THE INDEXING OF PERSONS IN NEWS SEQUENCES USING AUDIO-VISUAL DATA

Alberto Albiol

Politechnic University of Valencia, Spain
e-mail: alalbiol@dcom.upv.es

Luis Torres†*, Edward J. Delp*‡

Technical University of Catalonia, Spain†
e-mail: luis@gps.tsc.upc.es
Purdue University, USA‡
e-mail: ace@ecn.purdue.edu

ABSTRACT

In this paper, we describe a video indexing system that automatically searches for a specific person in a news sequence. The proposed approach combines audio and video confidence values extracted from speaker and face recognition analysis. The system also incorporates a shot selection module that seeks for anchors, where the person on the scene will be likely speaking. The system has been extensively tested on several news sequences with very good recognition rates.

1. INTRODUCTION

The last years have been characterized by a great interest in digital video indexing, management and storage. The recent release of the MPEG-7 standard [1] and the large number of related applications are good evidence of this interest. However, new video analysis tools are still needed to reduce as much as possible the amount of human indexing work.

Within this framework, we present an automatic video indexing system which aims to locate and recognize specific people in news sequences. Many previous approaches to this problem used audio or visual information independently, where speaker [2] and face [3] recognition techniques were typically applied.

Although, relatively high recognition rates were achieved, recognition results can be further improved if both audio and visual information sources are combined. This is explained because degradations on the information sources are uncorrelated [4]. Hence, the research community is now putting its efforts in developing multi-modal systems that combine different information sources [5, 6]. However, most of the presented results have been obtained under relatively controlled environments.

Our indexing system which is described in section 2, also uses audio and visual information to derive shot-based speaker and face recognition confidences respectively. These confidences are introduced in a linear classifier to decide if a specific person appears in a particular shot speaking. The results presented in this paper have been carried out on raw broadcast TV news sequences stored in our video database [7]. The rest of the paper is organized as follows, sections 3-6 explain the different modules of the system sketched in Figure 1, and finally the results of the audio-visual indexing are presented in section 7.

2. SYSTEM OVERVIEW

This section describes the global block diagram of our system which is depicted in Figure 1. The *shot selection* module aims to detect the subset of anchor shots where the person on the scene will be also speaking. The importance of this block is twofold, first the computational burden is greatly reduced since the number of shots to be processed will be much smaller, and second, recognition results will be also benefited since fusion of audio and video information does not make sense if the speech does not correspond to the face on the scene. Next, audio and image information are processed in parallel and two confidence values are extracted for each selected shot. Finally, the *fusion* module is used to decide, based on the audio and image confidence values, if a particular person P_i is on the scene.

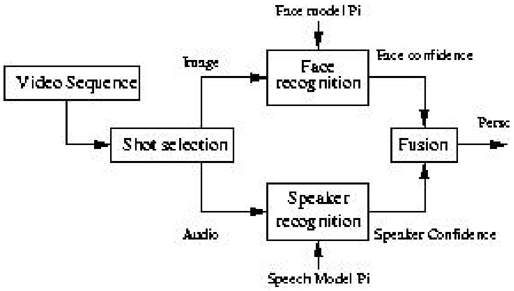

Fig. 1. System overview

This work has been partially supported by the grants TIC2001-0442, TIC2002-02469 of the Spanish Government and the US-Spain Joint Commission for Scientific and Technological Cooperation

3. SHOT SELECTION

We divide the people that appear in a news video sequence as two types: the first are news anchormen and reporters, the second are people that are the subject of news stories. The goal of the *shot selection* is to detect anchors shots in which these second type of people are also speaking. Our approach makes use of two simple clues, first the editing procedure used in news stories produces simultaneous transition on the audio and video cuts when an anchor shot is inserted [8], second the objects and camera motion in these shots is usually very low. Thus, our approach first searches for shots with simultaneous transitions in the audio and video tracks (sections 3.1 to 3.3) and then the shots with low motion activity are finally selected (section 3.4).

