COMBINING AUDIO AND VIDEO FOR VIDEO SEQUENCE INDEXING APPLICATIONS

Alberto Albiol

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

Luis Torres†, *Edward J. Delp*‡

Politechnic 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 address the problem of detecting shots of subjects that are interviewed in news sequences. This is useful since usually these kinds of scenes contain important and reusable information that can be used for other news programs. In a previous paper, we presented a technique based on a priori knowledge of the editing techniques used in news sequences which allowed a fast search of news stories. In this paper we present a new shot descriptor technique which improves the previous search results by using a simple, yet efficient algorithm, based on the information contained in consecutive frames. Results are provided which prove the validity of the approach.

1. INTRODUCTION

Recent advances in digital video coding have enabled the creation of a large number of digital videobases. However, due to the huge amount of audio visual data, special attention has to be paid to the design of systems used to access and retrieve information from these databases. Audio and video indexing play a key role in this process. The main objective of the indexing process is to assign labels to the audio visual data in order to describe its content. The data explosion problem can be alleviated if, before using force brute analysis tools on the entire video sequence, the relevant parts of the sequence are detected.

In particular, we want to locate those parts where the audio corresponds to the face (if any) present in the image. Based on experiments, we have estimated that approximately 85% of the time in broadcast news the audio and video do not match with respect to who is speaking, while in the remaining 15% the voice and face match, which justifies the interest of this work.

In this paper, some improvements with respect to our previous work are presented [1]. In that work, we took advantage of the editing techniques used in news sequences to discard a large portion of the news sequence. Now, those results are further improved using a new shot descriptor based on the shot activity.

In section 2 the main elements of a news sequence are described, this will give us the key idea to the presented approach. Sections 3 to 5 review the analysis tools which were used in [1]. In section 6 the new descriptor presented in this paper is examined and in section 7 we present our experimental results which will be also compared with the results of [1]. Finally section 8 draws some conclusions.

2. ELEMENTS OF A NEWS SEQUENCE

In this paper we divide the people that appear in a news video sequence as two types: the first are news anchors and reporters, the second type are people that are the subject of news stories. The goal of this paper is to detect shots that contain the second type of people in which these people are also speaking. News sequences are usually composed of the following elements:

1. Graphics and animations.

2. Shots where the news anchor or the reporter are speaking and they are either in the scene or narrating another scene.

3. News stories where either the anchor or reporter is narrating the scene or the person that is the subject of the story is speaking.

In this paper we are interested in locating scenes where people that are the subject of news stories are speaking. The editing procedure for this type of scene can be summarized as follows:

1. The anchor or reporter records his/her audio narration.

This work was partially supported by the US-Spain Joint Commission for Scientific and Technological Cooperation and the grant TIC1999-1361-CE from the Spanish Government.

2. If the news story is going to be inserted, then its audio and video are inserted together, creating a simultaneous audio and video cut.

3. If the reporter continues with speaking or more stories are to be inserted, the two first steps are repeated.

4. Finally, images are added where the reporter narrates the scene. Usually several shots are inserted for each audio segment.

In [1] we made use of this editing procedure to detect news stories by examining the matching of audio and video cuts. However, shots where the person who is speaking also appears on the image are usually characterized by a low activity. Now, this property will be also used to discard many false alarms. Finally, once the relevant segments are detected, more sophisticated (and computationally expensive) techniques, such as speaker recognition and/or face recognition, can be used to semantically index the sequence.

3. AUDIO SEGMENTATION

The goal of speaker segmentation is to locate all the boundaries between speakers in the audio signal. Some speaker segmentation systems are based on silence detection [2]. These systems rely on the assumption that utterances of different people are separated by significant silences. However, reliable systems would require cooperative speakers which is not the case for broadcast news. Other speaker segmentation approaches are based on speaker turn detection. These systems aim to segment the audio data into homogeneous segments containing one speaker only. Usually, 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. The second step is used then to validate or discard these possible turns. The techniques used for the first step can be classified into three different groups: phone decoding [3], hypothesis testing [4] and distance-based segmentation [5, 6]. Distance-based segmentation approaches have proved to be more robust for non-collaborative speaker segmentation, and thus, in this paper we will use an algorithm called DISTBIC (see [6] for details) to segment the audio data. DISTBIC is also a two-step segmentation technique, which is inspired on the BIC algorithm developed by IBM [5]. In the first step the distance between adjacent windows is obtained every 100ms. This result in a distance signal $d(t)$. In our implementation we use the symmetrical Kullback-Leibler distance [7]. Then, the significant peaks of $d(t)$ are considered as turn candidates. In the second step the turn candidates are validated using the ΔBIC criteria [5]. To that end, the acoustic vectors of adjacent segments are modeled separately using Gaussian models. The model of the union of the acoustic vectors of both segments is also computed, and then, the ΔBIC criteria is used to check if the likelihood of the union is greater than the likelihood of both segments individually. In the case that the likelihood of the union is greater the turn point is discarded. Otherwise the turn point is validated.

