionnelles sont définies par :

$$\frac{\partial U}{\partial t} = \left|\vec{N}\right|.|\nabla U|.|\cos\alpha| = |\nabla U|.|\cos\alpha| \quad \text{avec}$$

$$\alpha = \arctan\left(\frac{U_x}{U_y}\right) - \theta .$$

Ces opérations de base permettent de définir d'autres transformations intéressantes telles que les ouvertures, les fermetures ou les filtres connexes basés sur la reconstruction et leurs résidus.

La reconstruction discrète par dilatation et par érosion

Nous allons rappeler les définitions classiques de la reconstruction avant de présenter les formulations continues. Considérons une image initiale $I(x,y)$ et une image de référence ou un marqueur $g(x,y)$.

Avec la condition $(I > g)$, la reconstruction discrète par dilatation de l'image g dans I, est définie la manière suivante,:

$$\gamma^{(rec)} = \lim_{n\to\infty} \delta^n(I, g)$$

où $\begin{cases} \delta^0(I, g) = g \\ \delta^{n+1}(I, g) = Min(\delta_B(\delta^n(I, g), I) \end{cases}$

L'opérateur δ_B est la dilatation par un élément structurant B de taille 1.

De la même manière, Avec la condition $(I < g)$, la reconstruction discrète par érosion de l'image g dans I, est définie par l'équation suivante :

$$\varphi^{(rec)} = \lim_{n\to\infty} \varepsilon^n(I, g)$$

$$\begin{cases} \varepsilon^0(I, g) = g \\ \varepsilon^{n+1}(I, g) = Max(\varepsilon_B(\varepsilon^n(I, g), I) \end{cases}$$

L'opérateur ε_B est l'érosion par un élément structurant B de taille 1.

La reconstruction continue par dilatation et par érosion

Si $U(x,y,t)$ représente l'évolution de l'image référence g avec $U(x,y,t=0) = g(x,y)$, U est une solution «faible» de l'équation suivante :

$$\begin{cases} \dfrac{\partial U}{\partial t} = sign(U - I).|\nabla U| \\ U(.,t=0) = g \end{cases} \text{ avec } sign(r) = \begin{cases} -1 & si \quad r > 0 \\ +1 & si \quad r < 0 \\ 0 & si \quad r = 0 \end{cases}$$

Si $I \le g$, la fonction $U - I$ reste négative, l'équation définit alors l'érosion conditionnelle par rapport à l'image initiale I, la reconstruction continue par érosion est alors donnée par $U_\infty = \lim_{t\to\infty} U(x,y,t)$. De même si $I \ge g$, la fonction $U - I$ reste positive, l'équation exprime la dilatation conditionnelle par rapport à l'image initiale I, et la reconstruction continue par dilatation est donnée par $U_\infty = \lim_{t\to\infty} U(x,y,t)$.

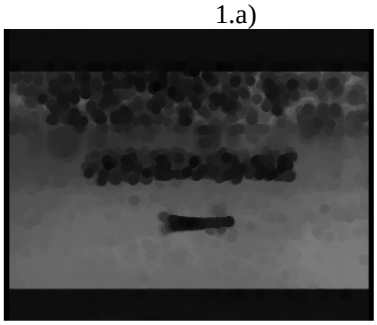
Si l'image $I - g$ peut changer de signe en fonction de la position des points dans l'image, $U - I$ peut changer de signe également, on aboutit alors à des filtres

analogues aux filtres de choc. L'opérateur érode dans les zones de l'image où le signe est négatif et dilate dans les zones où le signe est positif.
Les filtres de nivellement basés sur la reconstruction, permettent une simplification de l'image tout en préservant les contours. Ils favorisent également la création de zones plates de l'image. Ils permettent d'introduire de l'information a priori, à travers un terme pondérant dépendant de l'image. Ils constituent un puissant outil de prétraitement. Ils sont à la base d'un certain nombre de transformations morphologiques intéressantes comme les analyses résiduelles morphologiques, les minima de hauteur h, les filtres de contrastes que nous allons utiliser pour la détection du texte.

