Détection et reconnaissance de texte dans les documents vidéos

Et leurs apports à la reconnaissance de personnes

Johann Poignant

UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041 France

Johann.Poignant@imag.fr

RÉSUMÉ. *Cet article présente les différentes étapes de reconnaissance des caractères dans un système de reconnaissance multimodale de personnes dans des documents audiovisuels (défi ANR REPERE). La détection du texte est réalisée par une technique basée sur les caractéristiques du texte (texture, couleur, contraste, géométrie, suivi temporel, mesure du gradient cumulé). La reconnaissance du texte est ensuite effectuée avec le logiciel libre de* Google Tesseract. *La méthode a été évaluée sur un corpus de journal télévisé contenant 59 vidéos du journaux télévisés de France 2.*

ABSTRACT. *This article presents the different steps used to recognize characters for multi-modal person recognition systems in video (ANR REPRE challenge). Text detection is achieved by a technique based on the text features (texture, color, contrast, geometry, temporal information, measure of accumulated gradients). The text recognition is then performed by the free software* Google Tesseract. *The method was evaluated on a TV news corpus containing 59 videos from the France 2 French TV channel.*

MOTS-CLÉS : *Détection des textes, Vidéo Optical Character Recognition (OCR), reconnaissance de personnes*

KEYWORDS: *Text detection, OCR video, person recognition*

1. Introduction

Le défi REPERE[1] organisé par l'ANR propose de réaliser un système intégré de reconnaissance de personnes dans des émissions audiovisuelles, en s'appuyant sur les différentes sources d'information présentes dans ces émissions. La reconnaissance des personnes dans des documents vidéos a d'abord été un problème mono-modal (reconnaissance faciale, reconnaissance du locuteur) pour devenir un problème multi-modal en utilisant la fusion de ces deux principales modalités.
L'utilisation d'autres modalités, disponibles dans les vidéos, pourrait améliorer la qualité des systèmes de reconnaissance de personnes. Le texte incrusté dans une vidéo peut apporter des éléments informatifs quant à la présence ou à la citation d'une personne. Par exemple, le nom de la personne, sa fonction, un lieu où elle s'est rendue récemment, etc. La connaissance de ce texte peut permettre de désambiguïser un nom qui a été cité ou inversement. Elle peut permettre de comprendre le contexte de la vidéo, ou le sujet principal d'un reportage, et ainsi améliorer les systèmes de fusion multimodale de reconnaissance de personnes.
Plusieurs étapes sont nécessaires avant de pouvoir utiliser ces informations. En premier lieu, une détection du texte doit être faîte avec précision ; ensuite une reconnaissance du texte est effectuée à partir des images extraites lors de l'étape précédente. Un post-traitement peut-être appliqué pour améliorer la qualité de la transcription.
Ce papier présente une étude préliminaire à l'utilisation de la modalité texte pour la reconnaissance de personnes dans des documents audio-visuels. Il est organisé de la manière suivante : la section travaux connexes présente un tour d'horizon de l'état de l'art de l'OCR dans les vidéos. La section suivante présente notre système d'OCR. Suit une présentation de l'évaluation et des résultats connexes. Enfin la dernière section est dédiée aux conclusions et perspectives.

2. Travaux connexes

Les premières recherches dans ce domaine ont été effectuées par Lienhart (1996) qui propose une méthode basée sur la couleur, la texture, le contraste et la géométrie du texte. Depuis, de nombreuses méthodes, plus ou moins complexes, sont apparues. Parmi les différentes techniques, on peut citer (Smith *et al.,* 1997) et (Hua *et al.,* 2001) qui se sont concentrés sur la détection de densité de coins présents dans l'image. Par ailleurs, (Li *et al.,* 2002) utilise des méthodes d'apprentissage pour la détection du texte. Enfin, on trouve dans (Wolf *et al.,* 2001) un schéma de détection s'appuyant sur la mesure du gradient directionnel cumulé.
Deux solutions sont à envisager pour résoudre la problématique de reconnaissance du texte : utiliser un logiciel d'OCR classique à qui l'on envoie des images de tailles adaptées, ou produire un système OCR dédié à la vidéo. Ce dernier peut-être divisé en trois parties : 1) la segmentation des composantes d'un bloc texte ((Kahan *et al.,* 1987), (Tsujimoto *et al.,* 1991) (Casey *et al.,* 1982)), 2) la reconnaissance des caractères ba-

