Extraction de texte dans des vidéos : le cas de la binarisation

Text extraction in videos : About binarization *

C. Wolf[1] J.M. Jolion[1]

[1] Laboratoire Reconnaissance de Forme et Vision

Bât. J. Verne
INSA Lyon
69621 Villeurbanne Cedex
{wolf,jolion}@rfv.insa-lyon.fr

Résumé

Dans cet article nous abordons la problème de la binarisation de "boites", i.e. sous-image, contenant du texte. Nous montrons que la spécificité des contenus vidéos amène à la conception d'une nouvelle approche de cette étape de binarisation en regard des techniques habituelles tant du traitement d'image au sens large, que du domaine de l'analyse de documents écrits.

Mots Clef

Binarisation, seuillage, vidéo, textes artificiels.

Abstract

We present in this paper some researches on thresholding of "text boxes" (sub-images containing artificial texts and extracted from videos). We show that the particular context of videos leads to the formalization of a new approach of this step regarding the usual and wellknow techniques used in image analysis and more particularly for segmentation of written documents.

Keywords

Binarization, thresholding, video, artificial texts.

1 Introduction

Depuis quelques années, les documents audiovisuels numérisés sont de plus en plus fréquents. Des grandes bases de données audiovisuelles ont été crées par des entreprises, des organisations et aussi par des personnes privées. Cependant, l'utilisation de ces bases reste encore problématique. Tout particulièrement, ces nouveaux types de données, image et vidéo, ont conduit à de nouveaux systèmes d'indexation où la recherche par le contenu se fait à partir d'une image exemple (Qbic, VisualSEEK, SurfImage, MARS ...), grâce à des mots clefs ou les deux (WebSEEK, Two.Six, Virage).

Notre objectif n'est pas ici de traiter de la problématique de l'indexation en général. On peut cependant constater que les résultats ne sont pas encore à la hauteur des attentes des usagers potentiels. La principale raison se situe dans l'extrême difficulté rencontrée lors du passage des informations liées au signal à celles qui relèveraient de la sémantique portée par celui-ci. Cette difficulté n'est d'ailleurs pas nouvelle [2].

Cependant, entre le niveau du pixel, et celui de la sémantique, il existe des caractéristiques à la fois riche en information et cependant simples. Le texte présent dans les images et les vidéos fait partie de cette catégorie. Une fois extrait, ce texte peut alimenter les index qui caractérisent un plan ou une séquence au même titre que des indicateurs plus proches du signal.

Bien sûr, afin de soulager le travail du documentaliste, il est nécessaire de faire en sorte que la détection (et la reconnaissance) du texte soit la plus automatique possible.

Nous avons abordé le problème général de la détection de textes dans les vidéos dans notre projet ECAV (Enrichissement de Contenu Audiovisuel) en collaboration avec France Télécom. Celui-ci a donné lieu à un brevet [10]. Une présentation générale peut être trouvée dans [11].

Dans cet article, nous n'abordons qu'un des aspects de cette étude. En effet, au delà de l'aspect ingénierie inhérent au développement d'un système efficace et robuste, nous avons, au cours de notre étude, mis en évidence des manques en termes de modélisation formelle et donc d'outils qui en découlent.

Tout particulièrement, nous nous focaliserons ici sur le cas de la segmentation, *i.e.* binarisation d'une sous-image dans lequel un processus approprié de détection

*Cette étude a bénéficié du soutien de France Télécom Recherche et Développement dans le cadre du projet ECAV 001B575.

a mis en évidence la présence de textes artificiels avec une forte probabilité. L'étape qui nous intéresse ici se situe donc en aval de cette détection et en amont de la phase de reconnaissance du texte par un logiciel d'OCR qui nécessite une donnée de type binaire.
Nous allons tout d'abord présenter les spécificités des données de type vidéo. Nous verrons ensuite un rapide survol des principales approches de la binarisation et tout particulièrement de celles qui sont utilisées dans le traitement des documents. Nous pourrons alors voir en détails notre approche que nous comparerons sur des exemples variés.

