ROBUST VISUAL FEATURES FOR THE MULTIMODAL IDENTIFICATION OF UNREGISTERED SPEAKERS IN TV 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

[2]Institut national de l'audiovisuel
4 avenue de l'Europe
94366 Bry-sur-Marne cedex, France

ABSTRACT

In this paper we propose a novel multimodal method for identifying unregistered speakers in a TV talk-show using a semi-supervised learning approach based on Support Vector Machines. Our study highlights the fact that specific visual features prove to be very efficient for this particular type of video content which is edited from multi-camera recordings. These visual features, motivated by prior knowledge on the approach followed by the TV director in choosing the appropriate shots, are found to bring a significant improvement in identification accuracy when used together with classic audio Mel-frequency cepstral coefficients (+8% compared to various baseline systems, in particular a standard audio only system).

Index Terms— multimedia systems, pattern classification, image analysis, multimedia databases.

1. INTRODUCTION

The French National Institute of Audiovisual (Ina) is a repository of French radio and television audiovisual archives. Its main missions are to gather, store and share images and sounds. Content structuring or segmentation is for Ina an essential process in a number of application domains such as automatic video archiving and retrieval, scene categorization or automatic summary generation. In specific audiovisual scenes such as those of TV talk-shows, an automatic segmentation typically aims to obtain meaningful and semantically rich segments such as "musical passage", "film excerpt", "main guest on screen" or "arrival of a new guest".

From this viewpoint, speaker identification can be regarded as an important tool among others, contributing to the construction of a semantic segmentation. Our aim is to identify unregistered speakers (i.e. not known a priori). Hence, there is no training database with examples of every speaker available in advance that can be used to learn classifiers in a totally supervised fashion. Our approach can thus be considered as "semi-supervised" in the sense that the training data is collected on the fly by the operator of our system, for instance the Ina archivist handling a given talk-show, who is asked to arbitrarily select one short (4 to 15 s) video excerpt of each speaker involved. These short excerpts (a single random segment per speaker) are then used to learn classifiers that are subsequently applied on the whole show (about 3-h long) to perform speaker identification at every time instant. It is worth mentionning that the manual selection of excerpts by the operator is greatly simplified by the two following facts. First, the total number of speakers to look for is known a priori since a short textual notice, giving a list of all guests to be identified is available for every talk-show video. Second, the operator can quickly locate examples of the different speakers using visual inspection via a temporal slider and/or the fast forward video mode.

An alternative to our approach is to use a speaker diarization system [1]. However, since speaker diarization is totally unsupervised, the operator would still need to assign a speaker to each cluster with the inconvenience that some speakers may not be found among the clusters determined automatically (as these systems are not perfect), hence our approach is not necessarily more costly in terms of human operator efforts.

In this paper, the focus is put on the use of robust visual features to augment the standard set of audio features usually used for this task, i.e. Mel-frequency cepstral coefficients (MFCCs), in order to improve identification accuracy. Several speaker diarization studies investigated multimodal approaches [2, 3]. Other works in the multimodal field showed the benefits of measuring audio-visual synchrony for the detection of an active speaker [4]. However, in every case the authors used specific databases built for the task considered : broadcast news (National Institute of Standard and Technology Rich Transcription : NIST RT[1]), meeting videos with several cameras (Augmented Multi-party Interaction : AMI[2]), corpus of utterances of digits (Clemson University Audio Visual Experiments : CUAVE[3]), etc. The content exploited in this study presents the major difference of being edited, meaning that a selection of shots taken by professional cameramen were assembled by a TV director and/or editor. Moreover, these shots display great variations of camera angles and field size. Therefore, we are proposing a novel approach for speaker classification showing the usefulness of combining classical audio features such as MFCCs and specific robust visual features on TV talk-shows provided by Ina.

In Section 2 of this paper, we detail the motivation and extraction of our visual features, while in Section 3 we recall the principle of support vector machines (SVM) that we use for classification. In Section 4, we present our experimental study and discuss the results before we conclude and give the perspectives in Section 5.

