A Web-based Medical Video Indexing Environment

Engin Mendi
Computer Science Department
University of Arkansas at Little Rock
Arkansas, USA
esmendi@ualr.edu

Coskun Bayrak
Computer Science Department
University of Arkansas at Little Rock
Arkansas, USA
cxbayrak@ualr.edu

Abstract—Video indexing is an important process in video content management systems. In this paper, we extend our previous work by presenting a new web-based indexing environment for medical videos. The proposed system provides two ways to describe a video structure: Element Generation and Description Display for Browsing. After shot boundaries are detected, HTML documents combined with SMIL (Synchronized Multimedia Integration Language) elements are generated. SMIL timing and media synchronization elements and attributes manage the location of shot boundaries and allow the users to browse and preview each shot within the video. The system enables browsing the entire content of the video displaying video descriptions for each segmented shot using an embedded media player and Timed Text XML. Through the use of the Timed Text XML file, textual information of the content in every shot is displayed on the media player.

Keywords-component; Shot boundary detection, video indexing, SMIL, Timed Text XML.

I. INTRODUCTION

The growth of archived video material has made indexing and annotating the information crucial. Since manual indexing and annotating the video material are both computationally expensive and time consuming, automated systems that can efficiently perform these processes are needed. Video segmentation is the central process for automatic video indexing, browsing and retrieval systems. The first step of video segmentation is to divide a video sequence into shots, where each has the same content representing a sequence of frames.

Shot boundary detection has been active research area. Many automatic shot boundary detection techniques have been developed over the past few decades, targeting to break up the video into meaningful sub-segments by using pixel comparison between consecutive frames [1], edge changes [2], grey-scale and RGB color histogram features [3] [4].

In recent years, parallel to the growth of the digital video with the success of the Internet, efficient and effective video indexing and browsing techniques are key components to video content-based retrieval systems. In order to provide web users reliable and convenient access to visual information advanced technologies must be developed for indexing and browsing the vast amount of video available on the web. In this work, we describe a new web-based indexing environment for medical videos. The web environment consists of integration of HTML with SMIL and an embedded media player using Timed Text XML. HTML web documents combined with SMIL manages the structure of video content and gives the possibility to browse every segmented shot. Embedded video player using Timed Text XML enables browsing the indexed video by captioning the textual content of the shots. The rest of this paper is organized as follows. In the next section we give a brief overview of the previous work of shot boundary detection algorithms. In section 3, we present our web-based indexing environment. Finally, the conclusions of this paper are summarized in Section 4.

II. PREVIOUS WORK

In video processing, the segmentation video stream is an important problem with many applications. The video sequence is divided into meaningful segments (shots) which are the basic elements of indexing. Shot boundary detection targets breaking up the video into meaningful sub-segments [5]. We detected shot boundaries by using three approaches [5, 6, 7, 8]: 1) color histogram differencing, 2) self-similarity analysis, and 3) salient region detection and structural similarity. Although a brief overview will be given here, more detailed information including experimental results can be obtained in [5, 6, 7, 8].

A. Color Histogram Differences

The presented shot boundary detection method is based on absolute differences of HSV color histograms between frames. HSV color space in which colors are described by hue, saturation and value is more convenient for human perception. Moreover, features extracted in HSV space can capture the distinct characteristics of computer graphics better than of RGB space [9]. Therefore, RGB color space is converted to HSV space. Since the human visual system is unable to distinguish whole possible colors, HSV color space can be quantized to reduce the data - all while preserving the colors. It can also reduce the overall computation effort. Since human color perception is more tolerant to saturation and value deviations, the quantization should preserve more hue levels when compared to saturation and value [10]. A color quantization is done using 256 colors (16 levels for H channel, 4 levels for S channel and 4 levels for V channel). Finally, using Equation (1), the differences of HSV histograms between consecutive frames are computed to determine the peaks that represent the shot boundaries [6, 7].
where \( h_k \) is the color histogram of the \( k \)th frame of the video sequence with \( N \) bins.

An example is given in Figure 1. Peaks at frames 241 and 379 represent cuts where large discontinuities occur between histograms.

\[
D(k) = \sum_{i=1}^{N} |h_k(i) - h_{k-1}(i)|
\]

\[ \text{(1)} \]

B. Self-Similarity Analysis

HSV feature vectors of the frames within the video data are visualized with a two-dimensional matrix by applying a similarity metric \([7]\). The self-similarity matrix \( S \) is defined as:

\[
S = s_{ij} = s(v_i, v_j)
\]

(2)

where \( s \) is a similarity measure between feature vectors \( v_j \) and \( v_j \) containing any feature of \( i \)th and \( j \)th frames of the video sequence, respectively. \( S \) can be visualized by the comparison of feature vectors using a distance measure.

The self-similarity matrix depicted in Figure 2 is the portion of the full matrix corresponding to frames 150–310. There are dissolves from frames 197–203 and 291–297.

C. Salient Region Detection and Structural Similarity

The human visual system tends to seek visually important regions and actions in image or video sequences. Visual attention analysis is the process of selecting and getting visual information based on saliency in the image itself and simulating the human visual system by producing saliency maps of the attended regions in image or video sequences. A visual attention model is based on the extended frequency-tuned saliency model \([11]\), to extract shot boundaries \([8]\). Saliency maps are produced from the color and luminance features of the image. Saliency map \( S \) is formulated for the video frame \( F \) as follows:

\[
S(x,y) = \| F_{xy} - F_{xy}^* \|
\]

(3)

\( F_{xy} \) is the mean video frame feature vector, \( F_{xy}^* \) is the corresponding pixel vector value of the blurred frame from the original frame and \( \| \cdot \| \) is the Euclidean distance. Each pixel location is the Lab color space vector, i.e. \([L,a,b]^T\).

