Descripteurs visuels robustes pour l'identification de locuteurs dans des émissions télévisées de talk-shows

Félicien Vallet[1,2] Slim Essid[1] Jean Carrive[2] Gaël Richard[1]

[1]Télécom ParisTech CNRS/LTCI
46 rue Barrault,
75634 Paris cedex 13,
France
{firstname.lastname}@telecom-paristech.fr

[2]Institut national de l'audiovisuel
4 avenue de l'Europe,
94366 Bry-sur-Marne cedex,
France
{fvallet, jcarrive}@ina.fr

Résumé

Dans cet article, nous proposons une nouvelle méthode multimodale pour l'identification de locuteurs dans une émission de talk-show à base de machines à vecteurs supports. Notre étude met en évidence l'efficacité de descripteurs visuels spécifiques pour ce type de contenu vidéo, résultat de l'assemblage par un réalisateur des images prises par plusieurs caméras. Ces descripteurs sont motivés par des connaissance a priori sur l'approche suivie par le réalisateur dans la sélection des plans appropriés. Leur utilisation conjointement aux MFCCs (coefficients cepstraux à fréquence Mel) montre une amélioration significative du score d'identification (+8%) par rapport à plusieurs systèmes de référence dont un système audio standard.

Mots clefs

Système multimedia, analyse d'image, identification de locuteurs.

1 Introduction

L'Institut national de l'audiovisuel (Ina) a pour principales missions de regrouper, conserver et mettre à disposition les archives de la radio et de la télévision française. La classification et la segmentation automatique de ses contenus sont des étapes essentielles pour l'Ina dans de nombreux domaines tels que la recherche et l'indexation de vidéos, la structuration ou la génération automatique de résumés. Dans les scènes audiovisuelles, comme celles proposées dans un talk-show télévisé, le but d'une segmentation automatique est d'obtenir des segments pertinents tels que « passage musical », « extrait de film », « invité principal à l'écran » ou « arrivée d'un nouvel invité ».

De ce point de vue, l'identification de locuteurs est une tâche importante pouvant contribuer à une segmentation sémantique. Notre but est l'identification de locuteurs a priori inconnus. Il n'y a pas de données disponibles pour chaque locuteur pouvant être utilisées pour apprendre des classifieurs de façon totalement supervisée. Par conséquent, notre approche peut être considérée « semi-supervisée » en ce sens que les exemples sont collectés en ligne en demandant à l'utilisateur de notre système, par exemple un documentaliste de l'Ina traitant un talk-show, de sélectionner arbitrairement un court (4 à 15 s) extrait vidéo de chaque locuteur intervenant dans le talk-show. Ces courts extraits (un unique segment par locuteur) sont ensuite utilisés pour l'apprentissage de classifieurs pour l'identification de locuteurs sur l'intégralité de l'émission (environ 3 h). Il est important de noter que la sélection manuelle d'extraits par l'utilisateur est plus simple qu'il n'y paraît. Tout d'abord, le nombre total de locuteurs à chercher est connu a priori grâce à une notice, disponible pour chaque talk-show, détaillant la liste des invités. Ensuite, l'utilisateur peut rapidement localiser les extraits des différents locuteurs en utilisant le *slider* temporel et/ou le mode avance rapide.

Les systèmes de segmentation en locuteurs (*speaker diarization*) [1] sont des méthodes alternatives à notre approche. Cependant, ceux-ci étant totalement non-supervisés, l'utilisateur aura toujours à associer un locuteur à chaque cluster proposé avec l'inconvenient que certains locuteurs puissent être introuvables, ces systèmes n'étant pas parfaits. Par conséquent, notre approche ne semble pas plus coûteuse en termes d'efforts pour l'utilisateur.

