An Audio/Video Analysis Mechanism for Web Indexing

Marco Furini
Department of Computer Science
Via Bellini 25/G
Alessandria, Italy
furini@mfn.unipmn.it

Marco Aragone
Department of Computer Science
Via Bellini 25/G
Alessandria, Italy

ABSTRACT

The high availability of video streams is making necessary mechanisms for indexing such contents in the Web world. In this paper we focus on news programs and we propose a mechanism that integrates low and high level video features to provide a high level semantic description. A color/luminance analysis is coupled with audio analysis to provide a better identification of all the video segments that compose the video stream. Each video segment is subject to speech detection and is described through MPEG7 so that the resulting metadata description can be used to index the video stream. An experimental evaluation shows the benefits of integrating audio and video analysis.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Design, Experimentation

Keywords

Contents Indexing, Shot Boundary Detection, Video Indexing, MPEG7-DDL, Automatic Speech Recognition

1. INTRODUCTION

The advances in networking and multimedia technologies have contributed to the proliferation of video streams into the Web. The number of these videos increases day after day, and video indexing is becoming a must. Due to volume of video materials a manual indexing approach appears unreasonable and hence both commercial and academic researchers are studying a way to automatically describe and index video streams.

Most of the proposals are in the signal processing domain and are mainly based in low-level feature extraction, like color and luminance investigation, motion vector analysis and image texture [2, 5, 6]; others use the textual (for instance the Close Caption) and/or audio information to provide a high level semantic video description [4, 1, 3]. Regardless of the approach, the rough idea is to use low-level information to divide the video into several video segments, so that a more precise video description can be provided.

In this work we focus on video news programs and we propose a mechanism based on low-level video features extraction and on high level audio analysis to provide a video description suitable for web indexing. The novelty of our approach is in the integration of audio and video analysis to provide a more realistic video segmentation, by finding out video editing points. In fact, in news programs, most of the editing points should be not treated as cuts (points where a video segment begins or ends), as a news is usually coupled with a video clip that contains several editing points. If only low-level information are used, all the video editing points are treated as cuts causing a single news to be split into several (meaningless) videosegments. Our idea is to consider the audio energy associated to a cut. If silence is detected, the cut is usually the beginning (or the ending) of a news, otherwise it is very likely a simple editing point.

After identifying all the cuts, and hence all the video segments, a speech detector is applied and a high-level semantic description is given in MPEG7 using the audio transcript and the timing properties. In this way, search engines can easily index the video stream.

It is worth noting that our approach differs from the one proposed by *Hayashi et al.* [1]; Their idea is to perform a video segmentation by analyzing the sole audio stream. Conversely, our approach considers both audio and video analysis to provide a more realistic video segmentation.

2. OUR PROPOSAL

Three main steps are involved in our mechanism: i) Audio/Video analysis to identify cuts, silent audio segments and to perform speech analysis; ii) Videosegmentation to produce video segments; iii) Content description to produce an MPEG7-DDL document for web indexing purposes.

2.1 Audio/Video Analysis

The audio analysis is in charge of the speech detection. Since we are focusing on the indexing of news programs, we consider the adoption of an automatic speech recognition system trained for analyzing audio news and hence we based our audio analysis on the Sphinx Open Source Project [7].

The video analysis is in charge of identifying video cuts. A cut usually represents a camera break or an editing point and can be detected by analyzing the low-level features of any single frame. Different techniques are possible: histogram changes, edges extraction, chromatic scaling. Here, for each video frame i we combine the YUV components

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

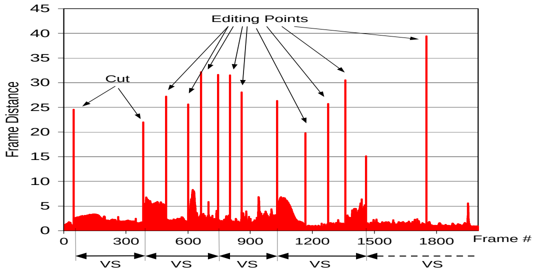

Figure 1: Histogram difference of YUV components.

