SEMANTIC ANALYSIS OF MULTIMEDIAL INFORMATION USING BOTH AUDIO AND VISUAL CLUES

A. Lukáč

University of Žilina, Faculty of Electrical Engineering, Department of Telecommunications and Multimedia
Univerzitna 1, 010 26 Žilina, Slovak Republic, lukac@fel.uniza.sk

Summary Nowadays, there is a lot of information in databases (text, audio/video form, etc.). It is important to be able to describe this data for better orientation in them. It is necessary to apply audio/video properties, which are used for metadata management, segmenting the document into semantically meaningful units, classifying each unit into a predefined scene type, indexing, summarizing the document for efficient retrieval and browsing. Data can be used for system that automatically searches for a specific person in a sequence also for special video sequences.

Audio/video properties are presented by descriptors and description schemes. There are many features that can be used to characterize multimedial signals. We can analyze audio and video sequences jointly or considered them completely separately. Our aim is oriented to possibilities of combining multimedial features. Focus is direct into discussion programs, because there are more decisions how to combine audio features with video sequences.

1. INTRODUCTION

Three types of modalities audiovisual information are preferred in these times [2]. We consider the following three information channels or modalities, within a video document [1]:

•**Visual modality:** contains everything, either naturally or artificially created, that can be seen in the video document.

•**Auditory modality:** contains the speech, music, and environmental sounds, which can be heard in the video document.

•**Textual modality:** contains textual resources that describe the content of the video document [1].

In this article we will deal only with 2 modalities - auditory and visual modality. Both modalities and the methods of their combinations are large section (detailed analyses, analysis of variance or level of features). However, ways of analyses and recognizing are similar and final stage is again semantically appraisal of multimedial document in collaboration with one of the above-mentioned modalities.

2. APPROACH OVERVIEW

The first step of multimedial analyze is signal segmentation into two independent streams (audio and video stream) [1]. We will assess these two modalities separately [3]. The audio/video analysis and its results are shown in Fig.1. It depends on fact, what is important to recognize. Convenient signals properties i.e. features are extracted after deeper analysis.

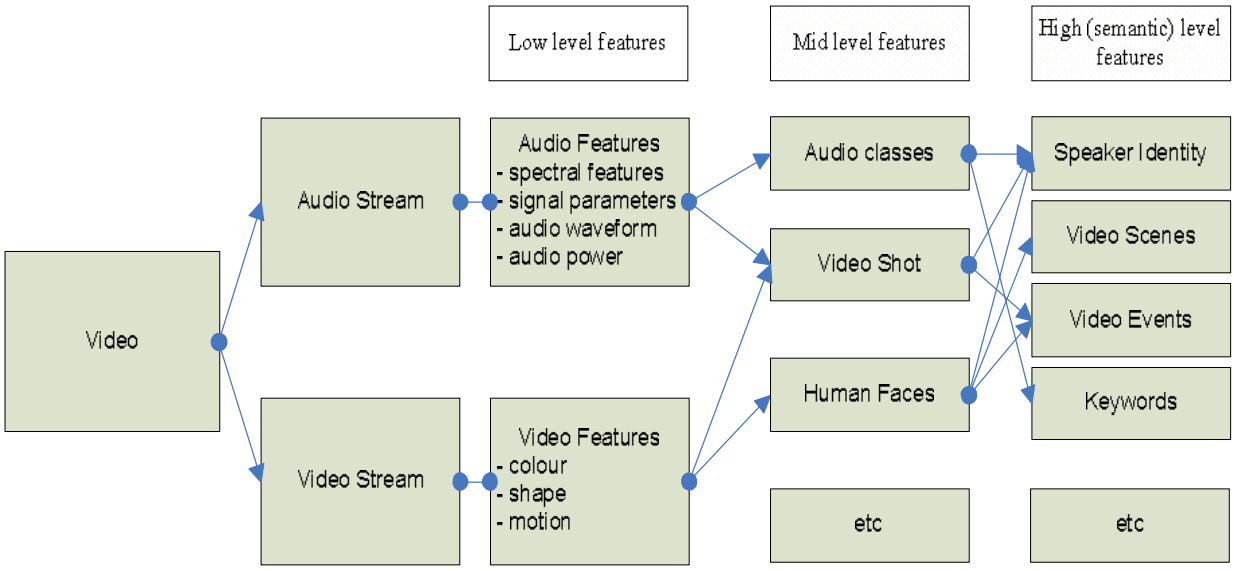

Fig. 1. Multimedial signal analysis

In general we can divide obtained features by three classes:

low level features – they define general importance in describing audio,

mid level features – from low level features, combination of several modalities,
high (semantic) level features – describing multimedial information in term of meaning and content.

