ACADI: Automatic Character (in Audiovisual Document) Indexing

Frédéric GIANNI and Julien PINQUIER
IRIT - Université Paul Sabatier
118, route de Narbonne
31062 Toulouse Cedex 9, FRANCE
{gianni, pinquier}@irit.fr

Ewa KIJAK
IRISA
Campus de Beaulieu
35042 Rennes Cedex, FRANCE
Ewa.Kijak@irisa.fr

ABSTRACT

We propose a system which permits to describe and structure audiovisual documents without training, nor corpus knowledge, and to visualize with an interface the principal interventions. It posts the most significant person list of the processed documents (news, TV games, variety programs, film, etc.). A person will be considered as significant if she/he speaks or appears on the screen during a minimum time lapse. This list is then presented with representative labels of the character (face or/and sound extract for example). Thanks to this person list, it is possible to listen and/or to view all interventions of each character by clicking on the representation of the selected one. The system, developped in the framework of the Network of Excellence (NoE) MUSCLE, is based on a face detection tool and speaker and costume segmentation tools. The interface allows to visualize (and/or to listen) the only segments where the character of interest appears, without a priori knowledge. We also have the statistics over the speaking time and appearance time of each character.

1. INTRODUCTION

By analogy with textual documents which can be easier to handle (storage, data mining, accessibility, etc), multi-media document processing is only at its beginning. For example, finding a video which contains the first steps of Armstrong on the Moon (without prior information) is rather critical. It requires finding semantics from the video and/or the audio and to use them jointly. Many works were carried out on the audiovisual content characterization, and particularly on person detection. The majority of these studies are monomedia and allow the detection of a person either by his visual appearance in a frame (like a face) or by his voice. On the one hand, the image study is based on visual features, like face detection: many applications are described in [15]. On the other hand, the audio analysis is based on homogeneous segments, which follow a speaker segmentation via the Bayesian Information Criterion (BIC) for example [2]. Sometimes, the objective is to improve exclusively audio-based systems with video (like Automatic Speech Recognition (ASR) [13]) or sometimes, it is the opposite [8]. More recent works start video content analysis by integrating both acoustics and visual features, which are the two inseparable parts of a video. Thus, an adaptive speaker identification system that employs audiovisual cues which is based on a probabilistic framework is proposed in [10]. In [16], a rather similar approach based on confidence values is presented. In [9], a person identification system, which is able to identify characters in TV shows, is constructed based on two different strategies. In the first strategy, speaker identification is used to verify the face recognition result. The second strategy consists of using face recognition and tracking to supplement speaker identification results. The goal of these applications is to find, starting from voice and face models, in which sequences a given person appears, each face being associated with a voice. However, in some applications, this audio/video association is not available. For example, in our case, we do not have any prior model: models are computed on the fly, when the persons appear. That is why, in this paper, we present a framework for audio/video association in order to compute automatically these association models.

Our goal was not to improve descriptor or segmentation method qualities. We first combine the audio and video indexes in order to make the information brought by each of them more robust (this fusion step permits to reduce audio and/or video oversegmentations in order to make the best association between voices and appearing persons). Then, we implement an interface which allows to visualize and to listen the principal interventions of the significant persons.

First, we describe in section 2 our three segmentation systems (visage detection, costume detection and speaker segmentation) and we present the voice/image association algorithm. Then, in section 3 we make a description of the interface.

2. AUDIO/VIDEO SEGMENTATIONS AND ASSOCIATION

In this section, we present our three segmentation systems: speaker segmentation in audio, and face and costume detections in video. Each segmentation provides an index (audio or video) in the form of a list of 3-tuples, each 3-tuples being composed of $< begin\_segment, end\_segment, label >$.

We propose also a method that permits to associate the voice to the face/costume.

2.1 Speaker segmentation

We give here an overview of the method which is based, first, on a segmentation that consists in partitioning the audio stream into segments where each speech segment must be as long as possible and must contain ideally the speech of only one speaker. This segmentation is followed by a clustering step that consists in giving the same label to segments uttered by the same speaker. Ideally, each cluster corresponds to only one speaker and vice versa. This method, without any *a priori* knowledge, is more described in [5].

2.1.1 Segmentation

We use a new approach for speaker segmentation using a combination of Generalized Likelihood Ratio (GLR) [14] and the Bayesian Information Criterion (BIC) [2]. This method searches for points of acoustic changes and processes without preliminary speech detection. It follows four main steps: a splitting step, a most probable point detection step, a pre-adjustment step and a definitive change detection step.

Splitting step consists in splitting arbitrarily the audio stream into windows of two seconds and then, detecting the most probable point of change in every window.