2. EXTRACTION OF ROBUST VISUAL FEATURES

2.1. Motivations behind the design of features

We deal with a particular type of content, that is TV talk-shows. Contrary to databases used in biometric and speaker diarization works, the video content is edited i.e. while many cameras are used

[1]http://www.nist.gov/index.html
[2]http://corpus.amiproject.org
[3]http://www.ece.clemson.edu/speech/cuave.htm

during the shooting, in general the images of a single one are shown at a time, corresponding to the current shot. The shot is chosen by the TV director (typically through a video switcher) who generally tries to follow the speaker. Hence, though the field size of the shot is varying (from long shots to close-ups), most of the time "you see who you hear". Therefore, we extract visual features from images of the persons appearing on the screen, assuming that they are the active speakers, as described hereafter. It is worth mentionning that the live nature of the talk-shows makes our task quite challenging, especially owing to the potentially noisy sound conditions and the complete spontaneity of the speech.

2.2. Face and dress tracking

The use of a facial recognition system as in [5] is rather difficult in our task due to the camera movements, the changing field size and angles of shot, the changing postures of the filmed persons and the varying lightning conditions. In order for the viewer not to confuse the participants on a TV set, their dress are usually carefully chosen. Our main assumption being that we often "see who we hear", we suppose that the information carried by the dress can help us for speaker identification and have the advantage that it can be extracted even more robustly than the persons' facial features. This assumption is reinforced by Figure 1 showing the correlation between the dress dominant color of the person on screen and the speech turns, the two speakers being those presented just above.

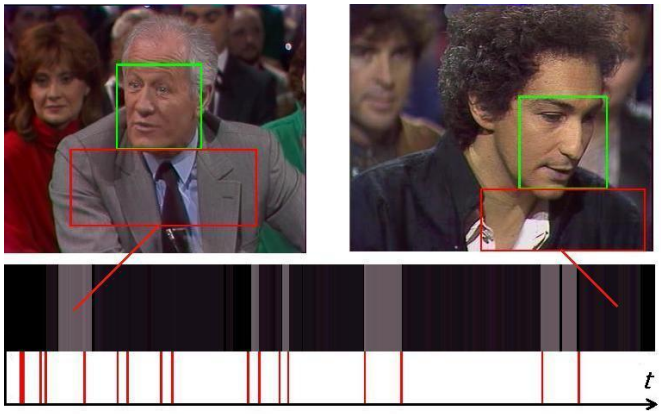

Fig. 1. Dress dominant color and speech turns (in red) for a 2min speech segment.

Therefore, as in [6], we decide to use the dress as a feature to automatically detect the appearance of a person on screen. To this end, we start by searching for faces within each video frame. We use the widely acknowledged algorithm presented by Viola and Jones in [7] and available in the OpenCV library [8] to detect faces. Then we determine the bounding boxes of the dress as in [6] by drawing rectangles below the faces detected.

Two Haar cascade classifiers are used to detect frontal and profile faces within each image. Also, we limit the number of detections in each frame and set a minimum face size, before keeping only the largest of the proposed bounding boxes. This process is supposed to avoid, as much as possible, detecting faces in the audience (in the background, hence often smaller).

However, frame by frame face identification introduces many false-alarms and misdetections. To alleviate this problem we use a simple heuristic procedure exploiting the temporal properties of the video, which turns out to be efficient on our data. We take advantage of the fact that we are anyway performing optical flow analysis to extract motion features (as described in Section 2.4) to save the need to use a sophisticated face tracker.

After implementing a basic shot boundary detector based on color histogram intersection, we extract corner points over the detected faces and dress, using the algorithm of Shi and Tomasi [9]. Then, these points of interest are tracked by running the Lucas and Kanade algorithm [10] between two shot boundaries. This two-frame differential method computes an estimation of the optical flow under the assumption of consistency of the motion in a local neighbourhood.

