Percol0 - un système multimodal de détection de personnes dans des documents vidéo

Frederic Bechet[1] Remi Auguste[2] Stephane Ayache[1] Delphine Charlet[3]
Geraldine Damnati[3] Benoit Favre[1] Corinne Fredouille[4] Christophe Levy[4]
Georges Linares[4] Jean Martinet[2]
(1) Aix Marseille Université - LIF (2) Universite de Lille, LIFL
(3) France Telecom Orange Labs (4) Universite d'Avignon, LIA
{frederic.bechet,stephane.ayache,benoit.favre}@lif.univ-mrs.fr,
{remi.auguste,jean.martinet}@lifl.fr ,
{delphine.charlet,geraldine.damnati}@orange.com,
{corinne.fredouille,christophe.levy,georges.linares}@univ-avignon.fr

RÉSUMÉ

Identifier et nommer à chaque instant d'une vidéo l'ensemble des personnes présentes à l'image ou s'exprimant dans la bande son fait parti de ces nouveaux outils de fouille de données. D'un point de vue scientifique la reconnaissance de personnes dans des documents audiovisuels est un problème difficile à cause des différentes ambiguïtés que présentent l'audio, la vidéo et leur association. Nous présentons dans cette étude le système PERCOL0, développé dans le cadre du défi REPERE, permettant de détecter la présence de personnes (audible et/ou visuelle) dans des documents vidéo, sans utiliser de modèles de locuteurs a priori.

ABSTRACT

Percol0 - A multimodal person detection system in video documents

The goal of the PERCOL project is to participate to the REPERE multimodal evaluation program by building a consortium combining different scientific fields (audio, text and video) in order to perform person recognition in video documents. The two main scientific challenges we are addressing are firstly multimodal fusion algorithms for automatic person recognition in video broadcast ; and secondly the improvement of information extraction from speech and images thanks to a combine decoding using both modalities to reduce decoding ambiguities.

MOTS-CLÉS : Reconnaissance Automatique de la Parole, Segmentation en locuteurs, reconnaissance de l'écriture, détection de visages.

KEYWORDS: Automatic speech reconnaissance, speaker diarization, Optic Character Recognition, Face Detection.

1 Introduction

L'étude présentée dans ce papier s'incrit dans le cadre du projet PERCOL [1] visant à proposer des outils novateurs d'identification de personnes dans des vidéos intégrant les flux images et

1. Projet PERCOL ANR 2010-CORD-102-01

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 553–560,
Grenoble, 4 au 8 juin 2012.

sons au sein d'une approche globale. L'analyse des différentes sources d'informations disponibles dans un contenu audiovisuel permet de reconnaître des personnes dont on dispose d'un modèle de voix et/ou visage préalable mais permet également d'identifier sans modèle préalable des personnalités dont l'information de l'identité se trouve dans la parole prononcée (la personne est nommée par son interlocuteur ou par le présentateur) ou dans le texte en incrustation (un bandeau contenant le nom de la personne apparaît en même temps que la personne). La difficulté manuelle (du point de vue applicatif et maintenance) de créer des dictionnaires, de les mettre à jour au fil du temps et d'y incorporer des nouvelles personnalités qui apparaissent dans l'actualité rendent l'approche de l'identification de la personne sans modèle proposée particulièrement attrayante. De plus, cette approche permet idéalement la création ou la mise à jour automatique de modèles de voix et de visage, pour permettre d'améliorer les taux de reconnaissance.

Nous présentons dans cette étude le système PERCOL0, développé dans le cadre du défi REPERE [2], permettant de détecter la présence de personnes (audible et/ou visuelle) dans des documents vidéo, sans utiliser de modèles de locuteurs a priori. Ce système est basé sur la coopération de plusieurs processus de traitement des signaux visuel et audio des documents vidéo :
– Audio : Segmentation et regroupement en locuteurs et Reconnaissance Automatique de la Parole ;
– Image : Détection et regroupement de visages et Reconnaissance Automatique de l'Ecriture (OCR Optic Character Recognition).