2 Sur la nature des données

2.1 Vidéo vs document numérique

Les recherches en détection et extraction du texte à partir des séquences vidéo sont encore confrontées à de sérieux problèmes. Pourtant la recherche dans le domaine de l'OCR des documents classiques (c.a.d. les documents écrits, les journaux etc.) a développé des méthodes performantes et des outils commerciaux produisent de bons résultats pour des images de documents. Le problème principal peut être expliqué par la différence entre l'information présente dans un document et celle donnée par une séquence vidéo ainsi que les méthodes de stockage de chaque type de données.
Les images de documents sont créées en vu de les numériser pour les passer ensuite à une phase OCR pour reconnaître la structure et le texte. Pour améliorer la qualité et le taux de la reconnaissance, les images sont numérisées à très haute résolution (200-400 dpi) donnant des fichiers de taille très élevée (un fichier de 100 Mo résulte d'une page A4 numérisée à 400 dpi). Les fichiers sont comprimés sans perte pour garder la qualité et empêcher des artéfacts de compression. Ces grandes tailles ne sont pas une limite puisque ni leur transport ni leur stockage ne sont prévus. La plupart du temps les images de documents sont bien structurées et contiennent un fond uniforme et la couleur du texte est également uniforme. Ceci permet de séparer les caractères du fond avec un seuillage fixe ou adaptatif. La majorité de la page contient des caractères structurés dans différents paragraphes, quelques images sont inclues dans la page.
Par contre, les images des séquences vidéo contiennent de l'information plus difficile à traiter. Le texte n'est pas séparé du fond. Il est soit superposé (le "texte artificiel" comme les sous-titres, les résultats de sport etc.) soit inclut dans la scène de l'image (le "texte de scène", par exemple le texte sur le tee-shirt d'un acteur). Le fond de l'image peut être très complexe ce qui empêche une séparation facile des caractères. De plus, contrairement aux documents écrits, les séquences vidéo contiennent de l'information très riche en couleurs. Enfin, le texte n'est pas structuré en lignes, et souvent quelques mots courts et déconnectés flottent dans l'image.

2.2 La faible résolution

Le but principal de la conception des formats de fichiers vidéo est de garder une qualité suffisante pour l'affichage, pour le stockage et le transport par des réseaux informatiques. Pour cette raison et pour une quantité de données vidéo plus haute, il est nécessaire de réduire fortement l'information avant son stockage dans le fichier vidéo. Pour limiter la taille de l'information deux méthodes sont souvent appliquées : la forte réduction de la résolution et le codage avec perte, c.a.d. avec élimination des données redondantes et perte d'une partie de l'information originale, en utilisant les algorithmes de compression JPEG et MPEG. La résolution spatiale de la vidéo dépend de l'application et prend des valeurs entre 160×100 pixels (diffusion par Internet) et 720×480 pixels (DVD vidéo codé en MPEG 2). Un format typique est le format CIF avec 384×288 pixels. Notons que la taille du texte affiché avec cette résolution est beaucoup moins grande que le texte résultant d'un document numérisé. Les frames d'une vidéo de cette résolution contiennent des caractères d'une hauteur de moins de 10 pixels, alors que dans les documents numérisés, les caractères ont une hauteur comprise entre 50 et 70 pixels.

2.3 Les effets d'aliasing

Pendant la production de la vidéo, le signal original est sous échantillonné plusieurs fois. Pour empêcher des effets d'aliasing, des filtres passe bas sont appliqués. Le signal résultant a alors une qualité suffisante pour la lecture mais il est complètement lissé.
Après la réduction de la résolution, la compression du signal ajoute également des artéfacts. L'information supprimée par le schéma MPEG est considérée comme redondante pour le système de visuel humain. Cependant, les artéfacts de l'encodage causent des problèmes pendant la reconnaissance (par exemple en perturbant l'uniformité des couleurs). Même si les nouvelles normes comme JPEG 2000 apportent un plus en regard de ce type de problèmes, l'existence de très nombreuses données codées selon le format JPEG classique justifie nos recherches dans cette direction.

