

Détection de texte dans des vidéos documentaires avec une cascade de classificateurs

Marc Lalonde – Langis Gagnon

Département de R-D, Centre de recherche informatique de Montréal (CRIM)

550 Sherbrooke Ouest, Suite 100, Montréal, Québec, Canada, H3A 1B9

`marc.lalonde@crim.ca`, `langis.gagnon@crim.ca`

Résumé : *Nous présentons les premiers résultats concernant le développement d'un module de détection de texte dans des films documentaires. L'approche suivie est basée sur l'utilisation d'une cascade de classificateurs entraînée par Adaboost : l'entraînement se fait à l'aide d'un large ensemble de caractéristiques et des tables de conversion jouent le rôle de classificateurs faibles. Caractérisé par une architecture simple, le module localise différents types de texte avec un taux de détection de 95% et un taux de fausses détections très faible (couvrant en moyenne 1% de l'image) et ce, malgré la complexité du contenu visuel des images clés tirées des vidéos documentaires analysés.*

Mots-clés : Détection de texte, Adaboost, indexation vidéo, vidéo descriptive, système multimedia

1 Introduction

Nous présentons une approche de localisation de texte dans des vidéos basée sur l'utilisation d'une cascade de classificateurs. Les vidéos sont des films documentaires avec un contenu visuel complexe et non contrôlé. La motivation première de ce travail est une contribution au développement d'un système d'assistance à la génération automatique de descriptions vidéo pour les mal-voyants ([GAG 06]). Un film renferme souvent des mots-clés qui s'avèrent importants pour le suivi et la compréhension de l'histoire (par exemple, un plan rapproché sur du texte, un texte de transition explicatif entre deux scènes, des sous-titres, des titres de pages de journaux, des noms de rues, etc.). La vidéo descriptive, également connue sous le nom d'audiovision, est une narration additionnelle ajoutée à la bande audio d'un film qui décrit quelques éléments visuels pour aider les malvoyants à mieux apprécier leur expérience d'écoute. L'automatisation d'un tel processus est une application particulière du domaine de l'indexation du contenu visuel.

Localiser n'importe quel type de texte ou partie de texte dans des environnements non-contrôlés comme un film est encore un domaine de recherche très actuel en indexation du contenu visuel ([FUR 04]). Bien que beaucoup de systèmes existants (académiques ou commerciaux) se concentrent toujours sur l'extraction des sous-titres seulement, des progrès importantes sur la détection sans contrainte du texte dans la vidéo ont été réalisés au cours des dix dernières années (par exemple [LIE 02], [ANT 01]). Nous proposons ici une variante de l'approche de Chen et Yuille [CHE 04], toujours basée sur l'utilisation d'une cascade de classificateurs mais appliquée à l'analyse vidéo. En particulier, nous n'imposons aucune contraintes au classificateur quant au choix des caractéristiques optimales. Les classificateurs faibles sont aussi de simple tables de conversion.

L'article est divisé comme suit. La section 2 donne un survol de l'approche utilisée, en particulier sur l'étape d'apprentissage et les caractéristiques utilisées. La section 3 présente différents résultats obtenus sur des films documentaires de l'Office National du Film (ONF) du Canada, avec un mesure de performance différente de celle utilisée précédemment ([LAL 06]). Nous concluons finalement sur les résultats obtenus jusqu'ici et la suite des travaux.

Ce travail fait partie du thème de recherche "Interaction et extraction du contenu audio-visuel" du nouveau Réseau de recherche canadien E-inclusion [CRI 06]. L'objectif du Réseau E-Inclusion est d'explorer et de développer des outils audio-vidéo pour améliorer la richesse de l'expérience multimédia chez les personnes ayant une déficience sensorielle. En particulier, (1) l'adaptation du sous-titrage pour les malentendants, en fonction de l'activité visuelle de la scène et de la densité de la narration et (2) le développement d'outils d'extraction automatique de contenu audio-visuel pour la génération assistée par ordinateur de descriptions vidéo pour les malvoyants, et la recherche par le contenu dans des archives de films.