1. REconnaissance de PERsonnes dans des Emissions audiovisuelles

sée sur les caractéristiques des caractères ((Belaid *et al.,* 1994), (Somol *et al.,* 1999)) ou sur des méthodes d'apprentissage (KNN[2] pour (Belaid *et al.,* 1992), HMM[3] pour (Belaid *et al.,* 1994), MLP[4] et cartes auto-organisatrices de Kohonen pour (Lecun, 1989) , SVM[5] pour (Lecun *et al.,* 1998) et (Chatelain, 2006)) et enfin 3) la correction des erreurs qui commence par l'utilisation de la distance d'édition de chaînes Levenshtein (1966) puis une de ses extensions (Wagner *et al.,* 1974). On trouve encore des post-traitements lexicaux et linguistiques (Takeuchi *et al.,* 2000).
(Zhao *et al.,* 2006) proposent une fusion multimodale pour la reconnaissance des personnes à partir de système de reconnaissance faciale, de reconnaissance locuteur, d'OCR et d'ASR[6]. En plus de ces modalités, ils ont utilisé des ressources externes (un corpus de texte AQUAINT, des sites de news et des moteurs de recherche) pour améliorer la reconnaissance. La fusion est basée sur l'algorithme RANKBOOST (Freund *et al.,* 1995).

3. Notre approche

Nous avons choisi de ne détecter que les textes fixes, écrits horizontalement. Une technique basée sur la texture, la couleur, le contraste, la géométrie du texte ainsi que la mesure du gradient directionnel est suffisante pour la détection. La reconnaissance du texte quant a elle sera effectuée par le logiciel Tesseract à qui l'on transmettra des images de résolutions adaptées.
Les différentes opérations pour la détection sont ordonnées comme suit (Fig.1) : A partir d'une trame (Fig.1a), un filtre Sobel[7] est appliqué pour détecter les bords des caractères que l'on seuille (Fig.1b). Ensuite, un traitement de dilatation et d'érosion permet de connecter les caractères entre eux (Fig.1c). Après suppression du bruit (Fig.1d), une détection des composantes connexes nous donne les coordonnées des boîtes rectangulaires comprenant le texte d'une ligne. Un filtrage sur la géométrique est effectué sur ces boîtes ainsi qu'un affinage des coordonnées des boîtes en fonction de la casse du texte (Fig.1e).
La détection est effectuée sur chacune des trames ; ce qui nous permettra d'avoir un suivi temporel pour connaître la trame d'apparition d'une boîte de texte et sa trame de disparition. Les boîtes qui ne sont pas suffisamment stables dans le temps, par exemple les zones de l'image détectées qui auraient une texture proche du texte mais se déplaçant dans l'image (mouvement de la caméra), sont ainsi supprimées.
Pour lisser les images de fond pouvant apparaitre derrière un texte en surimpression (texte sans fond uni), nous utilisons une image moyenne correspondant à la moyenne des 10 dernières trames (Fig.2a). La région d'intérêt est ensuite extraite selon les coordonnées calculées précédemment.