In the *most probable point detection step*, the resulting points of change $P_1 \dots P_m$, separate mono-Gaussian models existing in every window. However, those models are not very representative because they are affected by a window with a fixed size and fixed boundaries. So, we repeat the first step using windows that are chosen as following: to detect a change point P_i, we use the window $[P_{i-1}, P_{i+1}]$. Thus, the new models will be quite close to Gaussian distributions.

Re-Adjustment step consists in repeating the second step several times until the repartition of change points is stabilized. At the end of this step, the points detected are the points of change in their local area.

At the *definitive change detection step*, BIC criterion is applied to select, from previous detected points, points that are effectively points of acoustic changes.

2.1.2 Clustering

Our clustering method is based on the work of Tsai [17] which utilizes the Eigen Vector Space Model (EVSM) with a hierarchical bottom-up clustering. Briefly, this method consists in modelling every segment by a Gaussian Mixture Model (GMM) and then extracting, from each model, a vector that is re-dimensioned using PCA. Finally, similarity between two segments is calculated using cosine formula. For a stronger merging criterion, we add the F_0 feature in order to prevent some false grouped segments. The hierarchical grouping is done with a complete linkage manner.

2.2 Visage detection

The face detection approach is based on a convolutional neural architecture, designed to detect and precisely localize highly variable face patterns, in complex real world images. The system automatically synthesizes simple problem-specific feature extractors from a training set of face and non face patterns, without making any assumptions or using any hand-made design concerning the features to extract or the areas of the face pattern to analyze.

The face detector is designed to locate multiple faces of 20x20 pixel minimum size, rotated up to 20 degrees in image plane and turned up to 60 degrees. Once trained, it acts like a fast pipeline of simple convolutions and sub sampling operations, applied at various scaled versions of the original image, to handle faces of different sizes. This pipeline does not require any costly local pre-processing. It performs automatic feature extraction and classification of the extracted features, in a single integrated scheme. The full process is implemented via a convolutional neural network architecture, which offers the advantage of being trained to automatically derive all parameters, governing feature extraction and classification. The proposed scheme provides very high detection rate with a particularly low level of false positives, demonstrated on difficult test sets, without requiring the use of multiple networks for handling difficult cases.

All details concerning the face detection system can be found in [6] and [3].

2.3 Costume detection

The first step of our costume detection is a face detection, so as to detect the different possible characters who are present in the current frame, and their approximate position and scale. Then, the costume of each character is extracted from the image according to the location and the scale of his face. There are many methods for face detection in literature (see [18] for a review). For our purpose, we have tested two methods: the classical face detector of OpenCV [12] and our visage detector (presented in the previous paragraph).

The costumes are extracted according to the localization and the scale of the detected faces. At the moment, we estimate the costume by the area under the face. The size of this area is proportional to the size of the face. In our examples, we used a width size of 2.3 times the size of the face, and for the height size a ratio of 2.6. These coefficients were chosen experimentally from training images.

However, frame by frame face localization introduces many false alarms, due to some noise present in the data. Only one false detection in a frame is enough to involve a false alarm on costume detection. In order to reduce these false detections, we must exploit the properties of a video sequence by using a temporal approach. For each frame, we detect all the faces independently using a static approach. Then, we take a temporal window (subsequence) of $2N + 1$ frames. For each candidate face, we count its number of occurrences in the N previous frames, and in the N next frames. Then, we keep a candidate face if it appears at least N_2 times in this subsequence. In our application, we took $N = 2$ (which leads to a subsequence of 5 frames) and $N_2 = 4$.

We consider that two detected faces in consecutive frames correspond to the same face if there are roughly at the same location. The position parameters may slightly vary considering camera works or character motions. So, a small variation of these parameters is allowed to take into account these effects. Moreover, to avoid false detections, we consider that two faces correspond to the same face if the costumes detected from these faces are also identical.

For more details, see [7].

2.4 Audio/video association method

Some recent methods use both audio and video cues for improving person identification ([1], for example). Their goal is to identify persons using both visual features and speech recognition with the assumption that the current voice corresponds to a visual feature in the frame is made. In real sequences, this hypothesis is often violated. It is very common to find sequences where the appearing persons do not speak during many frames (or many shots). Moreover, it is also usual that the current voice belongs to a person whose visual feature is not in the current frame.

In our application, we propose to compute co-occurrences between audio and video indexes, i.e. we determine the matching between voices and images. This approach is suitable to take into account the cases where the usual assumptions are not verified. The scale of audio and video indexes are different, making impossible a direct comparison of the two indexes. That is why we propose to use a common scale for audio and video indexes in order to be able to directly compare them. Using the video frequency, a frame by frame decomposition of each index is made.