Dans cet article, l'accent est mis sur l'utilisation de descripteurs visuels robustes augmentant l'ensemble d'attributs audio habituellement utilisé pour cette tâche, i.e. les coefficients cepstraux à fréquence Mel (MFCCs) [2], afin d'améliorer le taux de détection. Plusieurs études dans le domaine de la segmentation en locuteurs ont proposé des approches multimodales [3, 4]. D'autres travaux dans le champ multimodal ont montré l'intérêt de mesurer la synchronie audio-visuelle pour la détection du locuteur actif [5, 6, 7]. Cependant, dans chaque cas les auteurs ont utilisé des bases de données très particulières : corpus de news (National Institute of Standard and Technology Rich Transcription : NIST RT[1]), vidéos de meetings en mode multi-camera (Augmented Multi-party In-

[1]http ://www.nist.gov/index.html

teraction : AMI[2]), vidéos de locuteurs prononçant des nombres (Clemson University Audio Visual Experiments : CUAVE[3]), etc. Le contenu exploité dans cette étude a la particularité d'être monté, un réalisateur de télévision sélectionnant des plans pris par plusieurs cadreurs. De plus, ces plans présentent de grandes variations de cadrage et d'angles de vue. De fait, nous proposons une nouvelle approche d'identification des locuteurs mettant en évidence l'utilité de combiner des attributs audio classiques comme les MFCCs avec des descripteurs visuels spécifiques et robustes sur les talk-show télévisés fournis par l'Ina.
Dans la section 2 de cet article, nous détaillons les motivations et l'extraction des descripteurs visuels et en section 3, nous rappelons le principes des machines à vecteur support (SVM) utilisées pour la classification. Dans la section 4, nous présentons notre étude expérimentale et discutons des résultats avant de donner des perspectives pour la suite de ce travail dans la section 5.

2 Extraction de descripteurs visuels robustes

2.1 Motivations pour la conception des descripteurs

Nous traitons un type de contenu particulier : le talk-show télévisé. Contrairement aux bases de données utilisées en biométrie et segmentation en locuteurs, le contenu vidéo est monté, c'est-à-dire que plusieurs caméras sont utilisées pour le tournage et qu'à chaque instant, les images d'une seule sont diffusées. Le plan est choisi par le réalisateur qui, généralement, essaie de suivre le locuteur à l'écran. Par conséquent, bien que le cadrage varie (de plan large à gros plan), la plupart du temps « on voit qui l'on entend ». Comme décrit plus bas, nous extrayons des descripteurs visuels à partir des images des personnes à l'écran, en supposant qu'elles sont les locuteurs actifs. Il est important de mentionner que, les émissions étant réalisées en direct, notre tâche est compliquée par les conditions sonores bruyantes et la grande spontaneité de parole des intervenants.

2.2 Suivi de visage et de costume

Dans notre étude, les mouvements de caméra, le cadrage, les angles de vue et les conditions variables d'éclairage rendent très délicate l'utilisation d'un détecteur de visage comme, par exemple, dans [8]. Afin que le téléspectateur ne confonde pas les participants présents sur un plateau télévisé, leurs costumes sont en général précautionneusement choisis. Notre hypothèse étant qu' « on voit qui l'on entend », nous supposons que l'information portée par le costume peut aider à l'identification des locuteurs et présente l'avantage de pouvoir être extraite de façon plus robuste que les attributs de visage. Cette hypothèse est renforcée par l'importante corrélation entre couleur dominante du costume et tours de locuteurs comme présenté dans la Figure 1.

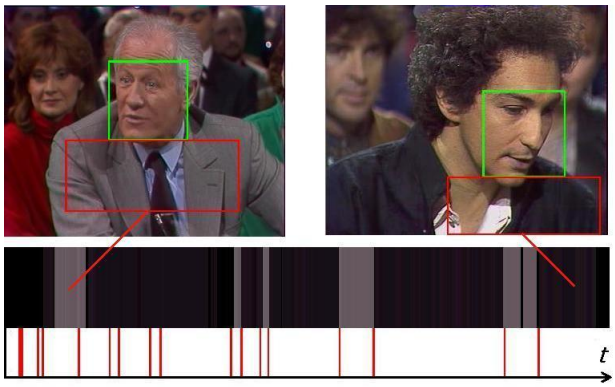

Figure 1 – *Couleur dominante du costume et tours de locuteur (en rouge) pour un segment de parole de 2min*

