Détection de textes de scènes dans des images issues d'un flux vidéo*

Christian WOLF Jean-Michel JOLION†

LIRIS

Bât. J. Verne, INSA Lyon
69621 Villeurbanne Cedex

{jolion@rfv.insa-lyon.fr}

Résumé

La plupart des travaux sur la détection de texte se concentre sur le texte artificiel et horizontal. Nous proposons une méthode de détection en orientation générale qui repose sur un filtre directionnel appliqué dans plusieurs orientations. Un algorithme de relaxation hiérarchique est employé pour consolider les résultats locaux de direction. Une étape de vote entre des directions permet d'obtenir une image binaire localisant les zones de textes.

Mots clefs

Détection de texte, indexation, multi orientation

1 Introduction

Lors de précédentes études, nous avons mis au point une technique permettant de détecter et reconnaître des textes artificiels dans des images et dans des séquences audio-visuelles [1, 2]. Dans la suite de cette démarche, nous présentons nos résultats sur la détection de textes de scène dans des images numériques.

Le passage de la notion de texte artificiel à celle de texte de scène s'accompagne d'une relaxation dans les contraintes qui définissent le texte artificiel :

Luminance : Un texte artificiel horizontal est assimilé à une accumulation de gradients forts dans cette direction. Cette hypothèse peut être maintenue mais il faut l'étendre à une direction inconnue.

Géométrique : Un texte artificiel obéit le plus souvent à une charte graphique et tout particulièrement en ce qui concerne la police donc la taille. Pour un texte de scène, cette contrainte n'est plus possible. Il est donc nécessaire de développer une approche multirésolution.

Forme : Un texte artificiel est un texte fait pour être lu, superposé au signal. Il ne subit pas de distorsion et sa forme est donc régulière. Ce n'est plus le cas pour un texte de scène qui peut subir les transformations/distorsions du support sur lequel il est imprimé (et surtout si ce support, comme dans le cas d'une banderole, n'est pas rigide). Compte tenu de la diversité des distorsions, nous supposerons encore une non déformation du texte.

Temporel : La principale caractéristique d'un texte artificiel est sa stabilité temporelle, indépendamment du contenu du signal. Dans le cas d'un texte de scène, cette contrainte doit être abandonnée.

La seule vraie contrainte d'un texte qui subsiste est la propriété texturelle (succession de gradients orientés). C'est pourquoi notre méthode s'appuie sur un détecteur de gradients cumulés similaire à celui présenté dans [1] mais déployé dans plusieurs directions.

La bibliographie relative à la détection de textes dans des directions quelconques est très réduites. Les seuls travaux connus utilisent des détecteurs de textes non directionnels. Li and Doermann [3] utilisent un réseau de neurones entraînés sur les coefficients de la transformée de Haar et estiment la direction du texte grâce aux moments d'inertie. Crandall and Kasturi [4] utilisent les coefficients de la DCT du flux MPEG pour détecter les textes et estiment son orientation par maximisation de son recouvrement par un rectangle. Ces deux approches utilisent donc des informations directionnelles de bas niveau et combinent ces informations pour créer un détecteur non directionnel de texte. Cependant, ces deux modèles restent très pauvres.

Nous pensons au contraire que le texte présent dans une vidéo a des caractéristiques directionnelles propres. Un détecteur directionnel est par conséquence nécessaire pour accéder à cette information particulière. L'objectif par cette recherche est, par l'utilisation de caractéristiques plus pertinentes, d'aboutir à une réduction du taux de fausses alarmes qui est un des principaux problèmes de la détection de textes (artificiels ou de scènes).

L'article est organisé de la façon suivante : La section 2 introduira le schéma général de la méthode. La section 3 présentera en détail la relaxation hiérarchique. Les résultats de nos expériences sont donnés dans la section 4, suivie par une conclusion.

