Segmentation en Caractères Individuels dans des Images de Scènes Naturelles

Céline MANCAS-THILLOU, Matei MANCAS, Bernard GOSSELIN

Faculté Polytechnique de Mons, Laboratoire de Théorie des Circuits et Traitement du Signal
Bâtiment Multitel - Initialis, 1, avenue Copernic, 7000, Mons, Belgique
celine.thillou@tcts.fpms.ac.be, matei.mancas@tcts.fpms.ac.be
bernard.gosselin@tcts.fpms.ac.be

Résumé – Grâce à l'essor des appareils miniatures surmontés de caméras basse résolution, parfois même un peu gadgets, de nouveaux challenges sont apparus pour analyser les images de scènes naturelles. Le texte présent dans ces images est difficilement reconnu par les logiciels OCR actuels en raison des nombreuses dégradations. Nous détaillons dans cet article une méthode de segmentation de caractères collés présents dans des images photographiques. Le principal but est de segmenter ces caractères en composants individuels pour faciliter et améliorer leur reconnaissance. Les premiers résultats de notre méthode éprouvée sur une base de données publique sont également présentés.

Abstract – Embedded applications are growing everyday with the multiplicity of tiny devices and cameras. Hence, new challenges appear in image processing with natural scene images. Characters present in this kind of images make off-the-shelf OCR softwares fail and we describe in this paper a method to handle touching characters in camera-based images. The main goal is to segment them into individual components to improve recognition rates, required in numerous applications. Using a public database, our first results are then detailed.

1 Introduction

La segmentation en unités distinctes et la reconnaissance de caractères ont été éprouvées depuis plus de deux décennies. Néanmoins, les systèmes commerciaux sont encore trop sensibles aux nombreuses dégradations dans les images. Les images de scènes naturelles, acquises ici par un appareil photo basse résolution comportent de nombreuses dégradations telles que le flou, la mauvaise illumination et la distorsion, absentes dans une acquisition par scanner, et une taille et police de caractères parfois extravagantes. Cet article traite principalement de la segmentation en unités distinctes des caractères collés. Dans les images de scènes naturelles, une majorité de ces caractères est présente si aucun traitement n'est effectué et la solution ne peut parfois pas être trouvée à l'étape précédente d'extraction de texte ou binarisation. Leur suppression est néanmoins indispensable pour obtenir des résultats de reconnaissance satisfaisants dans un environnement embarqué comme l'est l'acquisition des images de scènes naturelles et pour leurs nombreux usages dans les applications (aide aux personnes malvoyantes et aveugles, traducteur de poche...). En effet, les logiciels de reconnaissance actuellement disponibles sont lourds et très sensibles à un grand nombre de caractères collés.

Les algorithmes trouvés dans la littérature basent leurs tests sur du texte présent dans des flux vidéo principalement [9] en utilisant l'information présente dans les multiples trames disponibles. Cette information est manquante dans le cas d'images fixes, comme celles présentes dans notre base de données (issue de la compétition publique ICDAR 2003) illustrée avec quelques exemples dans Figure 1. La super-segmentation, exercée sur une analyse de documents acquis par scanner et qui a pour but de segmenter les caractères en primitives, a été testée

FIG. 1 – Exemples de la base de données publique Icdar 2003 [4]

pour donner des résultats satisfaisants lorsque la police des caractères n'est pas trop imaginative car alors le reconnaisseur de caractères peine à fusionner les différentes primitives. Des modèles [1] basés sur une étude préliminaire de la dégradation et des caractères présents ont donné des résultats encourageants mais la définition et application de ces modèles sont lourds en temps de calcul pour un usage embarqué et la diversité des polices accroît le nombre de modèles de manière exponentielle. Un état de l'art [2] basé sur les méthodes de segmentation de caractères acquis au moyen d'un scanner montre que ces techniques ne peuvent s'appliquer ici en raison des nouvelles dégradations qui sont absentes avec une numérisation par scanner. De plus, elles présentent trop de paramètres à définir, perdant la généricité, exigée par les images de scènes naturelles, qui, par

définition, sont aussi variées que possible. D'autres méthodes basées sur des modèles ou des marqueurs ont été utilisés [6] mais le plus délicat est d'approcher suffisamment le résultat pour initialiser les contours actifs ou autres marqueurs. En effet, si les marqueurs ou les modèles sont initialisés en utilisant une approximation trop éloignée de la réalité, le résultat ne sera pas satisfaisant.