Nous décidons donc, comme dans [9], d'utiliser le costume comme attribut pour détecter automatiquement l'apparition d'une personne à l'écran. La présence d'un individu est attestée par la recherche de visages dans chaque trame. Nous utilisons l'algorithme de Viola et Jones [10] (disponible dans la librarie OpenCV [11]) pour détecter les visages. Nous déterminons ensuite les régions des costumes en traçant des rectangles sous les visages détectés comme dans [9].
Deux classifieurs sont utilisés pour détecter les visages de face et de profil pour chaque image. De plus, nous limitons le nombre de détections dans chaque trame et indiquons une taille minimale pour les visages, pour ne garder finalement que la plus grande des régions d'intérêt proposées. Ce processus doit empêcher, autant que possible, la détection de visages dans le public (à l'arrière plan et donc souvent plus petits).
Cependant, la détection de visages trame à trame introduit de nombreuses fausses alarmes et non-détections. Pour s'affranchir de ce problème nous utilisons une heuristique simple exploitant les propriétés temporelles de la vidéo. Ainsi, nous proposons de faire l'économie d'un suivi de visage puisqu'une analyse de flux optique est effectuée pour l'extraction de descripteurs de mouvement (cf Section 2.4).
Après avoir implémenté un détecteur de changement de plan (*cuts*) basé sur l'intersection d'histogrammes de couleur [12], nous extrayons des points d'intérêt sur les rectangles contenant visages et costumes avec l'algorithme de Shi et Tomasi [13]. Ensuite, ces points d'intérêt sont suivis entre deux *cuts* grâce à l'algorithme de Lucas et Kanade [14].
Le suivi est initialisé avec un nombre maximum de 300 points d'intérêt au temps t_s (correspondant à la première trame contenant un visage après le dernier *cut* détecté) et arrêté au temps t_e, soit à la fin du plan courant ou avant le *cut* suivant si plus d'un tiers des points suivis sont perdus entre deux trames (indiquant généralement la présence d'un changement de plan non-détecté). Cette procédure est répétée tout le long de l'émission et les erreurs de détection

[2]http ://corpus.amiproject.org
[3]http ://www.ece.clemson.edu/speech/cuave.htm

de visage sont corrigées entre t_s et t_e de la manière suivante.

Si une trame f ne présente pas de visages détectés entre t_s et t_e, les regions d'intérêt du visage et du costume de la trame $f-1$ sont utilisées pour déduire celles manquantes par translation (déduites du déplacement des points d'intérêt suivis). De plus, si entre t_s et t_e le déplacement entre les trames f et $f+1$ des régions d'intérêt est trop important par rapport à celui des points d'intérêt, nous supposons que le visage détecté à $f+1$ est différent de celui détecté à f. Il correspond souvent à un visage au second plan qui peut ponctuellement devenir prépondérant pour la trame $f+1$ à cause de variations du cadrage. Par conséquent, les régions d'intérêt correspondantes pour le mauvais visage et le mauvais costume sont supprimées et d'autres valides sont créées par translation de celles de la trame f comme expliqué précédemment. Le procédé d'extraction proposé est résumé dans la Figure 2.

Bien qu'une évaluation directe de la procédure précédente soit délicate en raison de l'absence de vérité terrain appropriée, il est important de noter que celle-ci est implicitement évaluée par les descripteurs déduits par la suite et utilisés pour l'identification de locuteurs (voir Section 4).

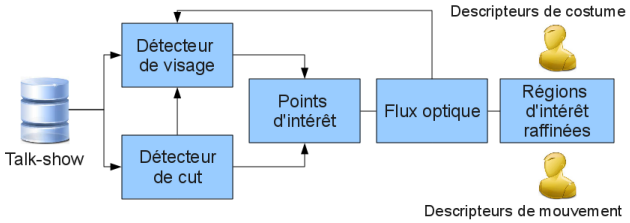

Figure 2 – *Procédé d'extraction des descripteurs*

2.3 Descripteurs de couleur du costume

Une fois obtenue une description robuste des occurrences des costumes des personnes à l'écran dans une vidéo, nous proposons deux attributs calculés sur les régions d'intérêt correspondantes et reposant sur le MPEG-7 Dominant Color Descriptor [15]. Celui-ci propose une description compacte de 1 à 8 couleurs dominantes pour une image ou région d'intérêt, une couleur dominante étant un vecteur de trois composantes RGB $[x_R, x_G, x_B]$, chacune quantifiée sur 32 bins.

