Cascade attentionnelle de classifieurs pour la détection du texte de scène dans les images

S. M. Hanif, L. Prévost, P. A. Negri

Cascade attentionnelle de classifieurs pour la détection du texte de scène dans les images *

Shehzad Muhammad Hanif – Lionel Prevost – Pablo Augusto Negri

UPMC Université Paris 06, F-75005, Paris, France
CNRS, FRE 2507, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France

shehzad.muhammad@lisif.jussieu.fr, lionel.prevost@upmc.fr,
pablo.negri@lisif.jussieu.fr

Résumé : *Dans cet article, nous présentons une méthode de détection et localisation de texte. Notre détecteur est une cascade attentionnelle et le localiseur emploie des méthodes de traitement d'images. Chaque étage de la cascade est entraîné en utilisant l'algorithme AdaBoost. nous avons étudié en détail l'influence des descripteurs utilisés pour coder l'information et des classifieurs mis en œuvre. L'évaluation de notre approche a été fait sur 250 images de la base ICDAR. Nous avons obtenu un taux de détection de 86% et un taux de fausses alarmes de l'ordre de* $3x10^{-3}$*. La cascade est capable de traiter une image 640x480 en moins de 2 secondes.*

Mots-clés : Détection et localisation de texte, cascade attentionnelle, AdaBoost, analyse et reconnaissance de document

1 Introduction

La détection, la localisation et la reconnaissance de texte dans les images sont les étapes principales d'applications destinés aux conducteurs de véhicules [CHE 04b], aux personnes mal-voyantes [EZA 04] ou non-voyantes [CHE 04a]. La communauté DAR (Document Analysis & Recognition) s'intéresse depuis peu à l'analyse des documents capturés à l'aide d'une caméra. Ce nouveau domaine pose beaucoup de challenges : faible résolution, dégradation du texte, fond très complexe etc. Le travail présenté dans cette communication se situe dans le cadre de ce dernier domaine de recherche.

2 Détecteur de texte

2.1 Codage

La recherche dans le domaine de la détection d'objet [VIO 04] a montré que des descripteurs locaux sont efficaces pour modéliser un objet et donnent d'excellents résultats. Dans notre approche, des segments de texte sont extraits des lignes de texte présentes dans une image. Chaque segment contient au mois deux caractères. Il est de taille 32x32 pixels ou son multiple entier. Nous avons considéré huit échelles différentes. Chaque segment est divisé en 16 bloc (voir figure 1). Pour chaque bloc, nous avons extrait trois différents types de descripteurs : la différence de moyenne d'intensité (MDF), l'écart type d'intensité (SD) et l'histogramme de gradient orienté en 8 directions (HoG). Nous avons obtenu 39 descripteurs (7 MDF, 16 SD et 16 HOG). Ces descripteurs peuvent être considérés isolement ou combinés en paires homogènes (SD+SD par exemple) ou hétérogènes (SD+HoG, par exemple). Nous obtenons 780 caractéristiques en considérant les descripteurs isolés et les paires de descripteurs, 9919 caractéristiques en ajoutant des triplets.

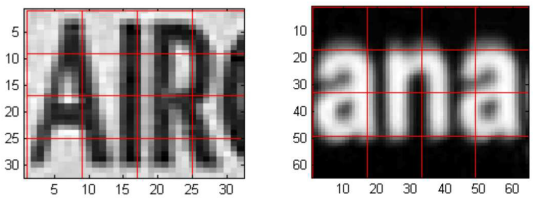

FIG. 1 – Division en blocs d'un segment de texte

2.2 Adaboost et la cascade attentionnelle

L'algorithme AdaBoost [FRE 96] combine un ensemble de classifieurs « faibles » (un peu meilleur que le hasard) en un classifieur « fort » performant. Dans les espaces de représentation de grande dimension définis précédemment, un classifieur faible est associé à chaque caractéristique. L'algorithme AdaBoost construit le classifieur en sélectionnant séquentiellement les clasifieurs faibles les plus performants, jusqu'à remplir un objectif fixé (taux de détection minimum, par exemple). Nous avons employé deux classifieurs « faible » : un classifieur discriminant linéaire (LD) et un test du rapport de vraisemblance (LRT). Pour ce dernier, les densités de probabilité des classes « texte » et « non-texte » sont supposées gaussiennes. Leurs paramètres (moyenne et matrice de covariance) sont estimés en maximisant la vraisemblance sur l'ensemble d'apprentissage.

Un problème complexe tel que la détection de texte nécessite un très grand nombre de caractéristiques et autant de classifieurs faibles, conduisant à des temps de traitements prohibitifs. Pour réduire la charge de calcul, [VIO 04] ont proposé de construire une cascade attentionnelle où chaque étage est un classifieur « fort ». L'idée générale de la cascade est de choisir un nombre faible de caractéristiques dans les premiers étages, qui rejetteront la plupart des zones de non-texte. Les cas difficiles ne sont traités que dans les derniers étages, en s'appuyant sur un grand nombre de caracté-

*Vous pouvez trouver la version détaillée de cet article sur http://isir.robot.jussieu.fr/?op=view_profil&lang=fr&id=73

ristiques.