The tracking is initialized with a maximum of 300 points of interest at time t_s (corresponding to the first frame where a face is detected after the last shot boundary), and stopped at time t_e, either at the end of the current shot or before the end of the shot if more than 1/3 of these tracked points are lost between two frames (generally indicating a shot change not detected by histogram intersection). This procedure is repeated over the whole show duration, then face detection errors are corrected on every segment between t_s and t_e.

Though a direct evaluation of the previous procedure is difficult to achieve owing to the absence of appropriate ground-truth, it is important to note that it is implicitly evaluated through the assessment of the features deduced after it and used for speaker identification (see Section 4).

2.3. Dress color features

Once obtained a robust detection of the occurences of dress in a video, we propose to compute two features based on the MPEG-7 Dominant Color Descriptor [11] over the corresponding bounding boxes. This descriptor provides a compact description of 1 to 8 main colors within an image or region of interest where a main color is a vector of three RGB component values, each one quantized on 32 bins.

The first feature which we compute is based on the association of the two main dominant colors i.e. the two colors with the highest fraction P_i of pixels, corresponding to a given dominant color i in the image. In case only one dominant color is detected, this one is duplicated. Our second feature is the average dominant color which is a weighted average of the n colors covering at least 40% of the bounding box according to :

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

with $\sum_{i=1}^{n} P_i \geq 40\%$ and x_{C_i} is an RGB component vector. This ratio has been found to be a robust estimate of the surface covered by the dress, taking into acount potentially noisy conditions, typically occuring when hands enter the dress bounding box.

2.4. Speaker motion features

Inspired by [3], we propose to compute several motion-related descriptors based on optical flow analysis. Our assumption is that each speaker possesses his/her own gestures and expressions which, described with the right features, can be very discriminative. Indeed, some speakers show a distinctive body-language, for instance the motion of their hands which are often visible in the face and dress bounding boxes. Thus, to encompass those characteristics in a coarse and robust manner, we propose to deduce from the optical flow, descriptors of the motion for the global image, the face and chest bounding box (which is also the dress bounding box).

The motion speed and motion acceleration intensities and orientations are computed as the first and second derivatives for points

of interest of the global image and the bounding boxes. Intensities and orientations are evaluated as the r and θ coordinates in a polar system. We also propose to compute the relative intensity presented as the ratio of the intensity in the face and chest bounding boxes over the whole image average intensity. Table 1 sums up the various features proposed in this section and gives their acronyms and dimensions :

Acronym	dim	Description
AvgDomCol	3	average dominant color of the dress
MainDomCol	6	two main dominant colors of the dress
SpeedOrient	3	global/face/dress motion speed orientations
AccelOrient	3	global/face/dress motion accel. orientations
SpeedInt	5	absolute and relative global/face/dress motion speed intensities
AccelInt	5	absolute and relative global/face/dress motion accel. intensities

Table 1. Description of the set of features proposed.

2.5. Feature interpolation on faceless frames

Faces are not detected on a number of frames, as a consequence of our design choices presented in Section 2.2. Therefore, in order to obtain temporally continuous features, we propose to interpolate empty feature frames. This is done by randomly picking the descriptor value of the previous or the following computed detection to be used on these frames. This strategy may seem as rather approximate but it showed good results (see Section 4).

3. SVM CLASSIFICATION

SVM classifiers have proven efficient for a wide range of classification tasks and have become very popular in various research areas. We refer the reader to one of the many good tutorials on this powerful tool [12] and merely recall here the basic concepts which are referred to in the sequel.

In bi-class problems, the SVM learning algorithm searches for the hyperplane $\mathbf{w}.\mathbf{x} + b = 0$ that separates the training samples $\mathbf{x}_1, ..., \mathbf{x}_n$ which are assigned labels $y_1, ..., y_n$ ($y_i \in \{-1, 1\}$) so that $y_i(\mathbf{x}_i.\mathbf{w} + b + \xi_i) - 1 \geq 0, \forall i$, under the constraint that the distance $\frac{2}{||\mathbf{w}||}$ between the hyperplane and the closest sample is maximal, ξ_i being positive slack variables used to account for outliers.