2. K-Nearest Neighbor
3. Hidden markov model
4. Multi-Layer Perceptron
5. Support Vector Machine
6. Automatic Speech Recognition
7. L'agorithme Sobel est un opérateur utilisé en traitement d'image pour la détection de contours

(a) Image originale (b) Detection des bords (c) Connection des caractères (d) Suppression du bruits (e) Image avec un masque

Figure 1. *Images extraites du journal de France 2 du 1er Février 2007, source INA*

Une interpolation bi-cubique est effectuée sur l'image moyenne de la boîte pour obtenir un texte de résolution classique utilisée pour un système OCR sur des pages scannées. Puis grâce à la méthode Otsu[8], un seuillage est déterminé pour obtenir une image binarisée (Fig.2b). Cette image est ensuite envoyée au logiciel Tesseract.
La reconnaissance du texte est effectuée toutes les 10 trames entre la trame d'apparition du texte et sa trame de disparition. Ces multiples détections nous permettent d'avoir plusieurs possibilités de retranscription de chacun des textes. Ces multiples possibilités sont fusionnées pour obtenir la transcription la plus probable.

(a) (b)

Figure 2. *Images moyennes et binarisées*

4. Évaluation

Le corpus utilisé est constitué de 59 vidéos du JT de france 2 du 1er février au 31 mars 2007. La durée moyenne de ces vidéos est d'environ 38 minutes, soit 37 heures de vidéos. 29019 images clés ont été extraites par segmentation. Les textes de ces images clés ont été annotés manuellement ainsi que la présence d'un nom de personne écrit à l'écran et la présence (à priori) de cette personne à l'écran.
La détection du texte n'a pas été évaluée pour sa position spatiale (qui demande une annotation précise) mais par la détection de la présence des boîtes dans les trames annotées. Notre système a un rappel de 0.85. Ce résultat un peu faible est dû au paramétrage manuel de notre système, une annotation précise des coordonnées des boîtes

8. Méthode de seuillage automatique à partir de la forme de l'histogramme d'une image

Type	boites	mots	caractères	err rate mots	err rate caractères
TC	9257	30912	154941	25.7%	15.2%
NS	1440	3230	19248	7.8%	3.7%
NM	1683	4263	23984	12.7%	6.5%
NMP	1491	3566	20484	10.2%	5.2%

Tableau 1. *Résultat de la reconnaissance des caratères*

TC : Texte complet. NS : Nom apparaissant seul (hors crédit). NM : Nom apparaissant seul ou avec d'autre mots (hors crédit). NMP : NM et personne à priori présente (seul ou accompagnée) dans la vidéo (hors crédit).

de texte devrait nous permettre d'améliorer ce résultat.
Le texte reconnu n'a subi que des post-traitements ad hoc automatiques corrigeant quelques erreurs. Nous avons utilisé comme métrique d'évaluation de la reconnaissance des textes la distance de Levenshtein, donnée par l'outil Sclite, portant sur les mots et sur les caractères. Cette distance est calculée entre une boîte de texte de l'image clé annotée et le texte reconnu sur la boîte correspondante.
Le tableau 1 nous montre que les noms de personnes ont tendance à être mieux reconnu par notre système d'OCR. Ils sont effectivement, au regard des vidéos que nous utilisons, écrits plus gros que la moyenne de l'ensemble des textes. Lorsque la personne est présente dans la vidéo, le système a encore plus de facilité à lire le nom. Cela est dû, peut-être, à une meilleure stabilité de l'image lorsque la personne est présente dans la vidéo, ce qui permettrait au système de mieux reconnaître le texte.

5. Conclusions et perspectives

Ces travaux nous ont permis d'évaluer l'apport que peuvent apporter les noms écrits dans la détection de personnes dans des documents audio-visuels. Par la suite, l'étude d'autres types d'entités nommées (nom de fonction, lieu, ...) liée à l'utilisation d'informations externes pourra encore apporter des informations utiles pour la reconnaissance de personnes.
Dans nos prochains travaux, nous allons aussi ajouter le pilotage du logiciel Tesseract, ce qui nous permettra d'avoir une sortie texte, avec plusieurs possibilités sur chaque caractère, couplée avec des post-traitements lexicaux et linguistiques. La qualité de retranscription devrait être améliorée.