2.4 Le fond complexe et détaillé

Le fond sur lequel est inscrit le texte ne peut être considéré comme constant. Un seuillage global, même déterminé localement, n'est le plus souvent pas suffisant pour clairement mettre en évidence le texte.
Enfin, afin de faciliter la lecture du texte, des artifices d'inscription du texte sont utilisés (e.g. des ombres artificielles qui augmentent le contrast avec le fond). Ces techniques ne sont malheureusement pas un avantage pour la détection. L'environnement du texte peut

être considéré comme bruité mais pas avec un modèle simple de type additif.

2.5 Quelques hypothèses

Dans ce qui suit, nous ferons les hypothèses suivantes (qui découlent des spécificités du texte dans les vidéos). Tout d'abord, nous assumerons une gamme fixe de taille de police. Plus exactement, nous supposerons que dans une boite de texte potentiel, il n'existe qu'une seule police de caractères. Ensuite, nous supposerons que le texte peut être discriminé du fond sur la composante luminance (ce qui est en fait l'hypothèse principale de toute méthode de seuillage). Sans que cela soit une contrainte trop forte, nous supposerons par la suite que le texte est plus foncé que le fond. Enfin, compte tenu de la variété du fond, nous imposons une recherche adaptative de seuil, *i.e.* par fenêtre locale. La taille de la fenêtre devra être fonction de la taille des caractères afin que toute fenêtre ne puisse être totalement incluse dans un caractère mais contienne toujours une part significative de pixels du fond.

3 Sur la binarisation

Proposer une nouvelle approche de la binarisation est, au premier abord, un peu surprenant tant ce domaine a été étudié par le passé et ceci dès le début des travaux sur les images numériques. En se référant à notre travail, on pourra citer deux types d'approches.

3.1 Les méthodes générales

Il s'agit là des méthodes à usage général. On peut y inclure les deux approches les plus célèbres. Tout d'abord, la méthode de Fisher/Otsu (issue du domaine de l'analyse de données et optimisée par Otsu [7]), encore dénommée méthode de minimisation de la variance (implicitement on fait ici référence à la variance intraclasse), s'appuie sur des caractéristiques statistiques des deux classes de pixels définies par le choix d'un seuil. Il n'y a donc pas vraiment de modèle sous-jacent sur la forme de l'histogramme. Par contre, la taille des populations doit permettre une estimation des paramètres statistiques, *i.e.* il y a explicitement référence à deux classes.
Nos tests ont montré que cette approche, bien que très générale, présentait de nombreux avantages pour les images qui nous intéressent, c.à.d des images bruitées de résolution basse (voir section 2). Cependant, les résultats sont en deçà de ceux obtenus avec les méthodes que nous présenterons par la suite. Nous ne l'avons donc pas inclue dans les comparaisons présentées dans cet article.
La méthode de minimisation de l'erreur de seuillage s'appuie quant à elle sur une modélisation de l'histogramme en terme de somme de distributions. La position optimale du seuil est alors définie par l'intersection de ces distributions. Une solution simple est connue dans le cas où les distributions sont normales [4]. Dans notre approche, il n'est pas possible d'obtenir des histogrammes pour lesquels une approximation de courbes donne de bons résultats. En effet, la taille des fenêtres est typiquement de l'ordre de 30×30 pixels. La forme de l'histogramme induit est très fortement bruitée et ne donne pas une bonne approximation des éventuelles distributions sous-jacentes.
Il est nécessaire de s'appuyer sur des critères plus globaux.
La principale critique que l'on peut faire à ces techniques est leur généralité, c'est à dire leur faible adaptabilité à un contexte particulier. C'est pourquoi il semble plus judicieux de s'orienter vers des techniques relevant d'un domaine similaire à la vidéo.