Comme on peut le voir dans la liste précédente, deux types de traitements comparables sont appliqués aux deux modalités : une phase de détection, segmentation et regroupement par similarité (au niveau de la voix pour la modalité audio, au niveau des visages pour la vidéo) ; une phase d'extraction de contenu (dans la parole pour l'audio et dans le texte incrusté pour l'image). La phase d'extraction de contenus permet de déterminer les noms des locuteurs et des visages potentiels intervenant dans la vidéo. Enfin, un processus de fusion multimodale est chargé de propager ces noms détectés vers les segmentations en visages et locuteurs afin de prendre les décisions finales sur la présence d'une personne dans le document.

2 Segmentation et regroupement d'hypothèses

2.1 Canal Audio : Segmentation et regroupement en locuteurs

Deux systèmes de segmentation et regroupement en locuteurs ont été utilisés dans le système Percol0, l'un suivant une stratégie ascendante, l'autre une stratégie descendante.

Stratégie ascendante Dans ce système la structuration en locuteurs est effectuée en 2 étapes : une étape de segmentation suivie d'un regroupement hierarchique ascendant basés tous deux sur le critère BIC, permet d'obtenir une structuration initiale dans laquelle les clusters sont assez purs et contiennent suffisamment de données pour permettre de modéliser le locuteur par un mélange de gaussiennes. Ensuite, avec une telle modélisation, un processus itératif de segmentation via un décodage de Viterbi, et de regroupement par le critère CLR (Barras *et al.*, 2006) est réalisé. Sur l'ensemble de développement de la phase0 du défi REPERE, en excluant les zones de double-parole de l'évaluation, le Diarization Error Rate obtenu est de 7.2%.

2. http ://www.defi-repere.fr

stratégie descendante Ce système est basé sur une stratégie de type "Top-down", incluant trois étapes distinctes. La première consiste en une détection parole/non parole basée sur un HMM à N états pour lequel les états représentent les événements acoustiques de type : "parole bande large", "parole bande étroite" (téléphone), "parole sur de la musique" et "musique". La seconde étape consiste, à partir des segments étiquetés "Parole", à appliquer une phase de segmentation basée sur l'approche e-hmm (Fredouille et Evans, 2008). Celle-ci permet l'obtention d'une segmentation en locuteurs pour laquelle les segments attribués à un même locuteur sont regroupés sous la même étiquette. Contrairement à la stratégie "Bottom-Up", il s'agit ici d'ajouter un à un les locuteurs détectés dans le signal audio à traiter suivant un processus itératif.

La dernière étape du système de segmentation et regroupement en locuteurs repose sur une étape de resegmentation permettant d'affiner la segmentation et de supprimer les locuteurs considérés comme peu pertinents (critère de durée minimale).

2.2 Canal video : Détection et suivi de visages

Les étapes de détection et de suivi de visages consistent à détecter les occurrences de visages dans les trames successives de la vidéo, et à les regrouper par similarité afin que les groupes constitués contiennent chacun les visages d'une unique personne. L'originalité du traitement des visages mis en place dans le système PERCOL0 réside dans l'exploitation de la dimension temporelle des vidéos, dans le but d'améliorer l'efficacité et la robustesse de système. En effet, le fait que les visages se présentent avec beaucoup de variabilités dans la pose, les conditions d'éclairage et les expressions faciales rend difficile leur détection et reconnaissance par un système automatique (Zhao *et al.*, 2003). L'approche mise en œuvre dans PERCOL0 a pour objectif de s'abstraire au maximum des variabilités, qui constituent une limitation des approches dites *statiques*. Les approches dites *dynamiques* intègrent et exploitent des informations temporelles de la vidéo. Bon nombre des techniques dynamiques existantes sont des généralisations directes des algorithmes de reconnaissance sur images fixes, appliquées indépendamment sur chaque trame, sans prendre en compte l'information temporelle.