When solving this optimisation problem the sum of the ξ_is is penalised in the objective function with a weight factor C (to be chosen) in order to control the total number of outliers. When handeling problems with imbalanced training sets, it is possible to use a different C factor for positive and negative training examples, which we will refer to as C_+ and C_- respectively, so that the solution is not biased by the overrepresentation of one class at the expense of another. Since the data is not linearly separable in the original feature space, a kernel function $k(x, y)$ can be used to map the d-dimensional input feature space into a higher dimension space where the two classes become linearly separable. A test vector $\mathbf{x}$ is then classified with respect to the sign of the function $f(\mathbf{x}) = \sum_{i=1}^{n_s} \alpha_i y_i k(\mathbf{s}_i, \mathbf{x}) + b$, where $\mathbf{s}_i$ are the support vectors, α_i are Lagrange multipliers, and n_s is the number of support vectors. In this work, we exploit the Gaussian kernel $k(x, y) = \exp\left(-\frac{||x-y||^2}{2d\sigma^2}\right)$.

We also explore the usefulness of transductive SVM (TSVM) [13] for our task. In addition to the set of labeled training examples used with SVM, the classifier is given a set of unlabeled examples to be classified. The algorithm proceeds iteratively to allocate a label to each one of the unlabeled examples. Besides the parameters C and σ, the user must specify the proportion p of test examples to be assigned to each class.

4. EXPERIMENTAL STUDY

4.1. Experiment description

The evaluation of our speaker-oriented visual features is done on a 3.5-h show, part of the corpus "Le Grand Échiquier" (over 50 TV talk-show programs of the 1980s). This database presents characteristics that made it popular among several European and national projects. Each show is dedicated to a special main guest who is interviewed together with other friend guests. The interviews are punctuated with film excerpts, live music or acting performances. The data is available in MPEG2 format and the audio is sampled at 12.8 kHz due to a very narrow bandwidth in the original files. Only the speech sections of the show are processed, representing more than 1.5 h.

As can be seen in Table 2, the speaker turns that are to be detected and identified are of very variable length. It is also worth noting that two persons share always about 70% of the load of the total speech : the host and the main guest.

total speech duration	5745 s
number of speakers	10
number of speech sections	64
number of turns	1048
turn average duration	5.3 s
turn duration standard deviation	7.1 s
longest speech turn	92 s
shortest speech turn	0.2 s

Table 2. Characteristics of the speech data.

The audio track is extracted from the video, converted to mono by averaging right and left channels and downsampled. Then, we extract every 10 ms the first 13 MFCCs, including the 0-th order cepstral coefficient. Since video features are computed every 40 ms (25 frames/sec), we use temporal integration, 4 consecutive audio frames being temporally averaged, so that the audio and visual features are available at the same rate (25 Hz). We also assume that they can be considered as synchronous (which will be assessed hereafter).

In addition to the robust visual features described in Section 2, we extract the YUV color histogram and the MPEG-7 ColorLayout descriptor [11] on each video frame to serve as reference features. Tested feature vectors are finally formed by simple concatenation of audio and visual features.

The selection of the short training excerpts by the archivist is simulated by randomly choosing a 4 to 15 s length video of each speaker. For SVM classification, which is carried out in a one vs one fashion, we use the *SVM*^{light}[4] toolbox with a Gaussian kernel. The σ parameter is set for all experiments through a 5-fold cross-validation over the training database while the C_-/C_+ cost-factors, coping with the classes imbalance within the learning samples, are set by considering the ratio of negative over positive examples. A 100-fold cross-validation is performed on the various candidate speech segments to assess the validity of the results.