3.2 Les méthodes du domaine des documents numériques

Comme nous l'avons évoqué, l'extraction d'une zone de texte en vue de sa reconnaissance/transcription par un logiciel d'OCR est une étape clef des logiciels d'analyse de documents. C'est pourquoi l'on y retrouve des techniques adaptées aux spécificités de ce type d'images (tout particulièrement la grande densité). On pourra se référer à [1] pour une étude comparative. La plus performante de ces méthodes est sans doute celle proposée par Niblack [6]. Il s'agit d'une approche adaptative où un seuil T est déterminé en tout point par la formule

$$T = m + k.s \tag{1}$$

où m, s et k désignent respectivement la moyenne des niveaux de gris sur la fenêtre centrée sur le point, l'écart type de cette distribution et une constante que Niblack a fixée empiriquement à -0.18. Le résultat est peu dépendant de la taille de la fenêtre même si l'on constate que les meilleurs résultats sont obtenus lorsque celle-ci englobe 2 ou 3 caractères. Bien que très simple, cette méthode ne permet pas de traiter le cas de la présence de textures dans le fond qui sont le plus souvent détectées comme du texte à faible contraste.
On doit à Sauvola *et al.* [8] une amélioration qui permet de tenir compte de cette remarque.

$$T = m.(1 + k.(\frac{s}{R} - 1)) \tag{2}$$

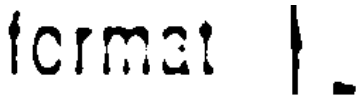
où R traduit la dynamique du paramètre s [1] et est fixée empiriquement à 128. Si le seuil se situe de nouveau proche de la moyenne, celui-ci s'adapte mieux au contenu de l'image sans pour autant répondre à toutes les contraintes que nous avons évoquées.
La figure 1 montre un exemple d'application de ces méthodes sur une boite de texte. Comme nous l'évoquions précédemment, la méthode de Niblack

[1] De plus Sauvola propose la valeur 0.5 pour la constante k.

extrait non seulement des structures importantes (à droite) ce qui est logique compte tenu de leur luminance très sombre, mais aussi des textures du fond (juste après la dernière lettre du mot, moins visible à l'oeil dans l'image originale) du fait de la localité du seuillage. Sur le même exemple, la méthode de Sauvola est moins sensible à ces bruits mais son seuillage, plus strict, conduit à des caractères extraits plus fins et présentant par endroit des coupures très préjudiciables à toute reconnaissance sans un post traitement de type morphologique.

(a) Image originale

(b) Niblack



(c) Sauvola

Figure 1: Exemple de binarisation par les méthodes de Niblack et Sauvola sur une image contenant du texte.

4 Notre proposition

Nous avons focalisé notre étude sur les approches statistiques, c'est à dire celles qui utilisent une caractérisation statistique de l'image à seuiller. Compte tenu de ce que nous avons spécifié dans nos hypothèses, nous supposerons qu'une fenêtre centrée en un point ne peut être que de deux types, soit elle contient uniquement des pixels du fond, soit elle contient à la fois des pixels du fond et des pixels du texte.
La formule de Sauvola peut être réécrite sous la forme

$$T = m + k\alpha m \tag{3}$$

où $\alpha = \frac{s}{R} - 1$.
Le terme correctif par rapport à la moyenne dépend donc de la dynamique locale mais aussi de la valeur de la moyenne. Plus celle-ci sera haute, plus le seuil sera diminué [2]. Les textures du fond qui seraient claires induiront donc un seuil bas qui induira une fenêtre binarisée appartenant totalement à la classe fond. Sous l'hypothèse que le texte est plus foncé que le fond, cela semble résoudre le problème des textures de fond. La motivation de cette amélioration trouve sa justification dans le domaine des documents écrits lorsque le contraste local est faible. Cette hypothèse est cependant trop stricte pour les vidéos où il est nécessaire de prendre en compte une plus grande variété qui ne peut se réduire dans une valeur constante de la dynamique de référence, *i.e.* $R = 128$.
La figure 2 montre un effet de surluminance de l'ensemble de l'image.