L'approche proposée consiste à mettre en correspondance les visages détectés ainsi que la partie de l'image correspondant à leur buste (s'il est visible) à l'aide d'un nouveau descripteur : les histogrammes spatio-temporels (HST) (Auguste *et al.*, 2012). Les HST sont des histogrammes contenant, en plus des données de comptage des pixels dans une vidéo, des informations sur leur position dans l'espace et dans le temps. Il permettent d'obtenir une plus grande précision que les histogrammes de couleurs classiques lors de la mise en correspondance des séquences. Les personnes sont détectées en utilisant le détecteur standard de Viola et Jones. Les détections ainsi obtenues permettent d'initialiser l'algorithme GrabCut paramétré pour détourer les bustes des différentes personnes de la trame. Un algorithme dédié, principalement basé sur des critères géométriques, décide de la correspondance des bustes entre chaque trame, une courte sous-séquence est ainsi crée incrémentalement pour chaque personne détourée. Les HST construits pour ces sous-séquences sont utilisés comme des signatures discriminantes, mises en correspondance via une mesure de similarité *ad hoc* (Auguste *et al.*, 2012). Le résultat de ces processus est le regroupement par similarité de l'ensemble des occurrences des personnes dans la vidéo, chaque groupe contenant idéalement une unique personne.

3 Extraction d'information

3.1 Canal audio : Reconnaissance Automatique de la Parole

3.1.1 Transcription automatique et extraction d'entités nommées

Le système de transcription automatique utilisé est très proche de celui qui a été engagé dans la campagne d'évaluation ESTER 2008. Le decodage comporte 2 étapes principales. La première réalise une segmentation du signal, dans laquelle les parties parlées sont extraite, la largeur de bande identifiée, puis la segmentation et le regroupement en locuteur réalisé. La seconde phase est la transcription à proprement parlée. Elle même est réalisée en plusieurs passes : un décodage rapide (3xRT, (Linarès *et al.*, 2007)) en 4-grammes, qui permet l'adaptation au locuteur des modèles acoustiques ; un décodage qui utilise ces modèles adpatés et produit des treillis de mots et enfin un décodage après transformation des treillis en réseaux de confusion.

Le système d'extraction d'entités nommées LIA_NEutilisé dans cette étude est décrit dans Bechet et Charton (2010). Il est basé sur une approche mixte utilisant tout d'abord un processus génératif à base de HMM pour prédire une étiquette syntaxique (POS) et sémantique à chaque mot d'un texte ; ensuite un processus discriminant à base de CRF est utilisé pour trouver les bornes et le type complet de chaque entité en utilisant le modèle *Begin-Inside-Outside* (BIO) pour représenter la position de chaque mot à l'intérieur ou à l'extérieur des entités. Les modèles HMM et CRF du système ayant été utilisé dans Percol0 ont été appris sur le corpus ESTER2.

3.2 Canal Vidéo : Reconnaissance Automatique de l'Ecriture

3.2.1 Présentation générale de l'approche

La reconnaissance automatique des textes incrustés dans les vidéos (VOCR) est réalisée à l'aide d'un processus de type multi-trame qui se décompose en quatre étapes : Détection des zones de texte à chaque trame, Reconnaissance des caractères à chaque trame, Suivi des zones de texte, Post-traitement des zones de texte reconnues successives.

La détection des zones de textes repose sur une approche de type réseau de neurone convolutionnel (Delakis et Garcia, 2008), appliquée aux pixels bruts, sans pré-traitement chromatique. La reconnaissance des caractères à proprement parler dans les zones préalablement détectées est réalisée à l'aide du système GOCR[3]. Le suivi des zones de textes repose sur des critères de position et de dimension des boîtes de texte détectées dans des trames successives. Chaque boîte détectée a une forme rectangulaire et est représentée par 4 coordonnées (abscisse X et ordonnée Y du sommet en haut et à gauche, largeur et hauteur). Deux boîtes détectées à deux trames successives sont considérées comme correspondant au même texte si les 4 coordonnées sont similaires. Une tolérance de 10 points est accordée pour X, Y et pour la hauteur et une tolérance de 15 points est acceptée pour la largeur (les boîtes ont généralement des marges plus importantes en largeur). En pratique la recherche de texte n'est pas réalisée dans l'ensemble de l'image mais dans des zones prédéfinies pour chaque type d'émissions. En effet, dans la mesure où nous nous intéressons particulièrement à la reconnaissance des noms de personnes incrustés

3. http ://www.jocr.sourceforge.net

pour présenter les personnes présentes dans la vidéo, il est possible de définir manuellement une zone d'intérêt qui reflète les choix éditoriaux de chaque émission.