We suggested possibilities of combination of features mentioned above to better facilities assignation and for following operations: parametrize, segmentation, indexing and retrieval.

3. THE PROPOSED SCHEME

We attempt audio recognition improvement of anchor person within video indexing. Long audio segment as speech of anchorperson (speaker) with associated short video shots is the example of person recognition. This detection can be used not only for person detection but for story boundary detection too.

Speaker or anchor person is the person who is in TV news most frequently begins and ends integrated unit (story). Scheme of proposed system is shows in Fig. 2. It describes two separate analyses. We use database TRECvid 2005. We classify audio into classes of which speech is the most essential.

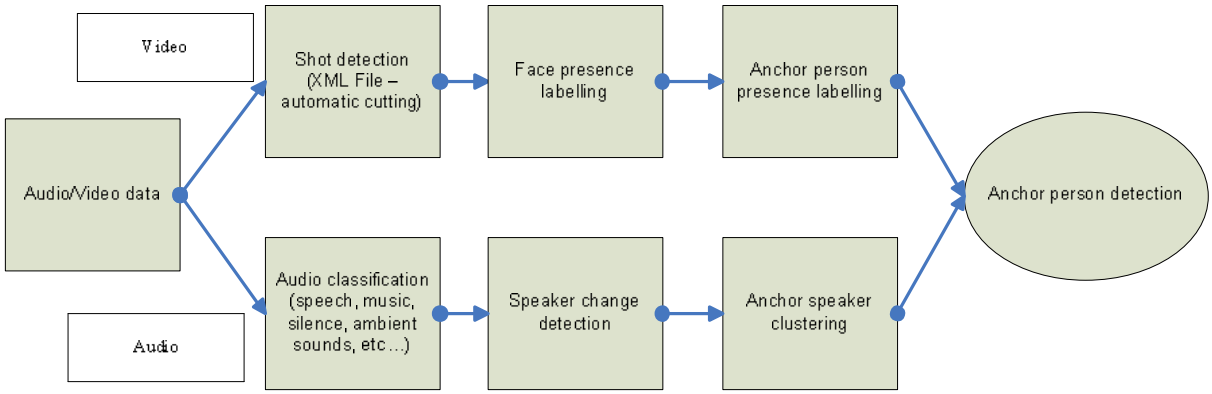

Fig. 2. Diagram of planned system

3.1. AUDIO FEATURES

We extracted an audio signal from video program. At first, we resample each audio to 22,05kHz / mono / 16bit and then transform it to the 12 Mel Frequency Cepstal Coefficients (MFCC). The 30 ms window with 20 ms shifting was applied. So each observation consists of 12 dimensional MFCC vector.

Anchorperson is modelled by the Gaussian Mixture model (GMM). All sill pauses from the training observations are removed. In the segmentation task we used automatically detected video shot boundaries and applied them as audio segments. We were looking for an anchorperson within a shot level. So classification task is pretty deterministic, shot may contain speech of an anchorperson or not. We considered that the minimum length of speech of anchorperson must be at least 2 seconds. Then we computated a logarithmical likelihoods (LL) for each shot using GMM of the anchorperson. So each observation within shot is represented by a scalar value of LL. The main problem is in the different size of each segment. For different shot lengths, summation of the LL can be confusing. Also mean value of LL may confuse classificator in the case when a shot contains speech of an additional non - anchorperson speaker.