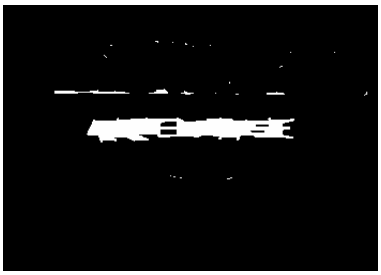
3. Détection et caractérisation du texte

Le problème de la détection de textes dans les images attire de plus en plus l'attention des chercheurs et reste un domaine en plein essor. En effet, avec l'avènement de la télévision interactive, en fonction des textes détectés des journaux télévisés, on pourra réaliser une association automatique avec les sujets d'actualités, et réaliser l'alignement du script textuel par rapport à la vidéo. Dans la littérature, différentes approches [7] [8] [9] traitent ce problème. Parmi ces travaux, on peut citer Jain et al qui ont proposé une méthode basée sur l'uniformité de la couleur pour la détection du texte, et Lienhart et al. dont la méthode est basée sur la recherche de contrastes locaux.
Notre approche combine les opérations de la morphologie mathématique et les critères de cohérence spatiale pour détecter et extraire le texte.
Les images sur lesquelles nous travaillons correspondent au format télévision aux dimensions 320 par 240. Ces images contiennent une foule d'informations dont seulement une petite partie, qui est la zone des textes, nous intéresse. Nous utilisons les opérations de morphologie mathématique pour éliminer les informations inutiles et mettre en valeur les zones susceptibles de contenir du texte.
Pour cela, la phase de détection commence par une érosion continue(I_{erod}) de l'image d'intensité I (**Fig.1.a**) qui va faire disparaître le texte, suivie d'une reconstruction de I_{erod} dans I (**Fig.1.b**). Le résultat est une image I_{recons} contenant uniquement les objets brillants d'une certaine taille.
Le résidu morphologique par reconstruction I_{res} (**Fig.1.c**) qui est la différence entre l'image d'intensité et l'image reconstruite permet de mettre en valeur les objets contrastés. On note qu'à cette étape, le texte est toujours présent et que le fond contient beaucoup mois d'informations que dans l'image d'origine.
Pour mener cette phase de détection de la zone de texte dans tous les cas même quand le texte est plus foncé que son environnement, l'image résidu est inversée puis érodée d'une façon continue avant de la reconstruire. Puis, l'image résultat I_R est une opération « Maximum » entre l'image résidu et son inverse pour ne garder que les informations susceptibles d'être du texte. Il ne faut pas perdre de vue que les images peuvent être de mauvaise qualité ou bruitées et les textes peuvent être placés sur des fonds non uniformes. L'image maximisée est binarisée selon un seuil déterminé automatiquement à partir de l'histogramme cumulé.
Pour détecter la zone de texte, nous avons utilisé la cohérence spatiale du texte pour faire en sorte q'une suite de mot soit considérée que comme un seul et même bloc. Pour cela, nous appliquons une fermeture morphologique directionnelle selon l'orientation du texte. Généralement, le texte a une orientation horizontale, l'élément structurant est horizontale (**Fig.1.d**).
Le calcul des rectangles exinscrits et leur labellisation permet de différencier les zones contenant du texte de celle n'en contenant pas (**Fig.1.e**). Un texte reconnaissable dans une image est généralement composé de plusieurs lettres qui possèdent des caractéristiques différentiables telles que la taille, orientation, contraste, proximité, et d'autres critères orientés objets. Tous les éléments connexes dont la taille est inférieure à un certain nombre de pixels (25 pixels) ne sont pas considérés comme étant du texte et sont éliminés.