3.2.2 Post-traitement des hypothèses consécutives

Les performances de la reconnaissance de texte peuvent varier significativement d'une trame à l'autre. Les textes étant placés en surimpression, ils sont parfois incrustés en transparence et peuvent donc varier en fonction de l'image de fond. De façon générale, le contraste de la zone avec un fond dynamique peut également induire des performances variables. Dans (Prasad *et al.*, 2008), deux approches sont explorées pour compenser cette variabilité : un pré-traitement consistant à générer une unique image synthétisant les trames successives, un post-traitement consistant à combiner les différentes hypothèses produites à chaque trame à l'aide de l'algorithme NIST ROVER. (Liu *et al.*, 2009) propose également un post-traitement des hypothèses successives pour former un réseau de confusion, sur lequel est appliqué un modèle de langage de lettres pour rechercher un meilleur chemin. Le système proposé ici repose sur une construction particulière d'un réseau de confusions (CN pour *Confusion Network*.

Pour la construction du CN, le choix de l'hypothèse pivot est réalisé en sélectionnant la séquence de caractères la plus fréquemment reconnue. Ensuite, les hypothèses différentes sont triées par ordre de fréquence et le réseau est construit en alignant itérativement une nouvelle hypothèse sur la meilleure hypothèse à l'itération courante. Le processus d'alignement doit être particulièrement adapté pour optimiser la construction du réseau. Ici nous utilisons un algorithme d'alignement classique reposant sur une distance d'édition mais en ayant recours à une matrice de coût particulière prenant en compte les confusions fréquentes et pénalisant différemment les insertions et les ommissions.

4 Identification multimodale de personnes

L'identification multimodale a pour objectif de recouper les indices trouvés dans chaque modalité afin de déterminer l'identité des personnes présentes, même si chaque modalité prise séparemment n'en est pas capable. Une telle incapacité peut provenir du manque de modèles a priori comme c'est le cas dans ce travail, de l'absence de données ou d'occlusions dans une modalité, ou d'erreurs des systèmes de détection et regroupement. Le processus d'identification multimodal de personnes se base sur les descripteurs produits lors des deux phases précédentes de segmentation et d'extraction d'information. La première phase consiste à produire des hypothèses d'identités, extraites à partir des sorties des modules d'extraction d'information présentés dans la paragraphe précédent. La deuxième phase propage ces identités aux segments obtenus lors des phases de segmentation audio et vidéo.

4.1 Extraction multimodale d'hypothèses d'identités

Etant donné que dans cette étude nous n'utilisons aucun modèle *a priori* de personnes (ni modèle de voix de locuteurs, ni modèle de visages), les seules sources potentielles permettant d'identifier une personne à un temps donné sont les mentions de noms de personnes dans le signal audio

et dans les incrustations de textes dans les vidéos. La mention d'un nom de personne par un locuteur n'est pas un indice suffisant pour prédire sa présence dans la vidéo au moment où il est mentionné, contrairement aux *cartouches* de texte incrusté nommant la personne apparaissant à l'écran. C'est pour cette raison que dans le système PERCOL0 nous recherchons systématiquement ces *cartouches* car ce sont elles qui vont permettre d'attribuer de manière fiable une identité aux personnes présentes.