The results shown in Section 4.2 are obtained through the detection error rate (DER) function used for the NIST Rich Transcription

[4] http://svmlight.joachims.org/

(RT) evaluations. This scoring function is specially designed for the speaker classification task. However, due to the specificity of our corpus we propose to use a new metric in parallel. Indeed, since the speech repartition is greatly imbalanced between the n speakers, the correct identification of only the main two speakers would lead to good results with the NIST RT function. Therefore, our new metric weights the speech amount of each person so that it is more sensitive to the correct identification of a speaker than to his/her speech load :

$$\text{New DER} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{number of misclassified frames for class } i}{\text{total number of frames for class } i}$$

4.2. Results and discussion

The results displayed in Table 3 are obtained after temporal integration of the SVM classifiers outputs over 0.5-s windows with a 50% overlap. They show the significant improvement with our robust visual features compared with a system based on MFCCs only. Besides, these features outperform the association of MFCCs and classical visual features. In fact, these last two deteriorate drastically the baseline obtained with the MFCCs. This could be expected since globally computed visual features tend to add noisy information due to their lack of focus. The addition of the average dominant color on the dress yields an accuracy improvement of 8% validating the idea of characterizing a speaker by his/her dress. The combined use of MFCCs, average dominant color of the dress and acceleration orientation obtains the best classification score.

Feature set	NIST DER	New DER
MFCC-AvgDomCol-SpeedOrient	29.5	46.0
MFCC-AvgDomCol-SpeedInt	27.9	44.5
MFCC-AvgDomCol-AccelOrient	27.4	44.3
MFCC-AvgDomCol-AccelInt	28.7	42.4
MFCC-AvgDomCol	27.5	43.4
MFCC-YUVColorHistogram	52.1	53.7
MFCC-ColorLayout	59.8	63.6
MFCC	35.4	52.1

Table 3. Speaker DER in percentage for several feature sets.

The visual features proposed show a great robustness. Indeed, they work collaboratively with MFCCs, ensuring locally a great discrimination but also allowing the MFCCs to take over in case of misdetection or incorrect interpolation, thanks to their smaller dimensions. We also notice that, the feature set performing best with the NIST DER is not necessarily the best with our new DER. This was expected, the task being really challenging particularly knowing that some people speak less than 10 s while other more than 30 mn. The fact that the visual features are potentially not always synchronous with the audio features turns out not to be critical for this task. This has been further validated by a preliminary study proving that our system was rather robust to the introduction of a delay between audio and video modalities.

Finally, the use of TSVM turns out not to improve the classification performance. While there is no reason for a correctly parametrized TSVM classifier to perform worse than a SVM one, the fine-tuning of the p parameter, relating to the speaker turns repartition, may be too restrictive. In fact, we do not expect the user to be able to make any strong assumption on the speech load of each person due to the a priori unknown structure of the show.

5. CONCLUSION AND PERSPECTIVES

We have shown that the extraction of robust visual features based on the motion and the dress dominant color of the onscreen person improves significantly the identification accuracy of unregistered speakers. Combined together, the two sets of audio and visual features ensure a very good speaker discrimination. These descriptors turn out to be particularly efficient when dealing with professionnally edited TV content. The live nature of the show and the imbalanced speech load among speakers make this task particularly difficult. In a future work, we will combine our approach with a speaker diarization system to further reduce the human intervention in the process of automatic speaker identification.

6. REFERENCES

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

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

[3] Gerald Friedland, Hayley Hung, and Chuohao Yeo, "Multimodal speaker diarization of real-world meetings using compressed-domain video features," in *International Conference on Acoustics, Speech and Signal Processing*, 2009.

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

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

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

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

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

[9] J. Shi and C. Tomasi, "Good features to track," in *Conference on Computer Vision and Pattern Recognition*, 1994.

[10] Bruce Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, 1981.

[11] B.S. Manjunath, Philippe Salembier, and Thomas Sikora, Eds., *Introduction to MPEG-7 - Multimedia Content Description Interface*, Wiley, 2002.

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

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