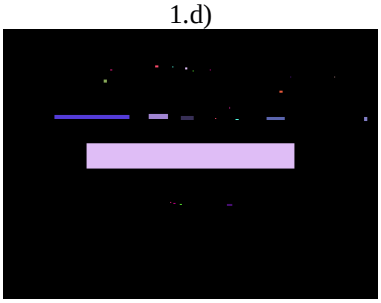
1.a)


1.b)

1.c)



1.d)

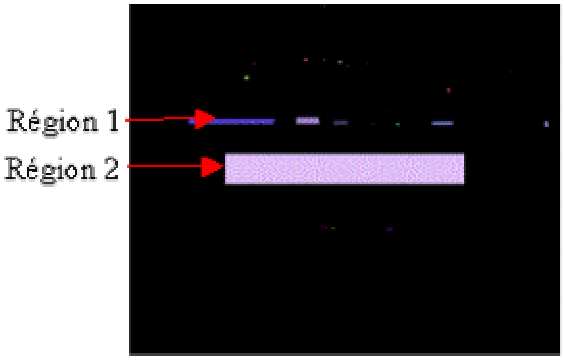


1.e)

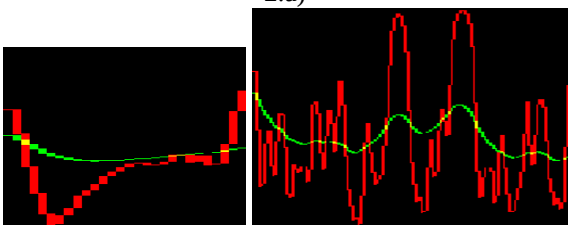
Figure .1. Différentes étapes de la détection de zones de texte. 1.a) Image initiale. 1.b) Image érodée. 1.c) Image reconstruite. 1.d) Binarisation et fermeture morphologique directionnelle. 1.e) rectangles exinscrits et Labellisation des régions.

Pour trier les zones susceptibles de contenir du texte par ordre d'importance, nous avons utilisé une série de critères pour caractériser les zones qui sont la rectangularité, l'allongement, la taille, la périodicité du texte et le profil.

1. *La rectangularité* : Ce critère n'est pas souvent caractéristique.
2. *L'allongement. :* Un mot ou une suite de mots possède généralement un rapport hauteur/largeur inférieur à 1. Cette technique fonctionne en fait très bien pour des images dont le texte est aéré et plutôt de grande taille. Si, au contraire, l'image contient un bloc de texte, la région correspondante pourra être plus haute que large et ainsi avoir un rapport hauteur/largeur supérieur à 1.
3. *La taille :* Un mot ne peut être lisible et donc reconnaissable par un O.C.R si le bloc qui lui correspond après l'opération dilatation-érosion directionnelle est d'une taille inférieure à un certain seuil. On élimine alors toutes les régions inférieures à 50 pixels.
4. *La périodicité du texte* : Les lettres de l'alphabet sont faites de plein et de creux. De plus un mot est composé de lettres non contiguës. A partir des les rectangles exinscrits pour chaque région de texte, la variance du rectangle est calculée. Plus la variance est grande, plus la région est hétérogène et donc susceptible de contenir du texte. Un texte est composé d'au moins trois lettres.
5. *Le profil:* Le profil d'une région correspond à une projection verticale sur une zone définie par le rectangle exinscrit à cette région. L'analyse de chaque région retenue (rectangles exinscrits) sur l'image initiale lissée par reconstruction permet d'établir une courbe correspondante au profil de la région.. La dérivée seconde est approximée par un lissage exponentiel. Le nombre de passage par zéro de la dérivée seconde correspond au nombre de lettres du rectangle. Un texte contenant au moins trois lettres, le profil du rectangle correspondant doit contenir au moins 5 passages par zéro. En dessous de 5, la région n'est pas considérée comme du texte et elle est éliminée.



2.a)



2.b) 2.c)

Figure .2. a) Image de rectangles exinscrits labellisés. b) Profil de la région 1. c) Profil de la région 2. En rouge, le profil de la projection verticale, en vert le profil de la dérivée seconde.