Le premier descripteur calculé est basé sur l'association des deux couleurs principales, c'est-à-dire les deux couleurs avec la plus grande proportion P_i de pixels, où i est l'indice des couleurs dominantes pour l'image. Dans le cas où une seule couleur dominante est détectée, celle-ci est dupliquée. Notre second descripteur est la couleur dominante moyenne qui est une moyenne pondérée des n couleurs recouvrant au moins 40% des pixels de la region d'intérêt selon :

$$x_{C_{avg}} = \frac{\sum_{i=1}^{n} P_i . x_{C_i}}{\sum_{i=1}^{n} P_i} \quad (1)$$

avec

$$\sum_{i=1}^{n} P_i \geq 40\% \quad (2)$$

x_{C_i} étant la valeur de la i$^{\text{ème}}$ couleur dominante. Ce ratio de 40% s'est avéré être une estimation robuste de la surface couverte par le costume dans la region d'intérêt, en prenant en compte les conditions potentiellement bruitées survenant par exemple lorsque les mains du locuteur entrent dans le rectangle englobant le costume.

2.4 Descripteurs de mouvement du locuteur

Inspiré de [4], nous proposons de calculer plusieurs descripteurs de mouvement basés sur l'analyse du flux optique. Nous supposons que chaque locuteur possède ses propres gestuelles et expressions qui, décrites avec les bons attributs, peuvent être très discriminantes (par exemple le mouvement des mains souvent visible dans la région d'intérêt du costume). Pour caractériser ces particularités de façon robuste et efficace, nous proposons de déduire du flux optique des descripteurs de mouvement pour l'image globale, ainsi que pour les régions d'intérêt du visage et de la poitrine (qui est la même que celle du costume) comme montré dans la Figure 3 a).

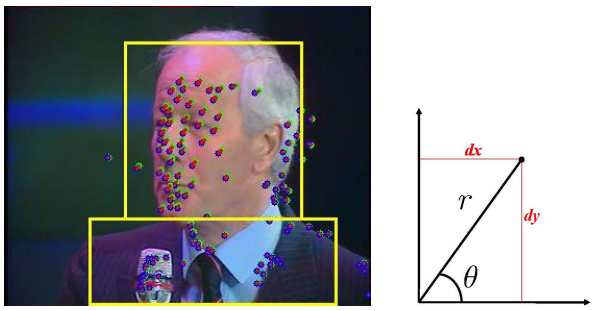

Figure 3 – *a) Mouvement entre deux trames consécutives pour une sélection de points d'intérêt. b) Amplitude de mouvement r et orientation θ pour un déplacement (dx, dy)*

Les amplitudes et orientations de vitesse et d'acceleration sont calculées comme les dérivées première et seconde des points d'intérêt de l'image globale et des régions d'intérêt, celles-ci étant évaluées comme les coordonnées r et θ dans un système polaire. Nous proposons également de calculer l'amplitude relative présentée comme le rapport des amplitudes pour les régions d'intérêt du visage et de la poitrine sur celle de l'image globale. Le Tableau 1 récapitule les différents descripteurs proposés dans cette section et donne leurs acronymes et dimensions.

Acronyme	dim	Description
DomColMoy	3	couleur dominante moyenne du costume
DomCol2	6	2 couleurs dominantes principales du costume
OrientVit	3	Orientation de vitesse pour image / visage / costume
OrientAcc	3	Orientation d'acceleration pour image / visage / costume
IntVit	5	Amplitude de vitesse absolue et relative pour image / visage / costume
IntAcc	5	Amplitude d'acceleration absolue et relative pour image / visage / costume

Tableau 1 – *Description du set d'attributs proposé*

2.5 Interpolation pour trames sans visage

Pour un certain nombre de trames, aucun visage n'est détecté à cause des choix de conception présentés dans la Section 2.2. Par conséquent, afin d'obtenir des descripteurs continus, nous interpolons les descripteurs sur les trames sur lesquelles aucun visage n'a été détecté. Cela est réalisé en choisissant aléatoirement pour ces trames la valeur des descripteurs précédents ou suivants. Cette stratégie peut sembler assez approximative mais donne de bons résultats (voir Section 4).