(a) Image originale

(b) Niblack

(c) Sauvola

Figure 2: Exemple de binarisation par la méthode de Sauvola lorsque l'image est globalement claire. L'ensemble de l'image est classée comme fond. La méthode originelle de Niblack fournit quant à elle un résultat correct.

Afin de coller au mieux aux données, nous proposons de normaliser les éléments intervenant dans l'équation qui définit le seuil T. Cela revient à ne plus raisonner sur les valeurs absolues mais sur les contrastes, ce qui est naturel si l'on regarde les différentes justifications des méthodes de Niblack [3] et de Sauvola qui argumentent sur la notion de contraste.
Comment peut-on définir ces contrastes ? Si l'on se réfère à [6], p. 45, le contraste, au centre de la fenêtre de niveau de gris I, est défini par

$$C_L = \frac{|m - I|}{s} \tag{4}$$

Comme nous avons supposé un texte sombre sur un fond clair, nous ne considérerons pas les points ayant un niveau de gris plus grand que la moyenne locale, la valeur absolue pourra donc être éliminée.
Si l'on note M la valeur minimale des niveaux de gris de l'image, la valeur maximale de ce contraste local, notée C_{max} est donnée par

$$C_{max} = \frac{m - M}{s} \tag{5}$$

Il est difficile d'appuyer une stratégie de seuillage sur la seule comparaison entre le contraste local et sa valeur

[2] Sauvola propose R fixé à 128.*Onadonc*s ¡ R qui conduit à des valeurs négatives pour α.

[3] Niblack situe l'origine de son travail dans les méthodes de transformation d'histogramme de niveaux de gris pour augmenter le contraste d'une image.

maximale car cela ne permet pas de prendre en compte la variabilité de la fenêtre centrée sur le point en regard du restant de l'image. C'est pourquoi nous proposons de définir un contraste plus global, le contraste de la fenêtre, noté C_F, par

$$C_F = \frac{m - M}{R} \tag{6}$$

où R désigne la valeur maximale des écarts type pour toutes les fenêtres de l'image.
Ce contraste permet de savoir si la fenêtre est plutôt sombre ou claire par rapport à l'ensemble de l'image (une valeur forte caractérisera l'absence de texte). Nous pouvons maintenant exprimer notre stratégie en fonction de ces contrastes.
Un critère de seuillage simple consiste à ne conserver que les points ayant un contraste local proportionnellement fort en regard de la valeur maximale corrigée par le contraste de la fenêtre centrée sur ce point.

$$I : C_L > a(C_{max} - C_F) \tag{7}$$

où a est un paramètre de gain.
En développant cette équation, on obtient la valeur du seuil

$$T = (1 - a)m + aM + a\frac{s}{R}(m - M) \tag{8}$$

Dans le cas où l'on se trouve sur la fenêtre de variabilité maximale, *i.e.* $s = R$, on obtient $T = m$. On incite le processus à conserver le maximum de points sur la fenêtre. Par contre, dans le cas où la variabilité est faible (*i.e.* $s \ll R$), il y a une probabilité forte d'absence de texte. On ne conservera un pixel que si il a un fort contraste local. Le seuil est obtenu par $T \approx (1 - a)m + aM$. Le paramètre a permet de régler la marge d'incertitude autour de la moyenne. Une solution simple consiste à fixer la valeur de ce paramètre à 0.5, le seuil se situant alors à mi-distance entre M et m.
La figure 3 montre un exemple d'application de notre méthode sur les deux images utilisées aux figures 1 et 2. Les résultats sont corrects comparativement aux autres méthodes.