Remerciements

Ce travail a été partiellement réalisé dans le cadre du programme Quaero, financé par OSEO, l'agence française d'État à l'innovation. Ce travail a été partiellement réalisé dans le cadre du projet QCompere financé par l'ANR, l'agence nationale de recherche française.

6. Bibliographie

Belaid A., Belaid Y., *Reconnaissance des formes - méthodes et applications*, Intereditions, 1992.

Belaid A., Saon G., « Use of Stochastic Models in Text Recognition », *KOSEF-CNRS French-South Korean Workshop on Text Recognition*, p. 79-98, 1994.

Casey R., Friedman T., Wong K., « Automatic scaling of digital print fonts », *IBM JRD*, vol. 26, n° 6, p. 657-666, 1982.

Chatelain C., Extraction de séquences numériques dans des documents manuscrits quelconques, PhD thesis, Université de Rouen, Décembre, 2006.

Freund Y., Schapire R. E., « A decision-theoretic generalization of on-line learning and an application to boosting », *Proc. of EuroCOLT'95*, p. 23-37, 1995.

Hua X.-S., rong Chen X., Wenyin L., Zhang H.-J., « Automatic Location of Text in Video Frames », *Proceeding of ACM Multimedia 2001 Workshops – MIR2001*, ACM Press, p. 24-27, 2001.

Kahan A., Pavlidis T., Baird H., « On the Recognition of Printed Characters of any Font and Size », *IEEE Trans. on PAMI*, vol. 9, n° 2, p. 274-288, March, 1987.

Lecun Y., « Generalization and Network Design Strategies », *Connectionism in Perspective*, Elsevier, 1989.

Lecun Y., Bottou L., Bengio Y., Haffner P., « Gradient-Based Learning Applied to Document Recognition », *Proceedings of the IEEE*, vol. 86, n° 11, p. 2278-2324, 1998.

Levenshtein V. I., « Binary codes capable of correcting deletions, insertions, and reversals », *Soviet Physics Doklady*, vol. 10, n° 8, p. 707-710, 1966.

Li H., Doermann D., Kia O., « Automatic text detection and tracking in digital video », *IEEE Transactions on Image Processing*, vol. 9, n° 1, p. 147-156, Janvier, 2002.

Lienhart R., « Automatic text recognition for video indexing », *MULTIMEDIA '96 : Proceedings of the fourth ACM international conference on Multimedia*, p. 11-20, 1996.

Smith S. M., Brady J. M., « SUSAN-A New Approach to Low Level Image Processing », *International Journal of Computer Vision*, vol. 23, n° 1, p. 45-78, 1997.

Somol P., Pudil P., Novovicova J., Paclik P., « Adaptive floating search methods in feature selection », *Pattern Recognition Letters*, vol. 20, n° 11-13, p. 1157-1163, 1999.

Takeuchi K., Matsumoto Y., « Japanese OCR Error Correction Using Stochastic Morphological Analyzer and Probabilistic Word Ngram Model », *International Journal of Computer Processing of Oriental Languages*, vol. 13, n° 1, p. 62-82, Mars, 2000.

Tsujimoto Y., Asada H., « Resolving ambiguity in segmenting touching characters », *Proc. of the 1^{rst} ICDAR*, p. 701-709, 1991.

Wagner R., Fischer M., « The String-to-String Correction Problem », *Journal of ACM*, vol. 21, n° 1, p. 168-173, janvier, 1974.

Wolf C., Jolion J.-M., Chassaing F., « Détection et extraction de texte de la vidéo », *In 7èmes Journées CORESA*, vol. 1, p. 251-258, Nov., 2001.

Zhao M., Neo S.-Y., Goh H.-K., Chua T.-S., « Multi-faceted contextual model for person identification in news video », *Multi Media Modeling*, 2006.