3.1. Audio segmentation

The goal of speaker segmentation is to locate all the boundaries between speakers in the audio signal. Earlier speaker segmentation systems were based on silence detection. However, these approaches require cooperative speakers which is not the case for broadcast news. More robust speaker segmentation approaches are based on speaker turn detection. Here, a two-step procedure is used, where the audio data is first split in an attempt to locate acoustic changes. Most of these acoustic changes will correspond to speaker turns. Then, a second step is used then to validate or discard these possible turns. In this paper, we use an off-the-self speaker turn detection algorithm called DISTBIC [9] to segment the audio data which has given satisfactory results.

3.2. Video segmentation

The objective of video segmentation is to segment the video sequence into parts called *shots* which correspond to a continuous set of frames taken from one camera. Cuts form the majority of shot transitions in news stories. Therefore, our system only focuses on the detection of cuts. Most of the techniques proposed for cut detection rely on the similarity between consecutive frames, and assume that a cut is produced when the similarity measurement is under some threshold. References to different similarity measurements and algorithms can be found in [8]. In this work, we use the *Mean Absolute Frame Difference* (MAFD) to measure the similarity between low resolution images obtained from the DC coefficients of the MPEG compressed video stream. The main problem of the MAFD measurement is that sometimes it is difficult to set a threshold because camera and object motion also increase the value of the MAFD. However, objects and camera motion normally last for more than one frame which produces wide pulses in the MAFD signal. We exploit this difference to distinguish motion and cuts in video. Hence, once the MAFD signal is obtained, we apply basic morphological operations, such as openings and closings [10], to reduce the contribution of object and camera motion.

We are well aware that more sophisticated video cut detection algorithms exist. However, this simple algorithm provides very good results as will be shown in section 7.

3.3. Audio and video correspondence

Once the audio and video are segmented, the next step is to locate those segments whose audio and video borders would ideally match. The problem at this step is that for real sequences, silence periods are usually located at the audio segment borders creating small inaccuracies. To overcome this problem, we define the overlap degree between the audio and video segments as:

$$overlap = \min\left\{\frac{l_{audio \cap video}}{l_{audio}}, \frac{l_{audio \cap video}}{l_{video}}\right\} \quad (1)$$

where l_{audio} and l_{video} are the time duration of the audio and video segments respectively and $l_{audio \cap video}$ is the duration of the intersection of a couple of audio and video segments. Finally, if the $overlap > 0.9$ then the audio and video segments are said to match.

3.4. Shot activity

As mentioned above, shots where the person that appears on the image is also speaking usually present low activity because the camera is placed on a fixed position focusing on the person who is speaking. This assumption is used here to further discard some of the selected shots obtained when the audio and video segments match. To measure the shot activity, we use the mean value of the MAFD signal within a shot. This measurement has the advantage of reusing the previously computed MAFD values. We have also tried other shot activity measurements [8] that use the MPEG motion vectors, however, the proposed measurement showed to be simpler and more robust.

4. SPEAKER RECOGNITION

Speaker recognition is formulated as a basic hypothesis test where given a speech segment S, we want to decide whether it was uttered by a specific person P_i or not. The optimum test is given by the log-likelihood ratio:

$$AC_{P_i} = \log\left\{\frac{p(S/P_i)}{p(S/BM)}\right\} \quad (2)$$

where $p(S/P_i)$ and $p(S/BM)$ are the probability density functions of the person P_i and the background model respectively [11]. These probability density functions are modeled using GMM's. The mixture models are built from feature vectors that consist of 12 mel-frequency cepstral coefficients and its corresponding delta coefficients. Cepstral vector coefficients are extracted each 17 ms. using a 34 ms. Hamming window. We also remove the silent frames and use cepstral mean subtraction to reduce linear channel distortions. In the training stage, we created a 32-GMM model for each person P_i using 2–3 min of clear speech. On the other hand, the background model is built from 1 hour of speech recorded from a variety of speakers extracted from our video database [7]. To make the background model as universal as possible, we took special care on the composition of the speaker's universe, trying to balance as much as possible the number of male/female utterances and different recording conditions. In the test stage we use AC_{P_i} as the confidence measurement that the utterance was spoken by P_i.