To compare the visual content of the video frames, the structural similarity (SSIM) index is used \([8]\). SSIM(\( F_1, F_2 \)) combines three comparison components, namely luminance- \( l(F_1, F_2) \), contrast- \( c(F_1, F_2) \) and structure- \( s(F_1, F_2) \) where \( F_1 \) and \( F_2 \) are to represent two video frames \([12]\).

\[
\text{SSIM}(F_1, F_2) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}
\]

(4)

where \( \mu_x \), \( \mu_y \), \( \sigma_x \), \( \sigma_y \) and \( \sigma_{xy} \) are means of \( F_1 \) and \( F_2 \), variances of \( F_1 \) and \( F_2 \) and correlation coefficient between \( F_1 \) and \( F_2 \). \( K_1 \) and \( K_2 \) are scalar constants that \( K_1, K_2 \ll 1 \) and \( L \) is the dynamic range of the pixel values. \( C_1 = (K_1L)^2 \), \( C_2 = (K_2L)^2 \) and \( C_3 = \frac{C_2}{2} \).

SSIM mainly focuses on local information and does not take global saliency features into consideration. In the human visual system - where the importance of a visual event increases with the information content and decreases with the perceptual uncertainty \([13]\) - we incorporated saliency map as a weighting function into the SSIM index. So saliency factors can be instated into the similarity measure. The weighting function is:

\[
w(F_1, F_2) = \| F_{xy} - F_{xy}^* \|
\]

(4)

We define saliency-based SSIM as S-SSIM as follows:
\[
S \cdot \text{SSIM} = \frac{\sum F_1 \sum F_2 w(F_1, F_2) \text{SSIM}(F_1, F_2)}{\sum F_1 \sum F_2 w(F_1, F_2)} \tag{5}
\]

Measuring S-SSIM index over time reveals that S-SSIM is higher between frames within a shot than instantaneous changes from one shot to another (cut) and during shot transitions where digital effects used. We determine shots, cuts and transitions by applying a threshold to the S-SSIM index of the video frames. The idea is summarized in Figure 3. Between frames 197–203 and 291–297, dissolving transition effects are used, where the current shot fades out, while the next shot fades in. As the S-SSIM indexes of these transitions are lower than threshold, they are determined as shot boundaries.

Figure 3. A part of neurosurgical video illustrating a dissolving transition effect, where visual content of the frames slowly change. The threshold is higher than the S-SSIM indexes of transition frames and lower than of shot frames.

III. WEB-BASED INDEXING ENVIRONMENT

Shot boundary algorithms have been applied to medical video data and a web environment has been developed to automate the indexing of these videos. The proposed web system offers two features: 1) SMIL-based Element Generation for Web Documentation and 2) Description Display for Browsing using TimedText XML and Media Player.

A. SMIL-based Element Generation

Indexing the web documents has been executed through the use of SMIL. SMIL is a W3C recommended XML markup language designed for describing multimedia presentations combining audio, video, text, and graphics in real-time [14]. It can be used to describe temporal and spatial specifications of a multimedia presentation with hypermedia support. In our system, we inserted SMIL elements supporting content control, linking, media objects, timing and synchronization into HTML documents. A simple example of a HTML file combined with SMIL elements is depicted in Figure 4.

After shot boundaries have been identified, an HTML document is generated using SMIL timing and media synchronization elements and attributes.

Figure 4. A simple HTML file combined with SMIL elements

The location of shot boundaries are managed by SMIL basic timing attributes controlling the beginning and the end of elements in XML structure. A sample of such document is given in Figure 7. The generated web file allows the users to browse and preview each shot within the video.

B. Description Display for Browsing

The proposed web system also allows users to browse medical video combined with captioning video descriptions for each segmented shots using Timed Text XML. Timed Text is the term of textual information synchronized with video and audio using the W3C’s Timed Text specification. The W3C Timed Text specification is intended to develop an XML-based format used for the representation of streamed text in synchrony with other timed media, like audio and video [15]. In our system, as shown in Figure 6, users are able to browse the video with a web browser using embedded media player. JW Player [16] is chosen as a media player. Descriptions of detected shots are embedded into Timed Text XML. A sample of Timed Text XML is given in Figure 7.

The accessibility plug-in enables the media player to display video descriptions in every shot through the use of the Timed Text XML file. Figure 8 shows first 4 shots of a medical video with their descriptions incorporated into XML tags.
Figure 6. Embedded video player

```xml
<tt xml:lang="en" xmlns="http://www.w3.org/2006/10/ttaf1" xmlns:tts="http://www.w3.org/2006/10/ttaf1#style"> <head></head> <body> <div xml:id="captions"> <p begin="0:00:00.000" end="0:00:07.800">Shot 1: Demonstration of the removal of a brain tumor that doctors believe is causing epileptic seizures in a middle-aged man.</p> <p begin="0:00:08.120" end="0:00:11.560">Shot 2: An MRI showed what appears to be a glioma (tumor) near a part of the brain that controls muscle movement, called the motor strip.</p> </div> </body> </tt>
```

Figure 7. A simple Timed Text XML

IV. CONCLUSION

In this paper, we presented a new web-based indexing environment for medical videos. The proposed indexing system generates HTML documents integrated with SMIL and consists of an embedded video player using Timed Text XML. After shot boundaries are detected, the location of the shot segments are managed by SMIL enabling users to browse and preview each shot within the video. The embedded video player using Timed Text XML allows shot-based video browsing through the presentation of textual description.

REFERENCES


Figure 8. Content descriptions of shots displayed on the player using Timed Text XML