2 Approche

La recherche de texte dans une trame video se fait en balayant l'image par une fenêtre d'analyse dans laquelle on extrait des caractéristiques de l'image. Une cascade de classificateurs est utilisée pour assigner la zone analysée à la classe 'texte' ou 'non-texte'. Cette cascade consiste en une serie de classificateurs entraînés de façon à ce que les premiers, de faible complexité, soient capables de rejeter une grande proportion de mauvais candidats (fausses alarmes) tout en s'assurant que les véritables zones de texte soient correctement classifiées. Les classificateurs en aval de la cascade sont généralement plus complexes puisqu'ils font face à des cas moins triviaux (mais peu nombreux). La figure 1 illustre le fonctionnement de la cascade.

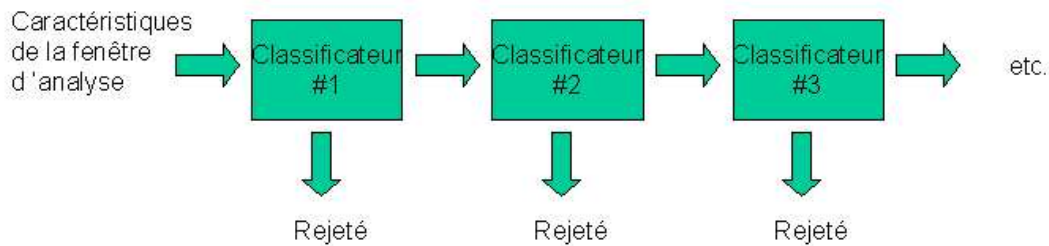


FIG. 1 – Cascade de classificateurs

2.1 Apprentissage

L'algorithme d'apprentissage Adaboost est utilisé pour construire chaque classificateur de la cascade à partir de la contribution de plusieurs classificateurs faibles (*weak classifiers*). Comme la stratégie usuelle est d'utiliser un classificateur faible pour chaque dimension du vecteur de caractéristiques, Adaboost se trouve à faire de la sélection de caractéristiques en recherchant les meilleurs classificateurs faibles. Pendant l'entraînement, pour chaque classificateur C_i de la cascade, on entraîne N classificateurs faibles selon les N dimensions des vecteurs de caractéristiques des données et on retient le plus performant (auquel on rattache un poids qui est inversement proportionnel à l'erreur de classification) ; si cette erreur est trop élevée, on entraîne à nouveau N classificateurs faibles pour choisir le meilleur (caractéristique d'Adaboost : les exemples d'entraînement bien classifiés par le ou les classificateurs faibles précédents ont moins d'influence dans ce nouveau calcul d'erreur). En bout de ligne, la décision que rend C_i est la combinaison pondérée des décisions de ses classificateurs faibles, qui ont été ajoutés en nombre suffisant pour que l'erreur de classification de C_i (faux négatifs et faux positifs) respecte des critères préétablis. De la même façon, les classificateurs C_i sont ajoutés à la cascade jusqu'à ce que l'erreur de classification de la cascade soit inférieure à une valeur désirée. Le type de classificateur faible choisi pour cette étude est basé sur une simple paire de tables de conversion représentées par des histogrammes normalisés H_p et H_n peuplés avec les valeurs de la caractéristique analysée tirées des exemples positifs et négatifs respectivement ; on prend une décision en comparant les colonnes correspondant à la valeur de la caractéristique utilisée : si la colonne de H_p est plus importante que celle de H_n , alors la 'forme' est assignée à la classe 'positif'.

2.2 Caractéristiques

Chaque vecteur de caractéristiques est construit à l'aide de mesures prises dans la fenêtre d'analyse qui balaie l'image. Contrairement à [CHE 04] qui orientent en partie la construction du vecteur, on divise cette

A
B
C

FIG. 2 – Caractéristiques extraites d’une fenêtre de trois blocs : $[\mu_A, \sigma_A, \mu_B, \sigma_B, \mu_C, \sigma_C, \mu_A - (\mu_B + \mu_C), \sigma_A - (\sigma_B + \sigma_C), \mu_B - (\mu_A + \mu_C), \sigma_B - (\sigma_A + \sigma_C), \mu_C - (\mu_B + \mu_A), \sigma_C - (\sigma_B + \sigma_A)]$, et ce, sur les images en niveaux de gris à variance normalisée et les images de dérivée en X et Y.

fenêtre en blocs verticaux à l’intérieur desquels on calcule une moyenne et un écart-type pour l’image en niveaux de gris, mais aussi pour les images de dérivées en X et Y correspondantes (approche similaire à [DLA 05]). Les caractéristiques retenues sont ces mesures ainsi que les variations de ces mesures entre les blocs (voir la figure 2). Le nombre de blocs varie de deux à cinq, et un vecteur de caractéristiques final compte alors 240 éléments.