In the second step, we used LL as a kind of new 1 dimensional observations. We modelled these observations with the 3 components Gaussian mixture (GMM-3) for each shot. Gaussian consists of mean vectors, variance vectors and weights. In the 1 dimensional case are all these parameters a scalar values. Each shot was also modelled by the 9 parameters (3 mixture weights + 3 scalar mean values + 3 scalar variances). These new parameters are independent of shot length and they are used in the final classification.

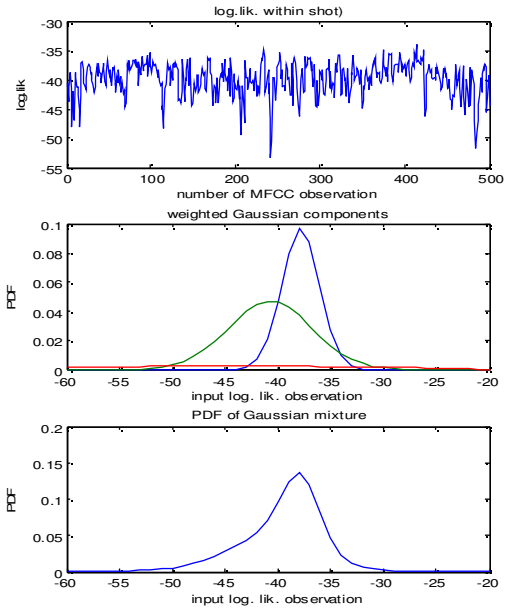

Fig. 3. One shot as GMM-3

3.2. VIDEO FEATURES

We have manually processed video shot. Video shot's boundaries are loaded from XML files, which are into TRECvid databases.

Video annotation is divided into two main sections:

- face detection,
- anchor person face detection,

We define parameters of detection for face detection. Face is one mouth, one nose and two eyes (definition by TRECvid).

We supplemented, that face has 5% from full screen and the duration is at least two second. If video shot satisfy above mentioned parameter, then it will be annotated by value 1. Value 0,5 was assigned to video shot, because we didn't know decision. Video shot had value 0, when it was outside parameters.

In final analysis, we use only one parameter namely anchor person face presence value. Decision fusion elaborates ten parameters, nine from audio analysis and one from video annotation. We try to connect these parameters with advisable method (stochastic optimizing algorithm, SVM – Support vector machine, etc…) and with them improve already existing audio recognition with the aid of video annotation [3].

4. CONCLUSION

This study has presented an approach for audio-video analysis. We have reviewed several important aspects of multimedia content analysis using both audio and visual information [2]. We have formulated an anchorperson detection technique by separate analysis of the audio and video signals. The overall results are formed by decision fusion. The TV news database TRECvid 2005 is chosen as the experimental materials and they are analyzed as point from different points of view: faces detection, anchor person detection, speech detection, silence detection, etc... . Aim of this experiment is the audio recognition improvement of anchor person detection with video annotation.

ACKNOWLEDGEMENTS

This paper has been supported by the VEGA grant agency no. 1/4066/07, “New systems and principles of semantic description and retrieval of multimedia content”.

REFERENCES

[1] ALBIOL, A., TORRES, L., DELP, E.J., Combining audio and video for video sequence indexing applications, ISBN 0-7803-7304-9, Y. 2002, 353 s.

[2] WANG, Y., LIU, Z., HUANG, J.-C., Multimedia content analysis, IEEE Signal Processing Magazine, nov. 2000, 12 pp.

[3] HALLER, M., KIM, H.-G., SIKORA,T., Audiovisual anchorperson detection for topic – oriented navigation in broadcast news, ICME 2006, 1817 s.

[4] LI, YING, NARAYANAN, *S.,* JAY KUO,*C.-C.,* Content-Based Movie Analysis and Indexing
Based on Audiovisual Cues, IEEE Transactions on Circuits and Systems for Video Technology, VOL. 14, 2004, 1073 s.

[5] KIM, H.-G., MOREAU, N., Sikora, T., MPEG-7 audio and beyond audio content indexing and retrieval, ISBN 0-470-09334

[6] KOSCH, H., Distributed multimedia database technologies supported MPEG-7 and by MPEG-21, ISBN 0-8493-1854-8