4. VIDEO SEGMENTATION

A great variety of techniques have been proposed in the literature for detection of transitions in video sequences. In this paper, we will focus only on cuts because gradual transitions are much less frequent in news sequences since they need much time to complete the edition process, and usually, in the context of news edition, time is a very important matter since the final product has to be broadcast as soon as becomes available.

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. Depending on the distance measurements, algorithms can be grouped broadly into three categories: pixel, block and histogram-based systems. Pixel based systems typically measure the difference between consecutive frames using the *mean absolute frame difference (MAFD)* [8]. Block based systems [9], aim to reduce the contribution of camera and objects motion. To that end, every frame is divided into a set of blocks that are compared to their corresponding blocks in the next frame. In this work, techniques that need motion vectors are not used because of the higher computational requirements. Finally, histogram-based systems [10] try to further reduce the negative effect of object and camera motion by comparing the histograms of successive images. The main problem of histogram techniques is that two images with completely different content may have similar histograms. In this work, low resolution images obtained from the DC coefficients of the MPEG compressed video stream will be used to compute the MAFD measurement. Figure 1.a shows the MAFD signal for a typical news sequence. The narrow and high peaks in the Figure correspond to scene transitions. On the other hand, objects and camera motion normally last for more than one frame which produces wide pulses in the MAFD signal (between the peaks). This difference is exploited here to distinguish motion and cuts in video. To that end, once the MAFD(n) signal is obtained, basic morphological operations, such as openings and closings [11], are applied to reduce the contribution of object and camera motion. Figure 1.b shows the MAFD(n) signal after taking the residue of an opening. It can be seen how the contribution of camera and object motion has been considerably reduced, and then, the choice of a suitable threshold becomes much simpler. It can also be noticed that the last peak of

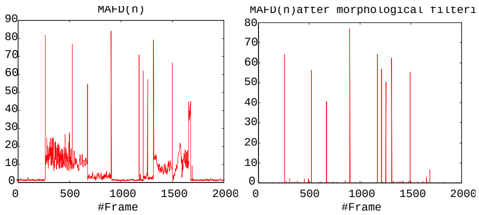

Fig. 1. Example of cut detection in a test news sequence. Left MAFD(n), right MAFD(n) after morphological processing

MAFD(n) in Fig. 1.a has been also removed by the morphological filtering because the peak corresponds to a gradual transition, and then, it lasts for several frames. However, this will not be a problem for our final goal, since although gradual transitions appear in news sequences, they are rarely found in news stories.

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.

5. AUDIO AND VIDEO CORRESPONDENCE

Once the audio and video segments are located, the objective is to find the correspondence between them. Figure 2.a shows the case that we are trying to find, i.e. the audio and video segments overlap. However, for real sequences the borders of audio and video segments do not overlap, as shown in figure 2.b. This is due mainly because silence periods are usually located in the audio segment borders creating a small inaccuracy. Figure 2.c shows an example of the typical situation for news stories, where a long audio segment coexists with short video segments. Given an audio segment in the time interval $[t_{min1}, t_{max1}]$ and a video segment defined in the interval $[t_{min2}, t_{max2}]$, the intersection interval is defined as:

$$[t_{min\cap}, t_{max\cap}] = [\max(t_{min1}, t_{min2}), \min(t_{max1}, t_{max2})] \tag{1}$$

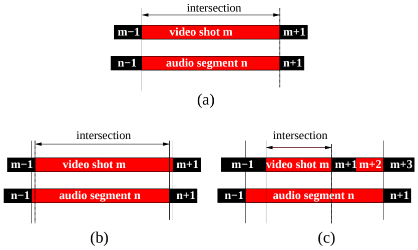

Fig. 2. (a) Audio and video borders match exactly. (b) Audio and video borders almost match (c) Audio segment contains several video shots

then, if $(t_{max\cap} - t_{min\cap}) > 0$ for a pair of audio and video segments, we define the overlap as:

$$overlap = \min\left\{\frac{(t_{max\cap} - t_{min\cap})}{(t_{max1} - t_{min1})}, \frac{(t_{max\cap} - t_{min\cap})}{(t_{max2} - t_{min2})}\right\} \tag{2}$$

If $overlap > 0.9$ then the audio and video segments are said to match and a new index entry is created.

6. SHOT ACTIVITY

Examining the matching of audio and video segments using the previously presented techniques, allows to discard a large portion of the video sequence in order to locate and recognize a speaker (as will be shown in section 7). However, results can be improved if additional information is used. As mentioned in section 2, 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 only the matching of audio and video is used. In order to measure the shot activity, we tested two shot activity measurements. The first measurement was based on the MAFD and the second measurement was based on the encoded MPEG motion vectors. We found that although both measurements were highly correlated, the second one provided noisier results, and therefore, the first measurement is used in this work. The measurement is defined as the mean value of MAFD(n) within the shot:

$$SA_1 = \sum_{n=ns}^{ne} \frac{MAFD(n)}{(ne - ns + 1)} \tag{3}$$

where ns and ne are the initial and final frame numbers of the analyzed shot.