2 Le schéma général

Pour rendre notre détecteur moins sensible à la taille du texte à détecter, nous employons une pyramide à plusieurs

*Cette étude est financée par France Télécom R&D dans le cadre du contrat ECAV2 001B575

†Correspondant

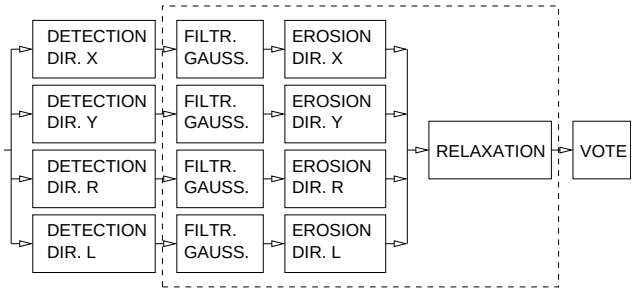


FIG. 1 – *Détection de textes de scènes de direction quelconque*

échelles (normalement 2 niveaux suffisants). La figure 1 montre le schéma du traitement appliqué à chaque échelle. Le texte est détecté dans plusieurs orientations, *i.e.* 4 (0°,45°,90°,135°). Chaque détecteur fournit une *probabilité* pour un pixel d'appartenir à une zone de texte dans une direction donnée. Cependant, un texte est une texture qui ne répond pas à une direction unique lorsque l'on examine les réponses au niveau du pixel. La notion de direction d'un texte ou d'un mot est une caractéristique plus globale. C'est pourquoi nous utilisons un schéma de relaxation pour consolider et/ou modifier les informations locales de direction. Une étape supplémentaire de vote entre des directions permet d'obtenir une image binaire localisant les zones de textes. Enfin, comme nous l'avons évoqué et pour tenir compte de la variabilité des tailles, l'ensemble de ce schéma est reproduit à plusieurs échelles grâce à une structure pyramidale.

Le filtre utilisé pour la détection directionnelle du texte est une amélioration du filtre introduit dans [1]. Ce détecteur utilisait une accumulation de gradients horizontaux, justifiée par le fait que les textes forment une texture régulière contenant des frontières verticales qui est alignée horizontalement dans le cas des textes artificiels. Cependant, ce type de détecteur réagit aussi en présence textures qui ne sont pas des textes. Nous avons donc ajouté une contrainte supplémentaire, l'alignement, majoritaire, des ruptures verticales. La nouvelle version de notre filtre tient donc compte de cette contrainte comme le montre la figure 2. Ce nouveau filtre est une combinaison entre un filtre à accumulation de gradient et un filtre répondant au coin. Comme détecteur de coin nous utilisons une forme particulière de la dérivée seconde (estimée par la sortie du filtre sobel dans la direction y appliquée sur la valeur absolue du filtre de sobel dans direction x) [1]. Ces valeurs sont propagées dans la direction des traits des caractères (direction y). L'orientation peut alors être estimée par le signe de la sortie du filtre de sobel dans la direction y. Une accumulation dans la direction x met en évidence les régions avec une concentration de coins alignés. La réponse finale du détecteur est le produit des deux parties du détecteur.

[1] Les détecteurs de coins connus (e.g. [5]) sont plus sophistiqués et produisent des réponses de coins plus ciblées et plus invariantes. Bien qu'il est facile de les régler pour produire un nombre specifique de points, nécessaire pour l'indexation ou la robotique, il est très difficile de seuiller leur réponse absolue.

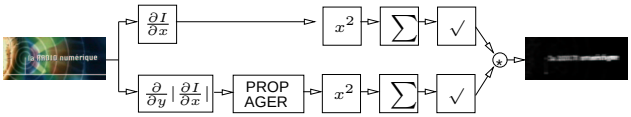


FIG. 2 – *Un filtre pour la détection directionelle du texte*

3 La relaxation hiérarchique

Une estimation de direction est obtenue classiquement par un filtre local. Une information plus stable est obtenue avec la prise en compte de voisinages de tailles significatives dont le coût est souvent prohibitif. Pour palier cet inconvénient, nous utilisons une structure pyramidale (différente de celle utilisée pour la prise en compte de textes de tailles variables). Dans notre approche, nous nous limitons à la détection de 4 directions principales. Cette détection est bien sûr accompagnée de bruit qui peut être atténué par la prise en compte de la cohérence spatiale des réponses par un processus de relaxation. Dans ce processus, chaque sortie de chaque filtre directionnel est réévalué par la prise en compte des directions détectées dans le voisinage du pixel considéré. Notre implantation dans une structure pyramidale permet de limiter l'accès au voisinage à un seul accès aux informations du parent du pixel dans la structure.