Let n_a be the number of different voices in the audio index, and n_v the number of different visual persons in the video index. $\{A_i\}_{i=1\ldots n_a}$ and $\{V_j\}_{j=1\ldots n_v}$ are respectively the set of voices and visual features of all persons. To compute the intersection matrix between audio and video indexes, we go through the two indexes, frame by frame. For each frame, if the voice A_i is heard and the visual person V_j is present, the number of occurrences m_{ij} of the pair (A_i, V_j) is incremented. Then, we obtain the following matrix:

$$m = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \ldots & V_{n_v} \end{array} \\ \begin{pmatrix} m_{11} & m_{12} & \ldots & m_{1n_v} \\ m_{21} & m_{22} & \ldots & m_{2n_v} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n_a1} & m_{n_a2} & \ldots & m_{n_an_v} \end{pmatrix} \end{array} \quad (1)$$

In this matrix, the value m_{ij} means that in all the frames where the voice A_i is heard, the visual person V_j appears m_{ij} times. Conversely, in all the frames where the person V_j is present, the voice A_i is heard m_{ij} times. To carry out the fusion, we compute two new matrices, m_a and m_v, where the frame numbers are replaced with percentage by rows and by columns:

$$m_a = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \ldots & V_{n_v} \end{array} \\ \begin{pmatrix} f^a_{11} & f^a_{12} & \ldots & f^a_{1n_v} \\ \boxed{f^a_{21}} & \boxed{f^a_{22}} & \ldots & \boxed{f^a_{2n_v}} \\ \ldots & \ldots & \ldots & \ldots \\ f^a_{n_a1} & f^a_{n_a2} & \ldots & f^a_{n_an_v} \end{pmatrix} \end{array} 100\,\% \quad (2)$$

$$m_v = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \ldots & V_{n_v} \end{array} \\ \begin{pmatrix} f^v_{11} & \boxed{f^v_{12}} & \ldots & f^v_{1n_v} \\ f^v_{21} & \boxed{f^v_{22}} & \ldots & f^v_{2n_v} \\ \ldots & \ldots & \ldots & \ldots \\ f^v_{n_a1} & \boxed{f^v_{n_a2}} & \ldots & f^v_{n_an_v} \end{pmatrix} \\ 100\,\% \end{array} \quad (3)$$

Matrix m_a gives the probability density of each voice A_i, whereas m_v gives the one of each visual feature V_j. From these matrices, we define the fusion matrix F, by computing, for each pair (i, j), a fusion between f^a_{ij} and f^v_{ij} with a fusion operator like maximum, mean or product. If we note $C(A_i, V_j)$ the fusion coefficient between A_i and V_j, expression of matrix F is given by:

$$F = \begin{pmatrix} C(A_1, V_1) & \ldots & C(A_1, V_{n_v}) \\ C(A_2, V_1) & \ldots & C(A_2, V_{n_v}) \\ \ldots & \ldots & \ldots \\ C(A_{n_a}, V_1) & \ldots & C(A_{n_a}, V_{n_v}) \end{pmatrix} \quad (4)$$

This matrix F can be directly used to realize the association. When the number of voices and visual features is the same ($n_a = n_v$), it is read equally by rows or by columns. For instance, for each row i, we search the column j which provides the maximum value: then the voice A_i is automatically associated to the visual feature V_j.

For more details about this association, see [4].

3. ENHANCING THE USE OF DETECTION TOOL WITH AN INTERFACE

We propose a graphical user interface, called *ACADI* and implemented during a MUSCLE [11] showcase, to provide a tool in a verification-aided fashion of the segmentation results. The segmentation tool produce XML files as result (see table 1). Even if XML files are human readable, the verification process is a painful task if done manually, as these files could contain lot of data. A post-process to summarize results is needed and also a convenient video/audio player in order to check the segmentation coherency.

```
<?xml version="1.0"?>
<Segmentation>
  <Header>
    <Video>/local/CorpusVideo/10052005.mpg</Video>
  </Header>
  <Body>
    <SEGMENT ID="1030">
      <COSTUME ID="Character 001">
        <ECHANCRURE>echancre</ECHANCRURE>
        <MANCHES>inconnu</MANCHES>
        <TEXTURE>inconnu</TEXTURE>
      </COSTUME>
    </SEGMENT>
  </Body>
</Segmentation>
```

Table 1: Annotations results for a video (named 10052005.mpg), the annotations are given for each frame, here is shown one frame (numbered 1030).

The character list is displayed in sorted appearance duration order (the main character first).The user can select a character and see the video and audio segmentation statistics related to this character (cf. figure 1). The segmentation is also displayed and audio/video segments can be played. Those statistics present the appearance duration and the speaking duration for a character during the sequence. To provide portability among different systems we used the Gtkmm library for the GUI, and frame access is managed by an embedded ffmpeg player. This one also provides random frame access. For now we only support mpeg PS file format.