Après cette brève description du travail déjà effectué, la première constatation est qu'il n'existe aucun algorithme de segmentation des caractères adapté aux images de scènes naturelles. Ainsi il convient de détailler notre méthode de segmentation des caractères collés et de présenter les premiers résultats obtenus. Enfin, nous discuterons les avantages de notre algorithme tout en présentant nos futurs travaux à ce sujet.

2 Méthode

2.1 Une approche haut niveau

Afin d'approcher le texte, nous utilisons une technique basée sur un débruitage effectué grâce à une décomposition en ondelettes suivie d'une classification non supervisée. Cette méthode d'extraction de texte (MET) a déjà été exposée lors d'articles précédents [8] et a prouvé sa généricité. Cette première approche permet d'avoir une boîte englobante précise des zones de texte et de voir le nombre de caractères collés diminuer pour n'être présent que dans 29.2% des images de notre base de données [4] grâce à un souci de segmentation future présent dans notre méthodologie.

Les principales étapes de MET sont :

- Débruitage avec les ondelettes de Daubechies en 16 niveaux. Conservation des hautes fréquences et de la plus basse fréquence.
- Discrimination binaire de la classe du fonds de l'image (simple, complexe) pour application de l'algorithme de seuillage global et rapide d'Otsu [7] dans les cas simples.
- Réduction du nombre des couleurs sans dégradation.
- Clustering K-moyennes non-supervisé des couleurs avec K=3.
- Combinaison éventuelle entre les clusters basée sur la distance entre les centroïdes. La combinaison s'effectue avec respect de la segmentation des caractères. Seuls les pixels ne connectant pas les caractères pré-existants sont ajoutés.

2.2 Complétion par une approche bas niveau

La deuxième étape qui est décrite dans cet article consiste dans l'analyse de chaque objet résultant de la méthode précédente. Ces objets sont soit des caractères uniques dans les images les moins dégradés, soit des caractères cassés dans de rares cas, soit enfin des caractères collés ce qui est l'erreur la plus courante. Dans cet article, nous traitons le problème des objets contenant des caractères collés tout en essayant de ne pas casser les objets contenant déjà un seul caractère. L'objectif prépondérant pour la suite des étapes est de réduire le taux de fausses alarmes pour ne pas alourdir les traitements à effectuer et dégrader l'image.

Nous avons besoin de localiser les informations à la fois spatialement afin de savoir où dans l'image se situe une éventuelle coupure entre les caractères et bien sûr fréquentiellement pour obtenir des informations sur les variations des niveaux de gris susceptibles de révéler les séparations entre caractères. Or la meilleure localisation simultanée dans l'espace image et fréquence est effectuée par un filtre de Gabor, qui applique un filtre de type cosinus possédant une certaine direction modulée par une fenêtre gaussienne.

Des études plus approfondies sur les images naturelles [3] ont montré qu'elles possédaient un spectre décroissant en $1/f$. Cela a permis d'introduire une échelle fréquentielle logarithmique au lieu de l'échelle linéaire afin de rendre compte de ce phénomène. La composante continue qui apparaît dans les filtres symétriques de Gabor lorsque la bande passante de ceux-ci est trop importante disparaît grâce à cette échelle ce qui donne plus de marge de manoeuvre pour la bande passante des filtres. Le filtre Log-Gabor présente aussi une similitude plus importante avec le système visuel humain qui montre une réponse symétrique sur une échelle de fréquences logarithmiques. Nous utilisons ici l'implémentation faite par Kovesi [5] de la convolution par un filtre Log-Gabor, ce qui nous permet de récupérer la phase φ de la convolution de l'image I. Le paramètre le plus important d'un filtre de Gabor est sa fréquence f_{Gabor}. Nous avons décidé de mettre ce paramètre directement en relation avec l'épaisseur moyenne des caractères. Pour la calculer, nous utilisons simplement le rapport entre le nombre de pixels constituant le squelette des objets de l'image I issus de la méthode MET et la somme de tous les points qui les composent (voir Figure 2-c) tels que :