(a) Image de la figure 1

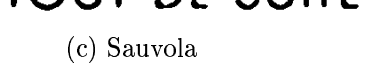
(b) Image de la figure 2

Figure 3: Exemple de binarisation par notre méthode pour les deux images présentées aux figures 1 et 2.

La figure 4 montre un autre exemple pour les trois méthodes que nous avons détaillées pour une image qui contient un fond majoritaire (donc influent sur les paramètres statistiques) ainsi qu'une dérive lumineuse. Notre approche combine la robustesse de la méthode de Sauvola vis à vis du fond et la qualité de délimitation des caractères de la méthode de Niblack.

(a) Image originale

(b) Niblack

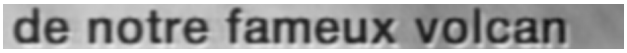


(c) Sauvola

(d) Notre approche

Figure 4: Exemple de binarisation par les différentes méthodes présentées dans cet article.

La figure 5 présente un exemple plus complexe de part la mise en forme des caractères (effet de relief) et la présence d'un fond très sombre. Le résultat confirme le comportement des trois méthodes.



(a) Image originale

(b) Niblack

(c) Sauvola

(d) Notre approche

Figure 5: Exemple de binarisation par les différentes méthodes présentées dans cet article.

Enfin, la figure 6 présente un dernier exemple où le fond a une texture complexe. Nous avons, pour cet exemple, ajouté la méthode de Fisher globale (seuil unique pour toute l'image) et locale (un seuil par fenêtre). La méthode de Fisher locale est très similaire à celle de Niblack de part son comportement sur les textures du fond. Notre approche fournit un résultat comparable à la méthode de Sauvola.

(a) Image originale

(b) Fisher

(c) Fisher local

(d) Niblack

(e) Sauvola

(f) Notre approche

Figure 6: Exemple de binarisation par les différentes méthodes présentées dans cet article.

5 Évaluation

Il est souvent plus simple de proposer une nouvelle technique que d'en prouver l'utilité en regard des techniques existantes. Comment pouvons-nous, en dehors de quelques exemples bien choisis, prouver l'intérêt de ce nouveau schéma de binarisation. Une approche possible est de créer artificiellement des images de référence perturbées en utilisant des modéles des dégradations déjà connus. Cependant, nous pensons que cette approche ne démontrera pas grand chose sur la réelle utilité de notre technique dans le cas de la vidéo. En effet, celle-ci n'a pas été conçue pour devenir une technique générique pour toutes les images mais pour répondre à un objectif bien précis, l'extraction de texte dans des vidéos **en vue de leur reconnaissance**. Nous jugerons donc de l'intérêt de la technique par la performance induite sur la phase de reconnaissance. En effet, meilleure sera la binarisation, plus facile sera la reconnaissance.

Le protocole que nous avons choisit est donc le suivant. Les boites de texte une fois binarisées [4] sont passées à un OCR (Finereader). Chaque chaîne de caractères est ensuite évaluée par la méthode de Wagner et Fisher [9]. Soient α et β deux caractères appartenant respectivement à une chaîne reconnue et à une chaîne de référence, *i.e.* la vérité terrain. La fonction de coût, γ, est définie par

$$\gamma(\alpha,\beta) = \begin{cases} 0 & \text{si } \alpha = \beta \ , \ \alpha,\beta \in X \bigcup \{\lambda\} \\ 0.5 & \text{si } \alpha = \beta \ , \ \alpha,\beta \in X \bigcup \{\lambda\} \\ & \text{et } \alpha \text{ et } \beta \text{ sont dans des formats} \\ & \text{(min./maj.) différents} \\ 1 & \textit{sinon} \end{cases} \tag{9}$$

où X est l'ensemble des caractères et λ le caractère "vide" pour lequel une fonction de coût spécifique est requise..