Cette recherche de cartouche est effectuée de la manière suivante :

1. Tout d'abord un lexique de noms de personnes susceptibles d'apparaître dans les vidéos a été extrait à la fois d'un corpus de dépêches de presse, ainsi que de listes de personalités (hommes politiques, journalistes) ; nous utilisons dans Percol0 un lexique de 160K expressions de noms de personnes.
2. Les noms de personnes de ce lexique sont tous cherchés systématiquement dans les réseaux de confusion de lettres provenant du module de VOCR présenté précedemment. Chaque fois qu'une forme peut être extraite du réseau, elle devient un candidat potentiel associé à un score de confiance provenant des scores des lettres du même réseau.
3. Les noms extraits par le détecteur d'entité nommée sont comparés à ceux obtenus sur les incrustations vidéos. Lorsque deux formes (l'une audio, l'autre vidéo) font référence au même nom [4], cette information est associée aux noms détectés.
4. Enfin chaque identité potentielle détectée dans les incrustations de texte est étiquetée ou pas comme *cartouche* à l'aide d'un classifieur Adaboost (Favre *et al.*, 2007) reposant sur des caractéristiques suivantes : position et taille du rectangle de texte, rapport entre la longueur du nom et la longueur du texte le contenant, mesure de confiance du système de VOCR , présence dans la canal audio.

4.2 Propagation de l'identité

Nous avons développé deux approches de propagation d'identité d'une modalité vers une autre dans PERCOL0. La première approche (propagation *locale*) consiste à propager localement l'identité reconnue : lorsqu'un nom de personne issu de l'incrustation vidéo est considéré comme une *cartouche* par le classifieur précédent, l'identité est associée au segment le plus long de la modalité cible recoupant le segment de l'incrustation. La seconde approche (propagation *globale*) exploite le regroupement préalable des segments (clustering de locuteurs ou clustering de visage) et affecte toutes les occurrences de l'identité dans la modalité cible. Si plusieurs identités se retrouvent candidates pour un même segment, un choix est fait en fonction de la fréquence des identités trouvées localement sur les différentes occurrences du cluster et du score de décision issu du module d'extraction multimodale d'hypothèses d'identités.

5 Expériences et premiers résultats

Les premiers résultats du système PERCOL0 présentés dans cette étude ont été obtenu sur le corpus de développement de la phase préliminaire de la campagne REPERE. Ce corpus contient

4. Un processus de résolution semi-supervisé permet de projer une forme vers une identité normalisée (ex : *B. Obama* → *Barack Obama*)

environ 3 heures de signaux vidéos provenant des chaînes LCP et BFM (émissions *Ca vous regarde*, *Entre les lignes*, *Pile et Face*, *Top Questions*, *LCP INFO*, *Planète Showbiz*, *BFM Story*). Nous possédons la transcriptions ainsi que la segmentation en locuteur de référence sur l'ensemble du corpus, mais uniquement 1088 images ont été annotées au niveau de la présence vidéo d'une personne et du texte incrusté. Les évaluations présentées ici seront donc menées uniquement sur ces 1088 trames. Ne disposant pas pour l'instant de corpus d'apprentissage, les modèles de segmentation et d'extraction d'information n'ont pas été entraînés sur des données provenant des mêmes chaînes et des mêmes émissions que le corpus d'évaluation. Seuls le classifieur et le processus de fusion ont utilisé le corpus REPERE, selon la technique du *Leave-One-Out* sur chaque fichier d'émission, afin de dissocier données d'entraînement et de tests.

TABLE 1 – Résultats sur le corpus REPERE de l'identification multimodale de locuteurs sans modèles a priori

Identification non supervisée de locuteurs (audio)					
méthode	nb tests	nb hyp	nb hyp correct.	rappel	précision
propagation locale	1013	406	292	28.8	71.9
propagation globale	1013	756	486	48.8	64.2