$$f_{Gabor} = \frac{\sum pixels_{squelette}}{\sum pixels_I} \quad (1)$$

Le calcul des convolutions de deux filtres, l'un horizontal et l'autre vertical, avec l'image, est effectué (voir Figures 2-d et 2-e). La séparation des caractères en unités distinctes est majoritairement verticale ainsi, seul le résultat de la convolution de l'image par le filtre vertical sera pris en compte. Par la suite nous ne considérons plus que la valeur absolue de la phase de ce résultat notée I_φ (voir Figure 2-f).

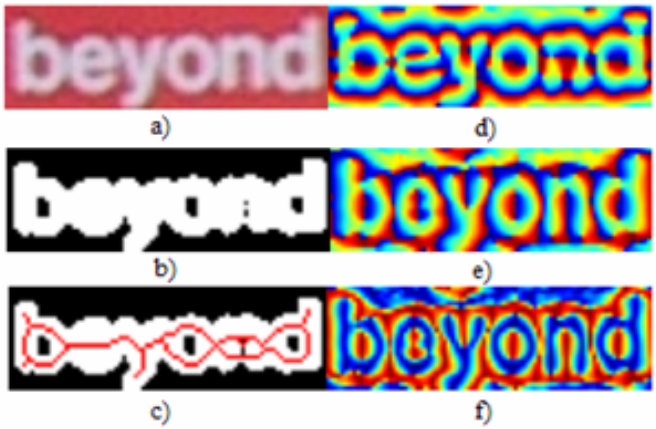

FIG. 2 – a) I, b) masque M, c) M et son squelette, d) et e) phases des filtres horizontal et vertical, f) valeur absolue de la phase e) I_φ

Enfin, nous multiplions le masque M par le résultat obtenu précédemment pour obtenir une image I_{masque} (voir Figure 3-a). Nous effectuons ensuite un seuillage S basé sur la méthode classique d'Otsu [7] qui va maximiser la variance inter-classes afin de trouver automatiquement ce seuil. Enfin, nous enlevons cette information inter-caractères éventuelle au masque M pour obtenir l'image finale I_{finale} par $I_{finale} = M - (I_{masque} > S)$ comme sur Figure 3-b.

FIG. 3 – a) I_{masque}, b) I_{finale}

3 Résultats

Sur la base de données d'exemples de mots segmentés [4], et après la méthode MET où 29.2% des images contenaient des caractères collés, l'algorithme de segmentation présenté dans cet article permet de diminuer le nombre de ces images de 76%, pour obtenir alors 7% d'images avec des caractères collés. De plus, seulement 2.3% de fausses alarmes sont constatées permettant de ne pas trop alourdir les traitements sur les caractères cassés dont le nombre n'augmente que légèrement. Un second résultat est illustré dans Figure 4 avec une comparaison avec l'algorithme de seuillage global Otsu [7] qui a également prouvé une grande généricité.

FIG. 4 – a) I, b) M obtenu par binarisation Otsu, c) I_{finale} par la méthode proposée

Sur Figure 5, un autre exemple est montré pour apprécier l'amélioration obtenue par rapport à la première approche MET. Ainsi cette approche en deux étapes permet d'avoir une bonne approximation de la segmentation des caractères.

Les résultats dépendent surtout de la bonne estimation de l'épaisseur des caractères. Celle-ci est, comme le montrent les premiers chiffres, assez satisfaisante dans 97.7% des cas. Elle est principalement erronée dans les images très bruitées où le masque M est imprécis.