5. FACE DETECTION AND RECOGNITION

5.1. Face detection

Our face detection system [12] has three main blocks. First, skin detection is used to locate regions which potentially might contain a face based on the color information. However, skin detection will likely produce non-homogeneous skin-like regions containing more than one object. The second block, unsupervised segmentation, aims to segment the skin detected pixels into a set of connected homogeneous regions containing one object. The unsupervised segmentation is performed in two stages where chrominance and luminance information are used consecutively. At each stage, we use an algorithm that combines pixel and region color segmentation techniques. However, the unsupervised segmentation can sometimes further divide the face region. Thus, region merging is used to iteratively extract a set potential face candidates. Next, we use simple constraints regarding shape, color and texture to discard many false candidates. This results in a much smaller set of face candidates for each test image, although we still might have some erroneous candidates. These candidates will finally discarded at the face recognition stage.

5.2. Face recognition

Face recognition is based on Principal Component Analysis (PCA) which has been modified to cope with the video indexing application [13]. The main difference with the normal eigenface approach [14] is that we model each person P_i with a different set of eigenfacs that we call *self-eigenfaces* [3]. The self-eigenfaces are built from a set of frontal views of the person P_i where the location of the eyes is used to normalize the size and rotation of each training view. In the test phase, each face candidate is projected and reconstructed using a particular set of self-eigenfaces. Then, reconstruction error is used to measure the confidence that the identity of the face candidate is P_i. Notice that this approach is more robust to changes in brightness that the normal eigenface approach. The proposed measurement gives a confidence value when just one face candidate is evaluated. However, we need a shot-based confidence measurement. Let e_i be the minimum reconstruction error for all the face candidates of frame i. Then, we define:

$$FC_{P_i} = \text{median}\{e_0, e_1, e_2, \ldots, e_N\} \tag{3}$$

The value FC_{P_i} is used as the shot-based measurement that the face of the person P_i appears in a particular shot.

6. COMBINED AUDIOVISUAL RECOGNITION

The audio and video confidences AC_{P_i} and FC_{Pi} respectively, are used in the *fusion* block of Figure 1 to make the final multi-modal decision.

Figure 2 shows the scatter plot of the bi-dimensional feature vectors $C_{P_i} = (FC_{P_i}, AC_{Pi})$, where it can be clearly seen that true and false candidates are better separated in the two-dimensional space.

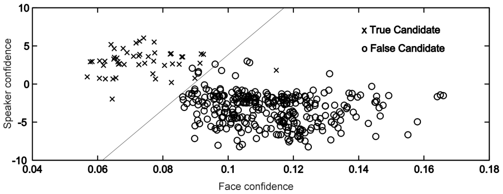

Fig. 2. Scatter plot of the face and speaker confidences.

From the structure of the data shown in the Figure 2, and following the Occam's Razor philosophy [15], we choose a simple linear classifier to separate true and false candidates. Thus, the linear discriminant function is found by MSE using the Pseudoinverse Matrix [15]. Figure 2 shows the decision boundary that is found using this method.

7. RESULTS

We carried out all the experiments on several TV news sequences stored on our video data base [7].

From these sequences we selected 10 people that appear frequently. This is important, because anchors in TV news stories usually do not last more than 15 seconds, and we need at least 2-3 minutes of audio (8-12 shots/person) to train each speaker model and we also need more appearances (5-10 shots/person) to test if we can find the selected candidate in different sequences.

The shot selection module achieves a detection rate of DR=90$\%$ (number of true anchors respect to the number of selected shots) with a false alarm of FAR=30$\%$ (number of false alarms respect to the number of selected shots). Almost all miss detections in the experiments are caused either by a miss in the audio transition, or because a more sophisticated edition process was used for a specific news story. Also, some misses are produced by flashlights which create false alarms in the video segmentation module. The high value for the FAR can be explained since shots where the TV anchorman appears are considered as a false alarm. Obviously, these scenes also fit our hypothesis and therefore we can not distinguish them with our simple approach. However, it should be noticed that the shot selection module allows to discard almost a 76% of the news sequence with minimal processing.