3 Classification SVM

En raison de leur efficacité à résoudre un grand nombre de problèmes de classification, les classifieurs SVM sont devenus très populaires dans de nombreux domaines de recherche. Nous conseillons au lecteur de se référer aux ouvrages de référence [16] et ne rappelons ici succintement que les principes de base.

Dans les problèmes bi-classes, l'algorithme d'apprentissage SVM recherche l'hyperplan $\mathbf{w}.\mathbf{x} + b = 0$ qui sépare les exemples d'apprentissage $\mathbf{x}_1, ..., \mathbf{x}_n$ assignés aux classes $y_1, ..., y_n$ ($y_i \in \{-1, 1\}$) tel que $y_i(\mathbf{x}_i.\mathbf{w} + b + \xi_i) - 1 \geq 0, \forall i$, sous la contrainte que la distance $\frac{2}{||\mathbf{w}||}$ entre l'hyperplan et les plus proches exemples soit maximale, ξ_i étant des variables d'écart positives prenant en compte les éventuels points abérrants ou *outliers*.

Lors de la résolution de ce problème d'optimisation la somme des ξ_i est pénalisée par un facteur de coût C (à définir) afin de contrôler le nombre total d'*outliers* autorisés. Il est possible d'utiliser différents facteurs de coût C_+ et C_-, respectivement associés aux classes positives et négatives, dans le cas d'ensembles d'apprentissage déséquilibrés et cela afin d'éviter que la solution ne soit biaisée par la sur-représentation d'une classe par rapport à l'autre [17].

Les données n'étant pas linéairement séparables dans l'espace des d descripteurs initiaux, une fonction noyau $k(\mathbf{x}, \mathbf{y})$ peut être utilisée pour projeter les données dans un espace de beaucoup plus grande dimension dans lequel les deux classes deviennent linéairement séparables. Un vecteur de la base de test est alors classé suivant le signe de la fonction $f(\mathbf{x}) = \sum_{i=1}^{n_s} \alpha_i y_i k(\mathbf{s}_i, \mathbf{x}) + b$, où $\mathbf{s}_i$ sont les vecteurs supports, α_i les multiplicateurs de Lagrange, et n_s le nombre de vecteurs supports. Dans cette étude nous utilisons le noyau gaussien $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2d\sigma^2}\right)$.

Nous testons également les performances obtenues à l'aide de SVM transductifs (TSVM) [18]. Dans ce cas, en plus de l'ensemble d'exemples étiquetés utilisé avec le SVM, on fournit au classifieur un ensemble d'exemples dont l'étiquette est non renseigné. L'algorithme alloue ensuite de façon itérative une étiquette à chacun de ces exemples. En plus des paramètres C et σ, l'utilisateur doit spécifier la proportion p des données de test à assigner à chaque classe.

4 Etude expérimentale

4.1 Description

L'évaluation des descripteurs visuels proposés est effectuée sur une émission de 3.5 h, appartenant au corpus « Le Grand Échiquier » (plus de 50 émissions de talk-show des années 1980). Cette base de données présente des caractéristiques qui l'ont rendue célèbre parmi plusieurs projets européens et nationaux. Chaque émission est consacrée à un invité principal entouré d'invités secondaires. Les interviews sont ponctuées d'extraits de film, de passages musicaux ou de scènes de théâtre. Les données sont disponibles au format MPEG2 et la piste audio est rééchantillonnée à 12.8 kHz en raison d'une largeur de bande très étroite dans les fichiers originaux. Seules les sections de parole sont traitées, représentant plus d'1.5 h.

Comme le Tableau 2 l'indique, les tours de locuteurs à identifier sont de longueurs très variables. Il est également bon de noter que deux personnes se partagent environ 70% du temps total de parole : le présentateur et l'invité principal.

temps de parole total	5745 s
nombre de locuteurs	10
nombre de sections de parole	64
nombre de tours de locuteurs	1048
durée moyenne d'un tour	5.3 s
écart type de la durée d'un tour	7.1 s
tour de parole le plus long	92 s
tour de parole le plus court	0.2 s

Tableau 2 – *Caractéristiques des données audio*

La piste audio est extraite du fichier vidéo, convertie en mono en moyennant canaux droit et gauche puis sous-échantillonnée. Ensuite, nous extrayons les 13 premiers coefficients MFCCs, en incluant le coefficient cepstral d'ordre 0. Les descripteurs vidéo étant calculés toutes les 40 ms (25 trames/sec), nous effectuons une intégration temporelle, 4 trames audio consécutives étant moyennées afin que descripteurs audio et vidéo soient disponibles à la même fréquence (25 Hz). Nous supposons également qu'elles peuvent être considérées comme synchrones (ce qui sera discuté par la suite). Les vecteurs de descripteurs sont finalement formés par simple concaténation.

