INTERVENANT CLASSIFICATION IN AN AUDIOVISUAL DOCUMENT

Jeremy Philippeau, Julien Pinquier and Philippe Joly
Institut de Recherche en Informatique de Toulouse, Equipe SAMoVA
UMR 5505 CNRS - INPT - UPS - UT1, 118 Route de Narbonne, 31062 Toulouse Cedex 9, France

Keywords: Intervenant, audiovisual indexation, multimodal descriptor, IN OUT OFF classification, multimedia modeling.

Abstract: This document deals with the definition of a new descriptor for audiovisual document indexing : the intervenant. We actually focus on its audiovisual localization, this is to say its place in an audiovisual sequence and its classification in 3 categories : IN, OUT or OFF. Based on the comparison of different analysis tools of both audio and video modes, we define a set of descriptors which can automatically be filled, potentially relevant to classify the intervenant localization. This decision is taken on the base of transition modeling between classes.

1 INTRODUCTION

A lot of works have already been done about automatic characterization of audiovisual contents, thanks to both audio and video descriptors, but most of the chosen orientations are actually improving exclusively audio-based systems with some video features (Automatic Speech Recognition (Potamianos et al., 2004) for example) or conversely (Kijak, 2003).

We studied that kind of consideration in order to create a new descriptor relevant enough to characterize an audiovisual content in an indexation framework. If we consider an intervenant, this is to say a speaking character localizable by its speech in an audiovisual sequence : we try to know if, at a T time, without any background knowledge on the type of the analyzed document, she is visible or not. Up till now, studies in that field are considering this : a speaker is IN when someone is detected on the screen during the locution, otherwise she is OUT. However, this arbitrary classification do not take into account the visible speech activity : the detected character on the screen is not necessary the talking one.

We want to precise this classification taking in account the visible aspect of the locution and create new classes of intervenant :

- visible speaking character is classified IN,
- invisible speaking character already or later filmed during its diction is classified OUT,
- speaking character never visible during its diction is classified OFF.

After having described the documents we worked on, we present which video and audio descriptors we have chosen to characterize an intervenant. We show some experiences related to them and, finally, we present the way we have used at the same time audio and video descriptors to instantiate this audiovisual descriptor.

2 APPLICATIVE CONTEXT

2.1 Corpus

For comparison matters, we chose to study sequences listed for the TRECVID2004 evaluation campaign (Kraaij et al., 2004). We also studied a french television game named ”Pyramide”. The low definition (352*264 at 29.97fps) of these videos and the relative bad quality of the frames (because of the MPEG encoding) is a sign of genericity of our tool. Speech is omnipresent and mainly uninterrupted in these documents, so we can process the speech signal as a mono-speaker one.

2.2 Audiovisual Segment

We define an audiovisual segment as a sequence in which an intervenant class remains stable. A segment

Philippeau J., Pinquier J. and Joly P. (2006).
INTERVENANT CLASSIFICATION IN AN AUDIOVISUAL DOCUMENT.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 185-188
DOI: 10.5220/0001570801850188

will then be defined between two boundaries corresponding to : a speaker change, a shot change, both of them or a silence. The accuracy rates stated in this paper have been computed on audiovisual segments extracted both from TRECVID2004 and Pyramide.

3 THE VIDEO POINT OF VIEW

3.1 Face Detection

A lot of works had been done about automatic face detection (quoted in (Jaffre and Joly, 2004)) and is based on several methods : on "low-level" characteristics (like color, texture or shape), on facial features detection (like the eyes, the nose or the mouth), or even on statistical approaches. The detector we used belongs to this last category : the Violat and Jones's face detector [1]. The analysis is performed frame by frame on all the considered segments, during the speech detection.

The decision of a face occurrence is taken when this face has been detected in at least 7 frames in a temporal window of 11 frames ((Jaffre and Joly, 2004)).

To know if a face is the same from a frame to another, we build a searching window around each detected face. If two faces are located in the same window have nearly the same proportions, then they are considered to be the same face.

This detector often "forgot" one or more faces on a whole segment. We so have to complete the missing ones. We have chosen to generate a non detected V_0 face by linear interpolation of the coordinates of V_1 (detected before V_0) and V_2 (detected after V_0), the two faces temporally nearest from V_0. This method gives us visually correct results. Wrong detections are relatively rare and partially evicted, thanks to the algorithm of the activity score calculation explained in section 3.2. A face presence detected during a whole segment on the video constitutes our first reliable descriptor.

Thanks to this, we obtain an accuracy rate for the detection of the IN intervenants of about 90.2%.

3.2 Lips Activity Analysis

We next look to the lips localization matter in order to quantify their activity. A lot of works have been done based on intrusive devices or/and on clean frames (well defined and high definition) in laboratory conditions (frontal films with constant illumination) (Potamianos et al., 1998). These methods are impossible to be implemented in our study. So we chose to localize them approximately, this is to say in the low third part of the face, between the second and the fourth fifth of the face width (figure 1). Besides the fact that this is a fast and easy implemented localization method, this always grasp lips, whenever the face is front or side presented.