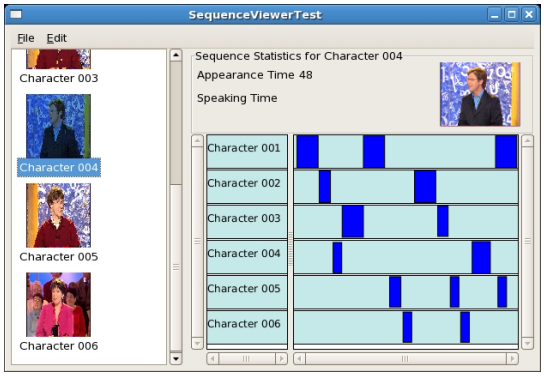

Figure 1: Main window of the interface. On the left the character list, upper right selected character statistics, bottom right segmentation results.

Oversegmentation, caused by resolution changes, for one actor can often happen. It results from an actor who appears, in the segmentation, several times with different labels. This can be solved manually through the interface. Renaming the label of several characters with the same name will merge the segments and thus quickly provide a segmentation without over-segmentation. The interface has been implemented following a modular conception, the audio/video segmentation tools are seen as "plug-in" for the interface. In this way several other treatments can be implemented independently, and used inside this interface, once they have been formatted as "plug-in".

4. CONCLUSIONS

We proposed in this paper a system for automatic association and visualization of audio and video indexes, within the framework of the NoE MUSCLE. This work is based on face and costume detections in video and speaker segmentation in audio. Without any *a priori* knowledge, nor training, we obtain good results ([4]) which permit to reveal the principal interventions of the most significant persons. Thanks to the interface, it is possible to listen and/or to view one (or all) intervention(s) of a selected character. We have also statistics over the speaking time and appearance time of each character: this application could be used to measure the audience of a TV program (for example an election).

5. REFERENCES

[1] A. Albiol, L. Torres, and E. Delp. Two are better than one: when audio comes to the rescue of video. In *Proceedings of the 5th European Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal, Apr. 2004.

[2] S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In *DARPA Speech Recognition Workshop*, 1998.

[3] M. Delakis and C. Garcia. Training Convolutional Filters for Robust Face Detection. In *Proceedings of the IEEE International Workshop of Neural Networks for Signal Processing*, pages 739–748, Toulouse, France, Sept. 1998.

[4] E. El Khoury, G. Jaffré, J. Pinquier, and C. Sénac. Association of Audio and Video Segmentations for Automatic Person Indexing. In *International Workshop on Content-Based Multimedia Indexing*, page to be published, Bordeaux, France, June 2007.

[5] E. El Khoury, C. Sénac, and R. André-obrecht. Speaker Diarization: Towards a More Robust and Portable System. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, page to appear, Honolulu, USA, Apr. 2007.

[6] C. Garcia and M. Delakis. A Neural Architecture for Fast and Robust Face Detection. In *Proceedings of the IEEE-IAPR International Conference on Pattern Recognition*, pages 40–43, Quebec, Canada, Aug. 2002.

[7] G. Jaffré. *Indexation de la vidéo par le costume*. PhD thesis, Université Toulouse III, 2005.

[8] E. Kijak, G. Gravier, L. Oisel, and P. Gros. Audiovisual Integration for Tennis Broadcast Structuring. In *International Workshop on Content-Based Multimedia Indexing*, pages 421–428, Rennes, France, Sept. 2003. GDR-PRC ISIS.

[9] D. Li, G. Wei, I. K. Sethi, and N. Dimitrova. Fusion of visual and audio features for person identification in real video. In *SPIE, the International Society for Optical Engineering*, volume 4315, pages 180–187, San Diego, USA, Aug. 2001.

[10] Y. Li, S. Narayanan, and C. Jay Kuo. Adaptive speaker identification with audiovisual cues for movie content analysis. *Pattern Recognition Letters*, 25(7):777–791, May 2004.

[11] Multimedia understanding through semantics, computation and learning. http://www.muscle-noe.org/.

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

[13] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-Visual Automatic Speech Recognition: An Overview. In G. Bailly, E. Vatikiotis-Bateson, and P. E. Perrier, editors, *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

[14] M. Siu, H. Gish, and R. Rohlicek. Segregation of speaker for speech recognition and speaker identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 873–876, Toronto, Canada, May 1991.

[15] C. Snoek and M. Worring. Multimodal Video Indexing: A Review of the State-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, Jan. 2005.

[16] C. Taskiran, A. Albiol, L. Torres, and E. J. Delp. Detection of unique people in news programs using multimodal shot clustering. In *Proceedings of the International Conference on Image Processing*, Singapore, Oct. 2004.

[17] W. Tsai, S. Cheng, Y. Chao, and H. Wang. Clustering speech utterances by the speaker using EigenVoice-Motivated vector space models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 725–728, Philadelphia, USA, Mar. 2005.

[18] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, Jan. 2002.