En plus des descripteurs visuels robustes décrits dans la Section 2, nous extrayons l'histogramme de couleur YUV et le MPEG-7 ColorLayout Descriptor [15] sur chaque trame pour servir d'attributs de référence.
La sélection des courts extraits d'apprentissage par le documentaliste est simulée par le choix aléatoire pour chaque locuteur de segments longs de 4 à 15 s. La classification SVM, effectuée en mode *un contre un*, est réalisée à l'aide de la toolbox *SVM^{light}* développée dans [19] avec un noyau gaussien. Le paramètre σ est fixé pour toutes les expériences par 5 validations croisées sur la base d'apprentissage tandis que les facteurs de coût C_-/C_+, traitant des déséquilibres à l'intérieur de l'ensemble d'apprentissage, sont fixés par le rapport du nombre d'exemples positifs sur celui d'exemples négatifs. 100 validations croisées sur les segments de parole candidats assurent la validité des résultats, c'est-à-dire que le tirage aléatoire des segments d'apprentissage et de test est effectué 100 fois.
Les résultats présentés dans la section 4.2 sont obtenus à l'aide de la fonction *detection error rate* (DER) utilisée pour les évaluations NIST Rich Transcription (RT). Cette fonction est spécialement conçue pour la tâche de classification de locuteur. Cependant, en raison des spécificités de notre corpus, nous proposons d'utiliser parallèlement une nouvelle métrique. En effet, la répartition du temps de parole étant très déséquilibrée entre les n locuteurs, l'identification correcte des deux principaux suffirait à l'obtention de bons résultats avec la fonction NIST RT. Par conséquent, la métrique que nous proposons pondère le volume de parole de chaque personne de sorte qu'elle est plus sensible à la correcte identification d'un locuteur qu'à la longueur totale de ses interventions :

$$\text{New DER} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{nombre de trames mal-classées pour le locuteur } i}{\text{nombre total de trames pour le locuteur } i} \quad (3)$$

4.2 Résultats et discussion

Les résultats du Tableau 3 sont obtenus après intégration temporelle de la sortie des classifieurs SVM sur des fenêtres de 0.5 s avec un recouvrement de 50%. Ils montrent qu'une amélioration significative est obtenue avec notre ensemble de descripteurs visuels robustes en comparaison avec un système basé sur les MFCCs uniquement. De plus, les performances de ces descripteurs dépassent largement celles des associations d'attributs visuels classiques et de MFCCs. En fait, ces dernières dégradent fortement le score de référence obtenu avec les MFCCs. Ceci peut être expliqué par l'ajout d'informations bruitées dues au manque de focus des descripteurs visuels calculés globalement. L'addition de la couleur dominante moyenne du costume entraîne une amélioration de +8%, validant l'hypothèse de caractérisation d'un locuteur par son costume.
Les attributs visuels proposés attestent d'une grande robustesse. En effet, ils fonctionnent de façon collaborative avec les MFCCs, en assurant une grande discrimination au niveau local mais également en permettant aux MFCCs de dominer la description grâce à leur plus grande dimension en cas de non-détection de visage ou d'interpolation incorrecte. Nous remarquons également que l'ensemble d'attributs donnant le meilleur score NIST DER ne donne pas le meilleur score avec la métrique que nous proposons. Cela était attendu, sachant que certains locuteurs parlent moins de 10 s alors que d'autres interviennent plus de 30 mn.