L'algorithme de relaxation peut être résumé par les étapes suivantes :

1. Initialiser les bases des pyramides directionnelles (une par direction testée) par les sorties des filtres directionnels.
2. Construire les 4 pyramides directionnelles des niveaux 1 à $N-1$.
3. Recalculer, pour chaque pixel, la sortie de chaque filtre directionnel par la prise en compte de son voisinage par une descente de la pyramide (du niveau $N-1$ jusqu'à la base).
4. Retourner à l'étape 2.

L'algorithme est itéré classiquement jusqu'à convergence, stagnation des valeurs ou épuisement d'un nombre limité d'itérations fixé *a priori*. En pratique, 10 itérations sont suffisantes pour permettre à la cohérence locale de s'affirmer.

Les 4 pyramides que nous utilisons sont construites de manière à tenir compte de leur objectif, *i.e.* un texte dans une direction particulière, dans leur processus de filtrage/sous-échantillonnage. Par exemple, un texte horizontal est un rectangle ayant son axe principal dans la direction horizontale. Le filtre gaussien utilisé dans la

construction de la pyramide horizontale a un support rectangulaire avec une variance dans la direction X plus grande que dans la direction Y. De même pour le filtre appliqué dans la pyramide verticale. Ces deux filtres sont séparables. Ce n'est plus le cas des directions diagonales où nous devons employer un filtre gaussien non séparable. Ces filtres anisotropics seront comparés avec des filtres gaussien isotropics de taille 3×3. Dans tous les cas, nous avons adopté une réduction 2×2 pour le passage entre deux niveaux consécutifs.
Lors de la descente du niveau $N-1$ à la base, la valeur G_i^l du niveau l pour chaque direction i ($i \in \Theta = \{X, Y, R, L\}$) est mise à jour à partir des informations du niveau précédent (nous omettons les coordonnées (x, y) pour des raisons de lisibilité) par

$$G_i^{l\prime} = \frac{G_i^l U_i}{Z} \quad , \quad Z = \sum_{j \in \Theta} P_j^l U_j$$

où Z est une constante de normalisation et U_i le facteur de mise à jour donné par

$$U_i = \sum_{j \in \Theta} \left(\sum_{Q \in Parents(Q)} W(Q) G_i^l(Q) \right)^{\alpha} c(i,j)$$

$c(i,k)$ désigne la compatibilité entre deux directions. Le coefficient α permet une accélération de la convergence. Classiquement, α augmente avec le nombre d'itérations. Le facteur de mise à jour U_i est calculé par une somme sur les 4 directions des produits entre l'avis des parents sur la direction j, et la compatibilité entre cette direction et la direction considérée i.

3.1 Le label "non-texte"

Le mécanismes de relaxation que nous venons de décrire impose une estimation de direction même pour les pixels qui se sont pas liés à du texte. De manière générale, tous les pixels des zones de textes sont perturbés par les pixels n'appartenant pas au texte et pour lesquels l'information de direction n'est pas pertinente. Pour palier ce problème, nous proposons de retirer du processus de relaxation les pixels dont la non appartenance au texte est quasi certaine. Pour cela, il suffit, dans le mécanisme de relaxation, de rajouter un nouveau label "N" qui désigne les zone de non texte. La décision sur la non appartenance à une zone de texte s'appuie sur la valeur maximale des réponses directionnelles. Une approche intuitive serait de procéder à une comparaison avec la valeur maximale théorique du filtre

$$G_N = G_{max} - \max_{i \in \Theta}(G_i)$$

où G_{max} désigne la sortie maximale du filtre. Malheureusement, celle-ci est liée à une configuration très particulière très peu probable dans une image, elle est donc inutilisable pour ce type de test. La seule mesure disponible est la valeur maximale observée dans l'image en cours de traitement.