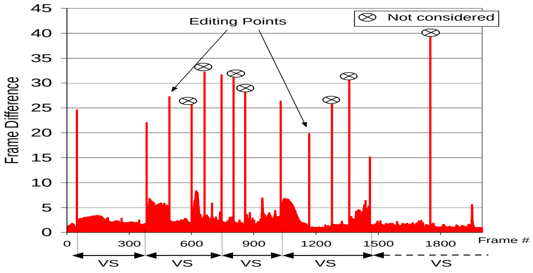

Figure 2: Cuts detected by analyzing YUV components and audio.

considering that human vision is more sensitive to brightness than to colors: $YUV(i) = 0.5Y(i){+}0.25U(i){+}0.25V(i)$. The perceptual difference between two consecutive frames is computed with $PD(i) = YUV(i) - YUV(i-1)$, and if this difference is above a pre-defined threshold (here equal to 10, obtained from analyzing several videos), a cut is detected.

2.2 Video Segmentation

The video frames between two consecutive cuts compose a video segment and our mechanism integrates audio and video analysis to find out editing points among all the identified cuts. In particular, for each identified cut, we consider a small number of frames that precedes the cut and an equal number of frames that follows the cut. An audio investigation is done to check whether these consecutive frames (the one with the cut and the others) are associated with audio information or with silence. If they are associated with audio, it is very likely that the cut is an editing point and hence it is not a video segment bound. After identifying all the video segments, a speech detector algorithm is applied to each video segment in order to transcript the audio content.

2.3 Content Description

To facilitate searching and indexing of videos, each video segment is described with a text-based description. MPEG-7 is used to produce the metadata video description.

3. EXPERIMENTAL RESULTS

Figure 1 shows results obtained from analyzing a video news. A perceptual difference between frames (based on YUV analysis) is carried out; If it goes above 10, a cut is detected. It is to note that most of the cuts are editing points. The real video segments (VS) are reported at the bottom of Figure 1. Nine editing points are treated as cuts.

By integrating the audio analysis, it is possible to find out editing points. Figure 2 shows that our mechanism finds out seven (out of nine) editing points and only two are erroneously treated as cuts. Each video segment is then passed to the speech detector and the audio information are then transcribed using MPEG7. Table 1 is an example of the video content description.

4. CONCLUSIONS

In this paper we presented a mechanism for web indexing, designed for video news programs. The novelty consists in the integration of audio and video analysis to provide a high-level semantic video description. The video segmentation process uses the audio analysis to find out editing points among all the video cuts identified by the video analysis. Since editing points are very common in video news, our mechanism provides a better video segmentation; each video segment is analyzed through a speech detector and a high level semantic description is given using MPEG7. Currently, we are working on identifying fade-in/fade-out cuts.

Table 1: MPEG7-DDL VideoSegment Description

```
<Mpeg7>
<Video>
 ...
  <VideoSegment>
   <TextAnnotation><FreeTextAnnotation>
    GARDNER AND OTHER MILITARY ANALYST WE
    SPOKE TO BELIEVE A CONVENTIONAL ATTACK
    ... ... ... ... ... ... ... ... ... ...
    STRIKING DISTANCE
  </FreeTextAnnotation></TextAnnotation>
  <MediaTime>
   <MediaTimePoint>0:03:17.480</MediaTimePoint>
   <MediaDuration>00:00:14.800</MediaDuration>
  </MediaTime>
 </VideoSegment>
 ...
 </Video>
</mpeg7>
```

5. REFERENCES

[1] Y.Hayashi et al., Speech-based and Video-Supported Indexing of Multimedia Broadcast News, Proc. of SIGIR03, July 28-August 1, 2003, Toronto, Canada.

[2] J.Zhou, X.P. Zhang, A Web-Enabled Video Indexing System, Proc. of MIR04, October, 15-16 2004 , New York, ACM

[3] M.R.Naphade, T. S. Huang, Extracting Semantics From Audiovisual Content: The Final Frontier in Multimedia Retrieval, IEEE Transaction on Neural Networks, Vol.13, No.4, July 2002.

[4] N.Babaguchi, Y. Kawai, T.Kitahashi, Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration, IEEE Transaction on Multimedia, Vol. 4, No. 1, March 2002

[5] J. Yuan, L.Y.Duan, Q.Tian, C.Xu, Fast and Robust Short Video Clip Search Using an Index Sructure, Proc. of MIR04, October 15-16 2004, New York, USA

[6] L.Y. Duan, M. Xu, Q.Tian, C.S. Xu, J.S.Jin, A Unified Framework for Semantic Shot Classification in Sports Video, IEEE Transactions of Multimedia, Vol. 7, No. 6, Dec. 2005

[7] Sphinx Project. Carnegie Mellon University, PittsBurgh. [online]. http://cmusphinx.sourceforge.net