L’entraînement s’est fait avec 3411 exemples positifs (tirés de la banque d’entraînement fournie pour un concours de recherche de texte dans le cadre de la conférence ICDAR 2003) et 12243 exemples négatifs tirés d’images de scènes naturelles. L’architecture finale de la cascade comprend trois classificateurs formés respectivement de 4, 5 et 4 classificateurs faibles. La figure 3 illustre les caractéristiques correspondant aux classificateurs faibles de chaque étage de la cascade.

3 Résultats et discussion

L’ensemble de test est formé de 147 images clés (*keyframes*) provenant de courts extraits de 22 films documentaires de l’Office national du film du Canada. Les images clés sont extraites manuellement. Le contenu des films est très varié : scènes urbaines et naturelles, en couleur et en noir et blanc, de qualité variable selon l’année de production. On y retrouve du texte sous différentes formes : générique, sous-titres, images de unes de journaux pour certains documentaires historiques, affiches commerciales, etc. En tout, 443 chaînes de texte constituent la vérité terrain.

Afin de tenir compte de la dimension variable du texte, trois tailles de fenêtres d’analyse ont été utilisées : 40x20 pixels, 80x40 pixels et 120x60 pixels pour des images de taille 640x480. Le temps de traitement d’une trame vidéo est d’environ 1 seconde sur un ordinateur de bureau (code C++ avec optimisations de base, p. ex. utilisation d’images intégrales [VIO 01]). La figure 4 donne des exemples de détection. Les exemples de la figure 5 sont également intéressants puisqu’ils contiennent du texte difficile à détecter.

L’évaluation plus formelle d’un algorithme de détection de texte n’est pas simple, comme le mentionne

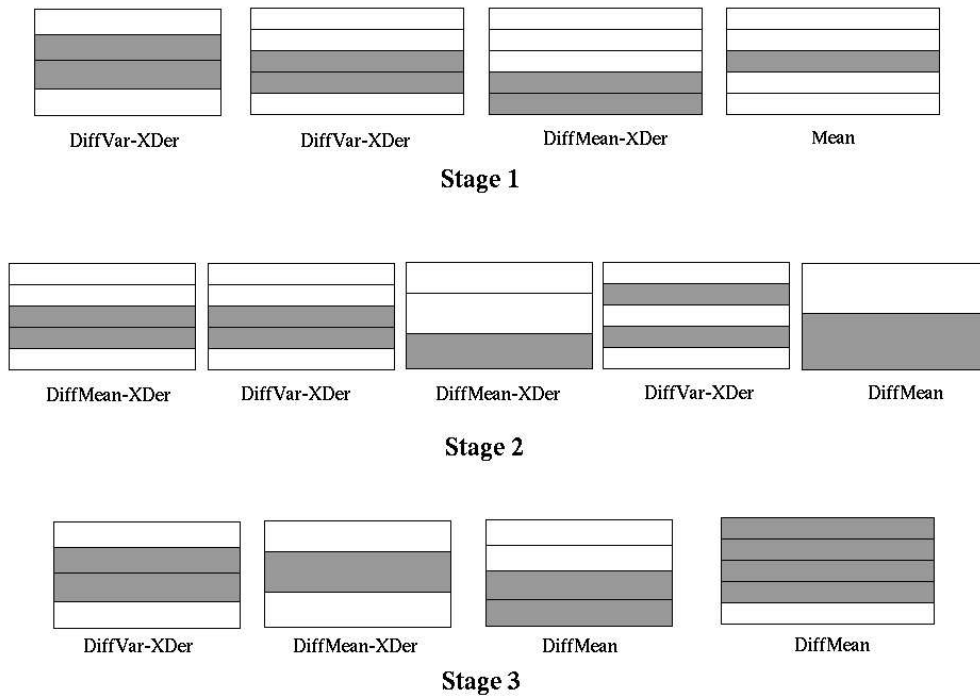
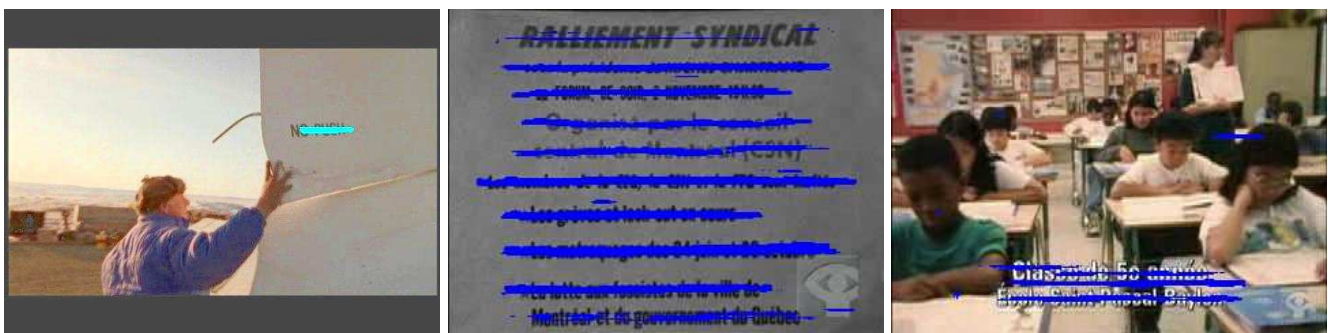