$$G_{max} = \max_{i \in \Theta}(\max_{x,y} G_i(x,y)))$$

Cette définition du label "non-texte" suppose que l'image contient du texte. Dans ce cas, on peut admettre que la valeur maximale observée est liée à une zone de texte observée dans la direction adéquate. Dans le cas contraire, notre hypothèse conduit à la production de fausses alarmes qui devrons être détectées par les étapes ultérieures de type filtrage morphologique. En effet, on peut attendre des composantes connexes issues de zones de non textes (typiquement des textures) qu'elles ne valident pas les contraintes géométriques et morphologiques associées à une zone de texte.
Le seuil séparant le texte du non-texte est obtenu classiquement par seuillage automatique [6] des distributions des sorties des filtres directionnels. Comme nous avons 4 filtres, nous retenons la valeur maximale du seuil comme seuil final (G_t).
Nous incorporons ces valeurs maximales et de seuil dans un processus de normalisation des sorties des filtres

$$G_i{}' = s(G_i) \quad , \quad i \in \Theta = \{X, Y, R, L, N\}$$

où s est la fonction de normalisation qui tient compte de G_t et G_{max} :

$$s(x) = \begin{cases} \dfrac{x - G_t}{2(G_{max} - G_t)} + 0.5 & \text{si } x \geq G_t \\ \dfrac{x}{2G_t} & \text{sinon} \end{cases}$$

Dans tout le processus de relaxation, le label "N" est traité comme les labels de directions. Sa pyramide associée est construite en utilisant un filtre gaussien isotropique 3×3.
La fonction de compatibilité $c(i,j)$ est proche de zéro pour des directions nettement différentes et proche de 1 pour des directions similaires. (cf. table 1). Un label non-texte est uniquement jugé compatible avec un autre label non-texte et très incompatible avec toutes les labels de directions.

Label	X	Y	R	L	N
X	1	0	0.5	0.5	0
Y	0	1	0.5	0.5	0
R	0.5	0.5	1	0	0
L	0.5	0.5	0	1	0
N	0	0	0	0	1

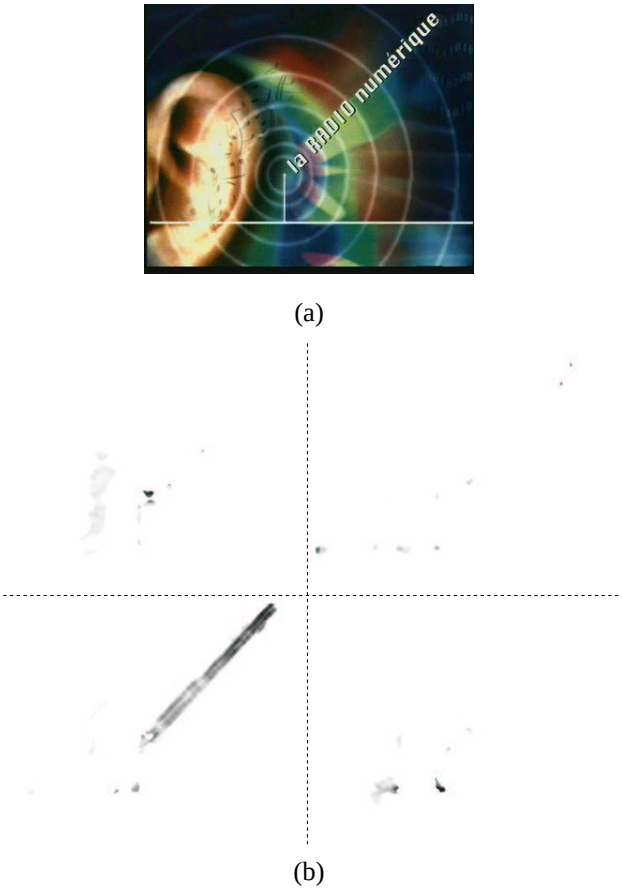
TAB. 1 – *La fonction de compatibilité entre les directions.*

3.2 Binarisation

Un simple système de vote permet de déterminer, pour chaque pixel, son label de direction, associé à la valeur maximale de la sortie des différents filtres. Cette segmentation est combinée avec une segmentation "frustre" obtenue par seuillage automatique de la sortie initiale de chaque filtre. La sortie finale pour chaque segmentation est un "et" entre les deux segmentations. La sortie associée à l'absence de texte est un "ou" sur les 4 directions testées.