Ensemble d'attributs	NIST DER	New DER
MFCC-DomColMoy-OrientVit	29.5	46.0
MFCC-DomColMoy-IntVit	27.9	44.5
MFCC-DomColMoy-OrientAcc	27.4	44.3
MFCC-DomColMoy-IntAcc	28.7	42.4
MFCC-DomColMoy	27.5	43.4
MFCC-YUVHistCol	52.1	53.7
MFCC-ColorLayout	59.8	63.6
MFCC	35.4	52.1

Tableau 3 – *Erreur de classification pour différents sets d'attributs*

La synchronisation des descripteurs audio et vidéo ne s'avère pas être critique ici. En effet, une étude préliminaire nous a montré une relative insensibilité du système de classification à l'introduction d'un retard entre modalité audio et vidéo.

Enfin, l'utilisation de TSVM n'a pas permis une amélioration des résultats de classication. Alors qu'il n'y a pas de raisons qu'un classifieur TSVM correctement calibré se comporte plus mal qu'un SVM classique, le paramètre p, relatif à la répartition du temps de parole pour chaque locuteur peut s'avérer trop restrictif pour notre approche. En effet, nous ne pouvons attendre de l'utilisateur de notre système qu'il soit capable de faire d'importantes suppositions quant au temps de parole de chaque locuteur en raison de la structure a priori inconnue de l'émission.

5 Conclusion et perspectives

Nous avons montré que l'extraction de descripteurs visuels basés sur le mouvement et la couleur dominante du costume de la personne à l'écran améliore significativement le taux d'identification de locuteurs dans des émissions de talk-show. Combinés avec des attributs classiques MFCCs, ces deux ensembles assurent une très bonne discrimination entre locuteurs. Ils s'avèrent d'ailleurs être particulièrement efficaces pour le traitement d'émissions télévisées. Ce type de réalisation et le déséquilibre entre les interventions de chaque locuteur rendent cette tâche particulièrement difficile. Dans un travail futur, nous combinerons notre approche avec un système de segmentation en locuteurs pour réduire encore l'intervention humaine dans le processus d'identification de locuteur.

Références

[1] S.E. Tranter et D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (5) :1557–1565, 2006.

[2] Lawrence Rabiner et Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice Hall PTR, 1993.

[3] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, et R. Kasturi. Audio segmentation and speaker localization in meeting videos. Dans *International Conference on Pattern Recognition*, 2006.

[4] Gerald Friedland, Hayley Hung, et Chuohao Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. Dans *International Conference on Acoustics, Speech and Signal Processing*, 2009.

[5] Harriet J. Nock, Giridharan Iyengar, et Chalapathy Neti. Speaker localisation using audio-visual synchrony : An empirical study. Dans *International conference on image and video retrieval*, 2003.

[6] Malcolm Slaney et Michele Covell. Facesync : A linear operator for measuring synchronization of video facial images and audio tracks. Dans *NIPS*, 2000.

[7] John Hershey et Javier Movellan. Audio-vision : Using audiovisual synchrony to locate sounds. *Advances in Neural Information Processing (MIT Press)*, 1999.

[8] Ming-Yu Chen et Alexander Hauptmann. Searching for a specific person in broadcast news video. Dans *International Conference on Acoustics, Speech and Signal Processing*, 2004.

[9] Gaël Jaffré et Philippe Joly. Costume : A new feature for automatic video content indexing. Dans *International conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2004.

[10] Paul Viola et Michael Jones. Robust real-time object detection. Dans *International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling*, 2001.

[11] G. Bradski et A. Kaehler. *Learning OpenCV : Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.

[12] Michael J. Swain et Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7 :11–32, 1991.

[13] J. Shi et C. Tomasi. Good features to track. Dans *Conference on Computer Vision and Pattern Recognition*, 1994.

[14] Bruce Lucas et Takeo Kanade. An iterative image registration technique with an application to stereo vision. Dans *International Joint Conference on Artificial Intelligence*, 1981.

[15] B.S. Manjunath, Philippe Salembier, et Thomas Sikora, éditeurs. *Introduction to MPEG-7 - Multimedia Content Description Interface*. Wiley, 2002.

[16] Bernhard Scholkopf et Alexander J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT-Press, 2001.

[17] K. Morik, P. Brockhausen, et T. Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. Dans *International Conference on Machine Learning*, 1999.

[18] Thorsten Joachims. Transductive inference for text classification using support vector machines. Dans *International Conference on Machine Learning*, 1999.

[19] Thorsten Joachims. *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*. MIT-Press, 1999.