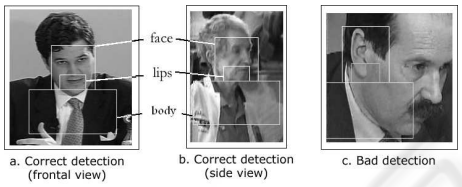

Figure 1: Face detector results.

To quantify lips activity, we proceed by pairs of frames to obtain a global result. So we consider two successive frames F_1 and F_2 containing the face of a same character. After lips localization, represented by the L(F_1) and L(F_2) regions, we build a searching window around L(F_1) and move L(F_2) in this zone. The matching and the value representing the difference between L(F_1) and L(F_2) pixels were both obtained while minimizing the Mean Square Error (MSE), normalized by the L2 size, on the luminance channel of the HLS color space. The mean of the MSE computed on all the video segments we considered gave us a quantitative value for the lips activity of the character. We called this the Lip Activity Rate.

We then consider a larger physical activity than lips, because a character not only moves its lips while speaking. So we calculate, with the same process, the Face Activity Rate and the Body Activity Rate (a rectangle placed just below the face) (figure 1):

We then built an activity score using a weighted sum of those rates. The accuracy rate of this score applied for speaker identification between two characters in a same segment or in two consecutive ones is about 95.7%.

4 AUDIO POINT OF VIEW

4.1 Cepstral Subtraction

The cepstral subtraction is usually used to remove noise coming from the recording source (microphone, telephonic channel...) on the speech signal (Mokbel et al., 1995). This is a relevant piece of information about the noise.

Mel Frequency Cepstral Coefficients (MFCC) are computed repositioning the signal spectrum in the

[1] http://www.intel.com/research/mrl/research/opencv/

Mel scale. To keep a relative independence towards the transmission's channel, it is usual to do a Cepstral Mean Subtraction (CMS) (Furui, 1981) from each MFCC. We went on processing the information contained into the MFCC's evolution between intervenant classes because of our previous researches on the cepstral subtraction study.

4.2 Descriptors Behavior

It is first necessary to list different possible configurations for transitions between classes and the expected behavior of the MFCC in each case, as the figure 2 illustrates it.

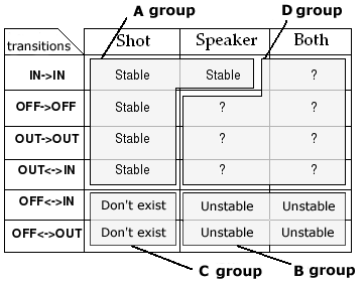

Figure 2: Expected behavior for the MFCC when used to characterize transitions between classes.

1. *The descriptors should characterize the stability of the audio environment in the case of a transition caused by a shot change if the same intervenant is speaking.* (**A group**). The particular case of a transition between 2 IN intervenants with a speaker change and without shot change has to be considered in this group.
2. *They hare expected to highlight an audio environment change in a transition between speakers evolving in different acoustic recording conditions.*(**B group**).
3. This is to notice that *some unusual transitions never happens* (**C group**). It concerns a change between an OFF voice and an IN or OUT one (and vice versa) without speaker change. That would imply that an OFF intervenant had been or would be in the field of view, refuting our previous definition.
4. The other cases (**D group**) *can not be useful if we only consider those descriptors*.

4.3 Experiments and Results

We vainly attempted to characterize in which acoustical environment the speaker was evolving. So we worked on the characterization of acoustical environment changes between two adjoining segments s^1 and s^2.

Given a 1 second segment s^k sampled at 16kHz, the cepstral analysis is computed on a 256 points windows with a 128 points covering. 125 vectors $y_i = (y_{i,1} \cdots y_{i,12})$ are obtained : they have got 12 dimensions (as much as the number of MFCC), with $i \in \{0, \cdots, 125\}$ as the vector index. If we process the 2 last seconds of the s^1 segment and the 2 first seconds of the s^2 segment, this gives us two collections of vectors, respectively $(y_1 \dots y_{250})$ and $(y_{251} \dots y_{500})$.

If we want to characterize a changing behavior of the MFCC between s^1 and s^2, we make the following suppositions :

- $(y_1 \dots y_{250})$ follows a multivariate Normal distribution $N(M^1, \Sigma^1)$ of dimension 12,
- $(y_{251} \dots y_{500})$ follows $N(M^2, \Sigma^2)$,
- $(y_1 \dots y_{500})$ follows $N(M^3, \Sigma^3)$.