2.3 Localisation

Nous combinons d'abord toutes les détections en utilisant une mesure de densité de contours par région. Plusieurs composantes connexes sont extraites, pouvant contenir des lignes de texte ou des mots. Puis, pour chaque composante connexe, nous employons des opérations morphologique sur l'image de contour obtenue par filtrage de Canny. Avec ces opérations, nous obtenons des mots ou des grands caractères sous la forme des composantes connexes. Des contraintes géométriques et spatiales, l'analyse des densités de contours par ligne, du profil etc. permettent de valider et de combiner les composantes connexes. Enfin, nous obtenons des boîtes englobantes autour des mots. Les seuils utilisés durant cette étape sont estimés sur la base d'apprentissage.

3 Résultats

Pour évaluer notre approche, nous avons utilisé la base ICDAR 2003 [ICD 03]. Les bases d'apprentissage et de test contiennent chacune 250 images. Le texte dans cette base varie selon la police, la taille, le style d'écriture et l'apparence.

D'abord, nous faisons des tests sur la combinaison des descripteurs et la sélection de classifieurs faibles. Nous avons entraîné trois classifieurs forts comprenant 200 classifieurs et utilisant des combinaisons de deux et trois descripteurs. La base de test contient 492 exemples de texte et 5378 exemples de non-texte. Nous constatons (tableau 1) que des combinaisons de deux descripteurs sont plus performantes que celle de trois descripteurs et que le test du rapport de vraisemblance est plus efficace (moins « faible ») que le classifieur discriminant linéaire . Les mesures de performance sont le taux de détection (TD) et le taux de fausses alarmes (TFA).

Type de classifieur	# total de classifieurs faible	TD	TFA
LD	780	92,9%	0,8%
LD	9919	91,7%	0,8%
LRT	780	**94,9**%	**0,6**%

TAB. 1 – Influence du choix des descripteurs et des classifieurs faibles sur les performances

Ensuite, nous avons comparé un classifieur fort comprenant 250 classifieurs et une cascade de 7 étage en utilisant le test du rapport de vraisemblance. La base de test contient 3842 exemples de test et 125000 exemples de non-texte. Les résultats (tableau 2) montrent que la cascade est plus performante que le classifieur simple au niveau des fausses alarmes et du temps de calcul (notons la légère baisse du taux de détection due au plus grand nombre d'exemples).

Type de classifieur	# Classifieurs faible	TD	TFA	Temps (secondes)
Simple	250	90,6%	1,1%	4
Cascade	266	86,6%	**0,3**%	**1,2**

TAB. 2 – Performances des détecteurs simple et en cascade

Finalement, la cascade et le localiseur sont évalués sur la base de test de 250 images. Le rappel et la précision sont utilisés comme des mesures de performance. Le critère ICDAR (critère 1) s'appuie sur le recouvrement entre le rectangle détecté et la vérité terrain. Le critère 2 est une méthode pixelique. Le tableau 3 détaille les résultats obtenus sur la base de test. Quelques exemples de détection sont montrés dans la figure 2. En moyenne, nous avons 2 fausses alarmes par images et des régions plus confondues sont les arbres, les bâtiments et des structures verticales.

FIG. 2 – Exemples de détection et localisation de texte

	Rappel	Précision
Critère 1	36,2%	24,0%
Critère 2	79,2%	37,3%

TAB. 3 – Rappel et précision sur la base de test

Références

[CHE 04a] CHEN X., YUILLE A. L., Detecting and Reading Text in Natural Scenes, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 366-373, 2004.

[CHE 04b] CHEN X., YANG J., ZHANG J., WEIBEL A., Automatic Detection and Recognition of Signs from Natural Scenes, *IEEE Transactions on Image Processing*, pp. 87-99, 2004.

[EZA 04] EZAKI N., BULACU M., SCHOMAKER L., Text Detection from Natural Scene Images : Towards a System for Visually Impaired Persons, *17th International Conference on Pattern Recognition (ICPR)*, vol. 2, pp. 683-686, 2004.

[FRE 96] FREUND Y., SCHAPIRE R., Experiments with a New Boosting Algorithm, *International Conference on Machine Learning*, pp. 148-156, 1996.

[ICD 03] ICDAR, ICDAR 2003 robust reading and locating database, *http ://algoval.essex.ac.uk/icdar/RobustReading.html*, 2003.

[VIO 04] VIOLA P., JONES M. J., Robust Real-Time Face Detection, *International Journal of Computer Vision*, vol. 57, pp. 137-154, 2004.