FIG. 5 – a) I, b) M obtenu par la méthode MET, c) I_{finale} par la méthode proposée

4 Discussion et Conclusion

Un algorithme de segmentation de caractères collés adapté aux images de scènes naturelles a été présenté ainsi que les premiers résultats assez prometteurs. La méthode basée sur les filtres de Log-Gabor n'est pas gourmande en ressources de calcul puisque seulement deux convolutions (filtres horizontal et vertical) sont effectuées, ce qui permet ainsi un usage dans un environnement embarqué de style PDA.

Dans l'avenir, nous chercherons à améliorer la méthode d'estimation de l'épaisseur des caractères pour réduire encore davantage le taux de fausses alarmes. Dans cet article, un accent particulier a été donné aux caractères collés ainsi nos travaux actuels s'orientent sur la continuité des propriétés décrites dans le cas des caractères cassés même si leur nombre est très inférieur aux caractères collés dans les images de scènes naturelles.

Afin d'améliorer notre algorithme global, nous pouvons envisager deux évolutions principales. La première consisterait dans une étude bas niveau des paires d'éléments connectés issus de la première étape MET au lieu d'analyser les éléments connectés indépendamment les uns des autres. Cette approche permettrait d'éliminer les erreurs dues à une mauvaise segmentation présente dès la création du masque M. Enfin, un deuxième axe permettrait de compléter notre chaîne de segmentation avec un raffinement du résultat final en utilisant le dernier masque pour initialiser des méthodes de contours actifs ou de lignes de partage des eaux basées sur des marqueurs. Ainsi nous obtiendrions une bonne segmentation des caractères, étape fondamentale préalable à l'extraction de caractéristiques pertinentes pour une reconnaissance de caractères. Ces évolutions doivent être prises en compte avec à l'esprit le compromis qualité-temps de calcul. Les résultats seront améliorés mais avec un temps de calcul plus prohibitif. Ainsi l'étude du compromis en regard de l'application pour un usage embarqué est indispensable.

5 Remerciements

Les auteurs remercient le minitère de la Région wallonne en Belgique pour les subventions attribuées pour réaliser ce travail, qui est partie intégrante du projet Sypole.

Références

[1] M. Cannon, J. Hochberg et P. Kelly, *Quality assessment and restoration of typewritten document images*, Int. Jour. Document Analysis and Recognition, vol.2, pp. 80–89, 1999.

[2] R.G. Casey et E. Lecolinet, *A survey of methods and strategies in character segmentation*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.18, n.7, pp. 690–706, 1996.

[3] D.J. Field, *Relations between the statistics of natural images and the response properties of cortical cells*, Jour. of the Optical Society of America, vol.4, pp. 2379–2394, 1987.

[4] Robust Reading Competition Database : Consulté le 15 juin 2005 du site web :
http ://algoval.essex.ac.uk/icdar/RobustWord.html.

[5] P. Kovesi, *Image Features From Phase Congruency*, Videre : A Journal of Computer Vision Research, MIT Press, vol.1, n.3, 1999.

[6] H. Nishida, *Restoring high-resolution text images to improve legibility and OCR accuracy*, Proc. of the Electronic Imaging Conference of the International Society for Optical Imaging, pp. 136–147, 2005.

[7] N. Otsu, *A thresholding selection method from gray-level histogram*, IEEE Trans. on Systems, Man, and Cybernetics, vol.9, pp. 62–66, 1979.

[8] C. Thillou et B. Gosselin, *Color Binarization for Complex Camera-based Images*, Proc. of the Electronic Imaging Conference of the International Society for Optical Imaging, pp. 301–308, 2005.

[9] C. Wolf, J-M. Jolion et F. Chassaing, *Text localization, enhancement and binarization in multimedia documents*, Proc. of Int. Conf. on Pattern Recognition, vol.2, pp. 1037–1040, 2002.