FIG. 3 – Meilleures caractéristiques retenues pour chacun des classificateurs de la cascade.



(a) Texte dans la scène

(b) Image d'affiche

(c) Sous-titres

FIG. 4 – Quelques résultats. Les pixels bleus représentent les centroïdes des zones de texte trouvées.



(a) bannière ISO9001 d'un véhicule en mouvement

(b) Affiche 'À louer'

(c) Ligne d'autobus : police de caractères atypique

FIG. 5 – Résultats intéressants dans des situations difficiles.

[ANT 01]. Une approche courante veut qu'on calcule une intersection de surfaces entre les zones de texte trouvées et la vérité terrain : plus leur surface commune est grande, plus l'erreur de détection est petite. Cette définition d'erreur fait peu de sens dans le contexte actuel avec une fenêtre d'analyse fixe : quelques fausses alarmes vont couvrir une portion significative de l'image, ce qui ne rend pas justice à l'algorithme. On opte plutôt pour la démarche suivante :

- pour chacune des trois échelles k , une image de 'succès' I_H^k est construite : si le contenu de la fenêtre d'analyse est jugé comme étant du texte, le pixel correspondant au centroïde de la fenêtre est marqué ;
- un léger filtre morphologique est appliqué aux images I_H^k : ouverture avec élément structurant rectangulaire 2×3 ;
- on crée, à partir de ces images, la liste des zones de texte potentielles et on fusionne celles qui se chevauchent à plus de 90% ;
- la comparaison avec la vérité terrain se fait selon les critères suivants : couverture de la vérité terrain, nombre de fausses alarmes nettes (zones sans chevauchement avec la vérité terrain) et proportion des zones qui chevauchent la vérité terrain.

Pour l'ensemble des 147 images de la banque de test, la vérité terrain est couverte en moyenne à 95.3%, ce qui correspond à un taux de détection élevé (peu de texte est manqué par le détecteur). La figure 6 donne deux exemples d'images pour lesquels le détecteur a trouvé peu ou pas de texte : l'image de gauche est de mauvaise qualité et celle de droite inclut des caractères écrits au trait fin dont les caractéristiques sont probablement très différentes de celles des caractères de l'ensemble d'apprentissage.