Speaker and face recognition is performed over the selected shots (about the 24% of the total time) for each selected candidate. Figure 3 shows the true positives and true negatives curves for each independent modality. These results are obtained by thresholding face or speaker confidences at different threshold values. It can be appreciated that for an equal error rate, face and speaker modalities achieve around a 93% and 91% of success respectively. These results are greatly improved if we combine the speaker and face confidences in the classifier described in section 6. Using the proposed classifier we obtain a 98% of true positives and a 99% of true negatives represents a big improvement respect to each separate modality. Most of the false negatives are produced because non-frontal views are available in a test shot, and thus a very low face confidence is obtained. On the other hand, background noise increases considerably the bias of the speaker confidence and this is probably the reason for most of the false positives and some false negatives.

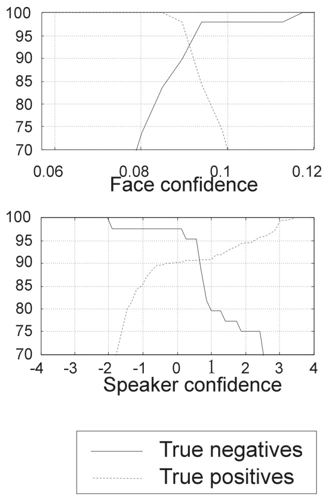

Fig. 3. Recognition results using only each modality separately.

8. REFERENCES

[1] ISO/IEC 15938-3: 2002, *Inf. Technology – Multimedia content description interface– Part 3: Visual*.

[2] S.Tsekeridou and I.Pitas, ,"Speaker identification for audio indexing applications " in *Int. Conf. on Telecom.*, Porto Carras, Halkidiki, Greece, June 1998.

[3] L. Torres and J. Vilá, "Automatic face recognition for video indexing applications," *Pat. rec.*, vol. 35, no. 3, Dec. 2001.

[4] C. Neti and Andrew Senior, "Audio-visual speaker recognition for video broadcast news," in *DARPA HUB4 Workshop*, Washington D.C., March 1999.

[5] M. Viswanathan, H.S. M. Beigi, and F. Maali, "Information access using speech, speaker and face recognition," in *ICME*, New York, August 2000.

[6] S. Ben-Yacoub, J. Luttin, K. Jonson, J. Matas, and J. Kittler, "Audio-visual person verification," in *CVPR*, Los Alamitos (CA), June 1999.

[7] C. Taskiran, J. Chen, A. Albiol, L. Torres, C. A. Bouman, and E.J. Delp, "Vibe: A compressed video database structured for video active browsing and search," *IEEE trans. on Multimedia*, vol. (Accepted for publication), 2003.

[8] A. Albiol, L. Torres, and E. J. Delp, "Combining audio and video for video sequence indexing applications," in *ICME*, Laussane, Swithzerland, August 2002.

[9] P. Delacourt and C. J. Wellekens, "Distbic: A speaker-based segmentation for audio indexing," *Speech comm.*, vol. 32, no. 1-2, Sept. 2000.

[10] Jean Serra, *Image analysis and mathematical morphology*, Academic Press, London, 1982.

[11] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speakers models," *Speech Communications*, vol. 17, 1995.

[12] A. Albiol, L. Torres, and E. J. Delp, "A simple and efficient face detection algorithm for video database applications," in *ICIP*, Vancouver, Canada, Sept. 2000.

[13] E. Acosta, L. Torres, and A. Albiol, "An automatic face detection and recognition system for video indexing applications," in *ICASSP*, Orlando, FL, May 2002.

[14] M.A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *CVPR*, June 1991.

[15] R. D. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Willey-interscience, 2nd ed. edition, 2001.