Type d. rel.	Labels	Rappel	Précision	H
Pas de Relax.	4	25.0	11.1	7.69
Pyr. isotr.	4	29.5	15.3	10.07
Pyr. anisotr. X,Y	4	34.1	23.7	13.98
Pyr. anisotr.	4	44.7	29.7	17.84
Pyr. anisotr.	5	**48.4**	**36.8**	**20.91**

TAB. 2 – *Résultats pour différents types de relaxation.*



(a) (b)

FIG. 3 – *Image source (a) et résultat du vote après une relaxation anisotropique avec 5 labels (b).*

4 Résultats

Nous avons crée une base de 232 images afin de déterminer des valeurs appropriées pour les différents paramètres de notre méthode et procéder à des tests. Ces images ont été partiellement extraites du corpus de l'INA [2] et partiellement d'une base fournit par notre partenaire France Télécom. Environ la moitié de ces images contiennent du texte artificiel et l'autre moitié des textes de scènes sans caractéristique constante. Les pixels de textes ont été marqués manuellement afin de construire une vérité terrain pour chaque type de texte (artificiel ou de scène) et pour les 4 directions.

Les résultats obtenus sont regroupés dans la table 2 en utilisant des mesures classiques de précision et de rappel. En plus, nous indiquons une mesure unique de performance (H) estimée par la moyenne harmonique des deux mesures précédentes [7]. On peut constater que la relaxation, surtout combinée avec des pyramides anisotropique et un cinquième label, augmente la performance de la détection de façon significative. La figure 3 montre le résultat de la relaxation appliquée à une image. On peut constater que les pixels détectés se concentrent sur quadrant bas-gauche de l'image, qui correspond au filtre diagonal, donc l'orientation du texte.

5 Conclusion et développements futurs

Les résultats de détection obtenus au niveau du pixel sont comparables avec ce que l'on obtient pour des textes artificiels. Il est maintenant nécessaire de prendre en compte des contraintes géométriques et de forme (sans prendre en compte des distorsions). Pour cela, il faut adapter les filtres morphologiques employés pour le texte artificiel (horizontal) à une direction variable. Enfin, un post-traitement pour réduire les fausses alarmes est possible en tenant compte de la dimension temporelle par *block matching* sous certaines hypothèses de mouvement. Ceci constitue le coeur de notre recherche actuelle.

Références

[1] C. Wolf, J.M. Jolion, et F. Chassaing. Procédé de détection de zones de texte dans une image vidéo. Brevet No. FR 01 06776, Juin 2001. France Télécom.

[2] C. Wolf et J.M. Jolion. Vidéo ocr - détection et extraction du texte. Dans *CORESA 2001, 7ème Journées d´Études et d´Echanges "COmpression et REprésentation des Signaux Audiovisuels"*, 12-13 Novembre 2001.

[3] H. Li et D. Doerman. A Video Text Detection System based on Automated Training. Dans IEEE Computer Society, éditeur, *Proceedings of the ICPR 2000*, pages 223–226, 3 Septembre 2000.

[4] D. Crandall and R. Kasturi. Robust Detection of Stylized Text Events in Digital Video. Dans *Proceedings of the International Conference on Document Analysis and Recognition*, pages 865–869, 2001.

[5] C. Harris et M. Stephens. A combined corner and edge detector. Dans *Proceedings 4th Alvey Visual Conference*. Plessey Research Roke Manor, UK, 1988.

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

[7] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd édition, 1979.

[2] http ://www.ina.fr