$$\gamma(\lambda,\beta) = \begin{cases} 0.5 & \text{si } \beta \text{ est un espace} \\ 1 & \textit{sinon} \end{cases} \tag{10}$$

et

$$\gamma(\lambda,\beta) = \gamma(\beta,\lambda) \tag{11}$$

En plus de cette fonction de coût, il est aussi possible de caractériser la reconnaissance du texte sur la base de test par les mesures de précision et de rappel.

$$\text{Précision} = \frac{\text{Nombre de caractères reconnus}}{\text{Nombre de caractères proposés par l'OCR}}$$

$$\text{Rappel} = \frac{\text{Nombre de caractères reconnus}}{\text{Nombre de caractères dans la référence}}$$

Les résultats qui sont résumés dans le tableau 1 ont été obtenus pour 4 vidéos extraites de la base mise à disposition par l'INA. Cela représente 60000 frames (environ 40 minutes) contenant 322 occurrences de textes. La figure 7 propose des extraits de ces vidéos qui montre bien la grande variété des apparitions de textes.
Dans cette étude, deux versions de la méthode de Sauvola ont été testées. La première correspond à celle que nous avons présentée (cf. Eq. 2). Dans la deuxième, nous avons simplement remplacé la valeur fixe de R (=128) par une valeur adaptative identique à celle que nous avons introduite dans notre approche, *i.e.* $R = max(s)$. Sur les trois indicateurs, rappel, précision et coût, cette simple modification induit un gain significatif (surtout sur le rappel) et lui permet de dépasser la méthode de Niblack. A l'exclusion d'un indicateur, la précision, et sur une seule vidéo, notre approche fournit les meilleurs résultats ce qui montre sa pertinence et sa robustesse.

6 Discussion

L'approche que nous avons proposée dans cet article permet, grâce à la prise en compte de paramètres

[4] Dans tous nos essais, la taille de la fenêtre a été fixée à 30×30 pixels.

AIM2 (INA) : publicités (France 3, TF1)

AIM3 (INA) : journal télévisé (M6, Canal+)

AIM4 (INA) : dessin animé et journal télévisé (Arte)

AIM5 (INA) : journal télévisé (France 3)

Figure 7: Extraits de la base vidéos de test.

Vidéo	Méthode	Rappel	Précision	Coût
AIM2	Niblack	67.4%	87.5%	499
	Sauvola R=128	53.8%	87.6%	616.5
	Sauvola R ad.	75.3%	87.8%	384.5
	Notre approche	**78.4%**	**90.4%**	**344.5**
AIM3	Niblack	92.5%	78.1%	196
	Sauvola R=128	69.9%	89.6%	206
	Sauvola R ad.	85.3%	92.5%	110
	Notre approche	**96.2%**	**95.3%**	**51**
AIM4	Niblack	78.5%	**92.0%**	252
	Sauvola R=128	48.6%	87.7%	490.5
	Sauvola R ad.	69.8%	84.6%	360.5
	Notre approche	**80.1%**	90.4%	**211.5**
AIM5	Niblack	62.1%	71.4%	501.5
	Sauvola R=128	66.7%	89.3%	324.5
	Sauvola R ad.	64.9%	90.1%	328
	Notre approche	**69.0%**	**91.0%**	**294.5**
Total	Niblack	73.1%	82.6%	1448.5
	Sauvola R=128	58.4%	88.5%	1637.5
	Sauvola R ad.	73.0%	88.4%	1183
	Notre approche	**79.6%**	**91.5%**	**901.5**

Table 1: Comparaison des méthodes de seuillage par la qualité de la reconnaissance des caractères sur les images binaires produites.

statistiques globaux, une meilleure adaptativité à la variété des contenus des vidéos. Le prix à payer se situe au niveau computationnel puisque ces paramètres supplémentaires requièrent un traitement en deux passes. Cependant, ce coût est très faible au regard, d'une part du gain sur les performances, et d'autre part du coût global incluant les phases de détection des boites de texte et de reconnaissance par le logiciel d'OCR.