Les performances du processus de propagation de l'identité sont présentées pour l'identification non supervisée des locuteurs dans le tableau 1. Pour l'instant les modalités utilisées pour produire ces résultats sont : pour la modalité audio, la segmentation et le regroupement de locuteurs ainsi que la reconnaissance automatique de la parole et l'extraction d'entités nommées ; pour la modalité vidéo la reconnaissance de texte incrusté. L'ajout de la détection et du regroupement de visage est en cours d'évaluation dans le cadre de REPERE. Pour cette première évaluation les performances sont bornées par deux facteurs : d'une part le nombre de locuteurs et de visages effectivement identifiés par une cartouche d'incrustation de texte dans la vidéo (car nous ne voulons pas utiliser de modèles de locuteurs ou de visage a priori) ; d'autre part les performances du module de VOCR sur ces mêmes cartouches. Sur le premier point, une étude sur le corpus REPERE nous a montré que seuls 492 sur les 981 (50,1%) locuteurs des trames annotées sont identifiés dans la vidéo. En obtenant un rappel de 48.8%, le système *propagation globale* est donc satisfaisant. Sur le deuxième point une étude du % de noms correctement reconnus dans les cartouches par le système de VOCR couplé au système d'identification multimodal donne un rappel de 63.7 pour une précision de 79.5%. Au vu de ces chiffres les résultats obtenus par les systèmes de propagation sur les locuteurs sont prometteurs.

La mise en oeuvre de la propagation sur la modalité *visage* couplée à un processus de fusion pouvant tirer partie de la propagation dans les deux modalités permettra d'améliorer ces performances.

6 Conclusion

L'analyse de la présence de personnes dans des émissions télévisés montre qu'il n'y a pas nécessairement de recouvrement exact entre le visage à l'écran et la voix audible. Cette non-

synchronie des flux audio et vidéo (du point de vue de la personne) oblige à mettre en œuvre des traitements de fusion particuliers. Le problème réside essentiellement dans l'obtention de segmentation en différentes classes propres à chaque modalité (voix et visage), dans l'obtention de labels sur ces segments (par analyse des entités nommées, par modèle de reconnaissance de voix et de visage), et dans la propagation de ces labels sur les différentes segmentations. Le système PERCOL0 présenté dans cette étude est une première étape dans ce processus de fusion multimodal, évalué dans le cadre de la campagne REPERE.

Références

AUGUSTE, R., AISSAIOUI, A., MARTINET, J. et DJERABA, C. (2012). Ré-identification de personnes dans les journaux télévisés basée sur les histogrammes spatio-temporels. *In Extraction et gestion des connaissances (EGC'2012)*, pages 547–548.

BARRAS, C., ZHU, X., MEIGNIER, S. et GAUVAIN, J.-L. (2006). Multistage speaker diarization of broadcast news. *Trans. on Audio, Speech and Language Processing*.

BECHET, F. et CHARTON, E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. *In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.

DELAKIS, M. et GARCIA, C. (2008). Text detection with convolutional neural networks. *In International Conference on Computer Vision Theory and Applications, 2008. VISAPP 2008.*

FAVRE, B., HAKKANI-TÜR, D. et CUENDET, S. (2007). Icsiboost. `http://code.google.come/p/icsiboost`.

FREDOUILLE, C. et EVANS, N. (2008). New implementations of the E-HMM-based system for speaker diarisation in meeting rooms. *In Proc. ICASSP'08*, Brisbane, Australia.

LINARÈS, G., NOCÉRA, P., MASSONIÉ, D. et MATROUF, D. (2007). The lia speech recognition system : from 10xrt to 1xrt. *In Lecture Notes in Computer Science, 4629 LNAI*, pages pp. 302–308.

LIU, A., FEI, J., TANG, S., FAN, J., ZHANG, Y., J., L. et Z., Y. (2009). Confusion network based video ocr post-processing approach. *In IEEE International Conference on Multimedia and Expo, 2009. ICME 2009.*

PRASAD, R., SALEEM, S., MACROSTIE, E., NATARAJAN, P. et DECERBO, M. (2008). Multi-frame combination for robust videotext recognition. *In IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. ICASSP 2008.*

ZHAO, W., CHELLAPPA, R., ROSENFELD, A. et PHILLIPS, P. J. (2003). Face recognition : A literature survey. *ACM Computing Surveys*, pages 399–458.