Le nombre de fausses alarmes nettes est faible : en moyenne, par image, environ deux zones de taille



FIG. 6 – Exemples d’images dont le texte est mal détecté



FIG. 7 – Étirement des zones de texte trouvées à cause de la sensibilité du détecteur. Le phénomène est particulièrement évident pour l’image de droite.

moyenne (totale) de 1856 pixels ne chevauchent pas la vérité terrain (i.e. ne contiennent pas de texte). Le filtrage morphologique joue un certain rôle en éliminant les détections isolées puisque sans filtrage, ce nombre monte à trois zones par image (détection de 96.2%). On note cependant la grande dimension des zones de texte trouvées quand on les compare à celles de la vérité terrain, puisque le chevauchement moyen mesuré est de 34% en moyenne. Ceci concorde avec l’inspection visuelle des résultats : on voit que dans beaucoup de situations (voir la figure 7) les zones de texte ont tendance à être étirées par rapport au texte présent. Le phénomène est causé par la grande sensibilité du détecteur de texte, qui donne une réponse positive même dans les cas où la fenêtre n’inclut qu’une partie du texte (une ou deux lettres).

4 Conclusion

Cet article portait sur la détection de texte dans des trames vidéos. Le détecteur proposé, constitué d'une cascade de trois classificateurs entraînés par Adaboost, offre des performances très intéressantes avec un taux de détection de 95%. De plus, le taux de fausses alarmes nettes est très bas (en moyenne, deux mauvaises régions par image couvrant 1856 pixels, soit moins de 1% de l'image). Le seul inconvénient est la taille des zones trouvées qui sont nettement plus grandes que le texte lui-même. L'étape suivante est naturellement de greffer les sous-modules requis pour la reconnaissance des caractères trouvés (segmentation, OCR). Nous pensons que leur travail sera de beaucoup facilité par le bas taux de fausses alarmes du sous-module de détection. Néanmoins, le détecteur étant capable de trouver du texte de faible dimension dans des trames vidéo à basse résolution, des algorithmes de superrésolution pourraient être utilisés pour extraire le maximum d'information textuelle de ces images. Finalement, afin d'exploiter la redondance temporelle de l'information, un suivi de la détection/reconnaissance sur plusieurs trames est également à considérer.

5 Remerciements

Ce travail est supporté en partie par (1) le Ministère du patrimoine canadien via le programme Culture canadienne en ligne, (2) le Conseil de recherche en science naturelle et Génie (CRSNG) du Canada et (2) le Ministère du Développement Économique de l'Innovation et de l'Exportation (MDEIE) du Gouvernement du Québec. Nous remercions aussi l'ONF de nous avoir donné accès à une partie de leur collection.

Références

- [ANT 01] ANTANI S. K., Reliable extraction of text from video, PhD thesis, Pennsylvania State University, 2001.
- [CHE 04] CHEN X., YUILLE A. L., Detecting and reading text in natural scenes, *Proc. CVPR 2004*, 2004, pp. 366-373.
- [CRI 06] CRIM, Réseau de recherche E-Inclusion, 2006, <http://e-inclusion.crim.ca/?q=fr>.
- [DLA 05] DLAGNEKOV L., Video-based Car Surveillance : License Plate, Make, and Model Recognition, Master's thesis, UCSD, mai 2005.
- [FUR 04] FURHT B., MARQUES O., Eds., *Handbook of Video Databases - Design and Applications*, CRC Press, 2004.

- [GAG 06] GAGNON L., FOUCHER S., LALIBERTE F., LALONDE M., BEAULIEU M., Toward an Application of Content-Based Video Indexing to Computer-Assisted Descriptive Video, *Proc. Canadian Conference on Computer and Robot Vision*, 2006, Submitted.
- [LAL 06] LALONDE M., GAGNON L., Key-text spotting in documentary videos using Adaboost, SPIE, Ed., *Proc. Electronic Imaging 2006*, 2006.
- [LIE 02] LIENHART R., WERNICKE A., Localizing and Segmenting Text in Images and Videos, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, n° 4, 2002, pp. 256-268.
- [VIO 01] VIOLA P., JONES M., Robust real-time object detection, *Proc. of IEEE workshop on Statistical and Computational Theories of Vision*, 2001.