Cette nouvelle méthode n'est *a priori* pas définitive en ce sens que d'autres paramètres epuvent être introduit. A titre d'exemple, nous avons évoqué (cf. Eq. 8) le cas où la variabilité locale est faible ($s \ll R$). Il serait intéressant de pas lier cette propriété à la seule comparaison de s à la valeur maximale mais également à la valeur minimale de l'écart type. Cette autre valeur extrême permet de quantifier le bruit additif à l'image et par là même d'accéder au rapport signal/bruit de l'image. De nombreuses techniques existent pour l'estimation de cette variance minimale (par exemple [5]). Elles sont souvent coûteuses en temps et leur adaptativité aux spécificités de la vidéo doit être préalablement étudiée.

Au delà du seul objectif de reconnaissance, nous souhaitons aussi étudier l'apport de ces indicateurs pour filtrer plus efficacement les fausses alarmes, *i.e.* les boites de texte qui ne contiennent pas de texte, qui sont encore trop nombreuses dans notre système (en conséquence de notre soucis de ne pas louper de

texte) et qui induisent un surcoût de traitement dans la phase de reconnaissance du texte.
Notre approche s'appuie sur une formalisation en termes de contraste et peut donc être reliée à tous les travaux, très nombreux en traitement d'images. En particulier, il sera nécessaire de faire le lien avec la méthode de Kholer [3].
Enfin, nous avons, pour le moment, focalisé notre étude sur le texte dit artificiel, c'est à dire incrusté dans une scène par surimpression. Afin de fournir des informations plus complètes pour construire des index plus informatifs, il est nécessaire d'étendre cette recherche aux textes dits de scène, c'est à dire totalement inclus dans le contenu de la scène. Ceci constituera le cadre de la poursuite de notre collaboration avec France Télécom Recherche & Développement.

References

[1] O. Due Trier & A.K. Jain, Goal-Directed Evaluation of Binarization Methods, *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 17, 12, pp. 1191–1201, 1995.

[2] J.M. Jolion, Indexation d'images: nouvelle problématique ou vieux débat ?, Rapport de recherche RR 98.05, RFV, INSA Lyon, Octobre 1998.

[3] R. Kholer, A segmentation system based on thresholding, *Comput. Graphics and Image Proc.*, vol. 15, pp. 241–245, 1981.

[4] J. Kittler & J. Illingworth, Minimum Error Thresholding, *Pattern Recognition*, vol. 19, 1, pp. 41–47, 1986.

[5] P. Meer, J.M. Jolion & A. Rosenfeld, A Fast Parallel Algorithm for Blind Estimation of Noise Variance, *IEEE Trans. on Patt. Anal. Mach. Intell.*, vol. 12, pp. 216–223, 1990.

[6] W. Niblack, *An Introduction to Digital Image Processing*, Englewood Cliffs, N.J., Prentice Hall, pp. 115–116, 1986.

[7] N. Otsu, A threshold selection method from gray level histograms, *IEEE Trans. Syst. Man Cyber.*, vol. SMC-9, 1, pp. 62–66, 1979.

[8] J. Sauvola, T. Seppänen, S. Haapakoski & M. Pietikäinen, Adaptive Document Binarization, *Proc. of the Intern. Conf. on Document Analysis and Recognition*, vol. 1, pp. 147–152, 1997.

[9] R.A. Wagner & M.J. Fisher, The string to string correction problem, *Journ. of Assoc. Comp. Mach.*, vol. 21, 1, pp. 168–173, 1974.

[10] C. Wolf, J.M. Jolion & F. Chassaing, Système de détection et de suivi de texte dans les images vidéos, *brevet France Télécom*, 01 06776, 2001.

[11] C. Wolf & J.M. Jolion, Vidéo OCR : détection et extraction de textes, Orasis 2001, Cahors, 5-8 Juin 2001.