La région 2 correspond à une zone de texte. En effet, on remarque que la projection comporte de nombreux pics de grande amplitude et surtout plusieurs passages par zéro. Tandis que la région 1 ne correspond pas à une zone de texte puisque son profil présente peu de pics et les passages par zéro sont peu fréquents. Dans cette image, seule la région 2 respecte les critères de cohérence et contient du texte. La **figure 3** illustre quelques résultats obtenus par notre méthode de détection de zones de texte dans les images.

Figure .3. Quelques résultats de détection automatique de textes

4. Conclusion et perspectives

L'expérience réalisée sur plusieurs images et sur plusieurs films et les conclusions positives issues de la comparaison avec les résultats d'autres méthodes ont permis une première validation de notre approche. Cette approche peut être facilement combinée avec des techniques de segmentation d'images [12] utilisant les contours actifs. De plus, des structures tels que les familles de R-trees peuvent être utilisées pour le stockage des zones de textes et aussi pour accélérer le processus de recherche. Nous avons présenté une méthode d'extraction de texte qui combine l'utilisation des opérateurs morphologiques et des critères de cohérence spatiale adaptables à toutes les régions dans diverses orientations. L'approche présentée dans cet article, est une contribution au développement d'outils d'indexation par le contenu des images fixes et animées. En effet, la détection, la segmentation et la reconnaissance du texte sont des tâches essentielles dans plusieurs applications telles que la segmentation des journaux télévisés [13]. N

5. Références

[1] J. Serra. Image analysis and mathematical morphology. Academic Press, London 1982

[2] M. Costner, J.L. Chermant, Précis d'analyses d'images. Presses du CNRS, Paris 1989

[3] F. Meyer, P. Maragos, Multiscale morphological segmentations based on watershed, flooding, and eikonal PDE. Lecture notes in computer science, scale-space theories in computer vision, vol. 1682. Springer 1999

[4] RW Brockett, P. Maragos, Evolution equations for continuous-scale morphological filtering, IEEE Trans. Signal processing, vol 42, 3377-3386, 1994

[5] R. Van den Boomgaard, A. Smeulders. The morphological structure of images: the differential equations of morphological scale-space. IEEE Trans PAMI, vol 16, 1101-1113, 1994

[6] A.B. Arehart, L. Vincent, BB Kimia. Mathematical morphology: the hamilton-jacobi connection. Int. Conf. On Computer Vision ICCV'93, Berlin, 11-14 may, 215-219, 1993

[7] H-K KIM - Efficient Automatic Text Location Method and Content-based Indexing and Structuring of Video Database - in Journal of Visual Communication and Image Representation. Special issue on "Indexing, Storage, Retrieval and Browsing of Image and Video". Vol. 7, Num. 4, pp 336--344, décembre 1996

[8] V.Wu, R. Manmatha and E.M. Riseman, Finding text in images. In ACM, editor, Proc. 2nd ACM Int. Conf. On Digital Libraries, 1997

[9] C. Wolf, J.M. Jolion , Extraction de textes dans les vidéos: le cas de la binarisation, RFIA'02, janvier 2002, vol. 1, 145-152.

[10] A.K. Jain and B. Yu. Automatic text location in images and video frames. Technical Report MSU-CPS-97-33, PRIP Lab, Dep. Of Computer Science, 1997

[11] R. Lienhart and W. Effelsberg. Automatic text segmentation for video indexing. Technical report, univ. of Mannheim, Prakt. Informatik, 1998

[12] R. Lienhart and W. Effelsberg. Automatic text segmentation for video indexing. Technical report, univ. of Mannheim, Prakt. Informatik, 1998

[13] R. Lienhart. Localizing and segmeting Texte in Images and Videos, IEEE Transactions and Systems for Video Technology, Vol. 12, n°4, pp 265-266, 2002.