If we consider that the MFCC are independent (Tianhao, 2006), we can state the two following hypothesis :

- hypothesis (h_1) : *there is an acoustical environment change between s^1 and s^2*. This is to say :

$$P(y_1 \dots y_{500}/h_1) = P(y_{1 \dots} y_{250}/N(M^1, \Sigma^1)) \cdot P(y_{251 \dots} y_{500}/N(M^2, \Sigma^2)) \quad (1)$$

- hypothesis (h_2) : *the acoustical environment of s^1 is the same than s^2*. This is to say :

$$P(y_1 \dots y_{500}/h_2) = \prod_{i=1}^{500} P(y_i/N(M^3, \Sigma^3)) \quad (2)$$

The hypothesis test is based on the likelihood ratio :

$$\Delta(s^1, s^2) = \frac{P(y_1 \cdots y_{500}/h_2)}{P(y_1 \cdots y_{500}/h_1)} \quad (3)$$

Setting a threshold θ, we can then make a decision for one hypothesis or the other. We noticed that, using the logarithmic form of this test, an experimental set threshold at $-68.5 * 10^{-3}$ worked well for 92.8% of the studied cases.

5 JOINED IMPLEMENTATION OF THE DESCRIPTORS

We decided to use these audio and video descriptors :

- **Presence**$_t \in \{yes, no\}$: character presence or absence during the segment t (section 3.1).

- $\Phi_{t,t+1}$: compared activity score between the two characters having the strongest Lips Activity Rate and living in the two adjoining segments t and t+1 (section 3.2).

- $\Delta_{t,t+1} \in \{yes, no\}$: stability or instability of the acoustical environment between the two adjoining segments t and t+1 (section 4.3).

- **Transition** $\in$ {S,L,S+L} : audio and/or video boundaries. S for Shot change detection, L for Speaker change detection and S+L for both change detection (figure 2 section 4.3).

We chose to create a 4 stated automaton : **IN**, **OUT**, **OFF**, and a **DOUBT** state used both as an initial state and as a temporary escape if the information extracted from the sequence is not sufficient to classify the intervenant (figure 3). We take in consideration that a state remains stable on each analyzed segment, and we define transition of this automaton like possibilities to explore each time a decision has to be taken, this is to say how the chosen descriptors were evolving.

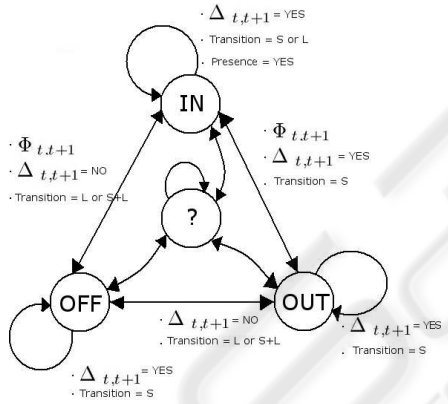

Figure 3: Automaton.

As far as there is no corpus where the ground truth take into account the IN/OUT/OFF classification, we have developed our own evaluation contentset of about 21 minutes. Here is a presentation of results we obtained with our automaton :

- if we consider **DOUBT** as a correct classification, we obtain an accuracy rate about 87.1%,

- if we consider **DOUBT** as a bad classification, we obtain an accuracy rate about 55.8%,

- if we do not take the doubt into account, this is to say if we only consider segments that are not classified as **DOUBT** cases, we obtain an accuracy rate about 82.6%.

- the automaton enters into **DOUBT** state in 24.2% of the cases,

6 CONCLUSION

We presented videos descriptors that allowed us to compare visual speech activity between intervenants from a segment to another, to determinate which character speaks inside a same segment, and finally to avoid the Viola and Jones's face detector deficiencies if it is used into a face following way.

We also showed that MFCC variations considered at the frontiers of the transitions between classes, represents a reliable descriptor to characterize change or stability between two acoustical environments.

Finally, these information joined in an automaton allowed us to create a reliable audiovisual descriptor to get an original IN, OUT and OFF classification for an intervenant.

REFERENCES

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. In *IEEE Trans. Acoust. Speech Signal Process.*, volume 29, pages 254–272.

Jaffre, G. and Joly, P. (2004). Costume: A new feature for automatic video content indexing. In *RIAO 2004*, pages 314–325, Avignon, France.

Kijak, E. (2003). *Structuration multimodale des videos de sports par modeles stochastiques*. PhD thesis, Universite de Rennes 1.

Kraaij, W., Smeaton, A., Over, P., and Arlandis, J. (2004). Trecvid 2004 - an introduction. In *Proceedings of the TRECVID 2004 Workshop*, pages 1–13, Gaithersburg, Maryland, USA.

Mokbel, C., Jouvet, D., and J., M. (1995). Blind equalization using adaptitive filtering for improving speech recognition over telephone. In *European Conference on Speech Communication and Technology*, pages 817–820, Madrid, Spain.

Potamianos, G., Graf, H., and Cosatto, E. (1998). An image transform approch for hmm based automatic lipreading. In *Proceedings of the Internationnal Conference on Image Processing*, volume 3, pages 173–177, Chicago.

Potamianos, G., Neti, C., Luettin, J., and Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. In Bailly, G., Vatikiotis-Bateson, E., and Perrier, P., editors, *Issues in Visual and Audio-Visual Speech Processing*. MIT Press.

Tianhao, L., Q.-J. F. (2006). Analyze perceptual adaptation to spectrally-shifted vowels with gmm technique. In *10th Annual Fred S. Grodins Graduate Research Symposium*, pages 120–121. USC School of Engineering.