7. RESULTS

The previous algorithms have been tested on several 30 minutes news sequences recorded directly from the Spanish television. In order to evaluate the proposed algorithms the following parameters have been defined.
Detection Rate (DR):

$$DR = 100 \times \frac{\text{num. detected elements}}{\text{num. actual elements}} \tag{4}$$

False alarm rate (FAR):

$$FAR = 100 \times \frac{\text{num. false alarms}}{\text{num. actual elements + num. false alarms}} \tag{5}$$

also, an additional parameter has been defined to evaluate the detection of people speaking in news stories.

Selected Time (ST) is defined as:

$$ST = \frac{\text{total duration of the selected shots}}{\text{Sequence duration}} \qquad (6)$$

For the audio segmentation using the DISTBIC, we have obtained a MDR=11$\%$ with a FAR=20$\%$, which also confirms the results presented in [6]. Most of the missed detections are usually short segments (less than 4 seconds), since the algorithm is tuned to detected longer speaker segments. The high FAR value is explained since usually the reporter voice is recorded over background noise which greatly varies as the background scenes change. The cut detection procedure described in section 4 was able to detect all the cuts in our test sequences (DR=100$\%$) with a small false alarm (FAR=2$\%$). It is important to note that all the false alarms are produced by flashlights.

In [1] we achieved a DR=94$\%$ with a FAR=41$\%$ for the detection of people speaking in news sequences, taking into account only the matching of audio and video segments. These results are improved here using the shot activity descriptor as discussed in section 6. In this case, a DR=90$\%$ with a FAR=30$\%$ is achieved, which proves the usefulness of the shot activity measurement. Almost all miss detections in both experiments are caused either by a miss in the audio transition, or because a more sophisticated edition process was used for a specific report. Then, the audio and video cuts do not match. Also, some misses are produced by flashlights which create false cuts as mentioned before. On the other side, the high value for the FAR on both experiments, can be explained since shots where a TV news anchors appears are considered as false alarm in our results. Obviously, these scenes also fit our hypothesis and therefore this approach can not distinguish them. However, it can be noticed how using the shot activity descriptor allows to greatly reduce the FAR value. Finally, the Selection Time (ST) obtained in [1] was ST=31$\%$ without using the shot activity measurement which reduces to ST=24$\%$ when the measurement is used. These results show how it is possible to discard a 76$\%$ of the news sequence with minimal processing.

8. CONCLUSIONS

We have presented a fast algorithm to detect people speaking in news stories without the need of analyzing in detail the whole sequence. The proposed algorithm is based on our previous work [1] which has been improved here using the shot activity descriptor. Once the segments of interest are located, more sophisticated analysis tools can be used such as speaker or face recognition techniques.

9. REFERENCES

[1] A. Albiol, L. Torres, and E. J. Delp, “A fast anchor person searching scheme in news sequences,” in *3rd Int. conf. on audio and video-based biometric person authentication*, Halmstad, Sweden, Jun. 2001, pp. 366–371.

[2] M. Nishida and Y. Ariki, “Speaker indexing for news, articles, debates and drama in broadcasted TV programs,” in *IEEE Int. Conf. on Multimedia, Computing and Systems*, Florence, Italy, Jun. 1999, pp. 466–471.

[3] D. Liu and F. Kubala, “Fast speaker change detection for broadcast news transcription and indexing,” in *Proc. ESCA Eurospeech’99*, Budapest, Hungary, Sept. 1999, vol. 3, pp. 1031–1034.

[4] S. Wegmann, P. Zhan, and L. Gillick, “Progress in broadcast news transcription at dragon systems,” in *Proc. of Int. Conf. on Acoustics Speech and Signal Processing*, Phoenix, AZ, May 1999, vol. 1, pp. 33–36.

[5] S. S. Chen and P. S. Gopalakrishnan, “Speaker environment and channel change detection and clustering via de bayesian information criterion,” in *DARPA Speech Recognition Workshop*, Landsdowne, VA, Feb. 1998, pp. 127–132.

[6] P. Delacourt and C. J. Wellekens, “Distbic: A speaker-based segmentation for audio indexing,” *Speech communication*, vol. 32, no. 1-2, pp. 111–127, Sept. 2000.

[7] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, “Automatic segmentation, classification and clustering of broadcast news audio,” in *Proc. of the DARPA speech recog. workshop*, Chantilly, VA, Feb. 1997, pp. 97–99.

[8] A. Albiol, V. Naranjo, and J. Angulo, “Low complexity cut detection in the presence of flicker,” in *Proc. of the Int. Conf. on Image Processing*, Vancouver, Canada, Oct. 2000, pp. 957–1000.

[9] Behzad Shahraray, “Scene change detection and content-based sampling of video sequences,” in *Proc. of SPIE Conf. on Digital Video Compression: Algorithms and Technologies*, San Jose, CA, Feb. 1995, vol. 2419, pp. 2–13.

[10] B. Gargi, S. Oswald, D. Kosiba, S. Devadiga, and R. Kasturi, “Evaluation of video sequence indexing and hierarchical video indexing,” in *Proc. SPIE Conf. Storage and Retrieval in image and video databases*, San Jose, Ca, Feb. 1995, pp. 1522–1530.

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