VideoAnnEx: IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning

Ching-Yung Lin, Belle L. Tseng and John R. Smith
IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532

ABSTRACT
The VideoAnnEx, a.k.a. IBM MPEG-7 Annotation Tool, assists authors in the task of annotating video sequences with MPEG-7 metadata. Each shot in the video sequence can be annotated with static scene descriptions, key object descriptions, event descriptions, and other lexicon sets. The annotated descriptions are associated with each video shot or regions in the keyframes, and are stored as MPEG-7 XML files. The tool allows customized lexicons to be created, saved, downloaded, and updated. It can also be used to generate storyboards by saving all keyframes using JPEG. An annotation-learning component is encompassed in the tool to speed up the annotation task.

1. INTRODUCTION
The growing amount of digital video is driving the need for more effective methods for indexing, searching, and retrieving of video based on its content. While recent advances in content analysis, feature extraction, and classification are improving capabilities for effectively searching and filtering digital video content, how to reliably and efficiently index multimedia data is still a challenging issue. Besides, in order to learn audio-visual concept models, supervised learning machines also require labels being associated with training videos.

We implemented a VideoAnnEx MPEG-7 annotation tool to allow authors semi-automatically annotate video content with semantic descriptions [1][2]. It is one of the first MPEG-7 annotation tools being made publicly available. The tool explores a number of interesting capabilities including automatic shot detection, key-frame selection, automatic label propagation to similar shots, and importing, editing, and customizing of ontology and controlled term lists.

Given the lexicon and video shot boundaries, visual annotations can be assigned to each shot by a combination of label prediction and human interaction. Labels can be associated to a shot or a region on the keyframe. Regions can be manually selected from the keyframe or injected from the segmentation module. Annotation of a video is executed shot by shot without permuting their time order, which we consider an important factor for human annotators because of the time-dependent semantic meanings in videos. Label prediction utilizes clustering on the keyframes of video shots in the video corpus or within a video. By the time a shot is being annotated, the system predicts its labels by propagating the labels from the last shot in time within the same cluster. Annotator can accept these predicted labels or select new labels from the hierarchical controlled-term lists. All the annotation results and descriptions of ontology are stored as MPEG-7 XML files.

Some other MPEG-7 annotation tools are available in public. MovieTool is a developed by Ricoh for creating video content descriptions conforming to MPEG-7 syntax interactively [3]. While the use of MPEG-7 in VideoAnnEx is transparent to the users, MovieTool requires users being familiar with MPEG-7 and editing XML files directly using MPEG-7 tags. The KnowCenter released a MPEG-7 based annotation and retrieval tool for digital photos [4]. IBM Multimedia Mining Project released a Multimodal Annotation Tool, which is derived from an old version of VideoAnnEx with an additional manual audio signal segmentation function [5].

2. SYSTEM DESCRIPTION
Four major components describe the annotation process and are depicted in Figure 1. First, video segmentation is performed to cut up the video sequence into smaller video units. Second, semantic lexicon is defined in order to regulate the video content descriptions. Third, an annotator labels the video segments with the semantic. An automatic annotation-learning component can be used to speed up the annotation task. Fourth, the MPEG-7 descriptions of the annotation process are directly outputted from the VideoAnnEx. The goal of the video annotation is to categorize the semantic content of each video unit or regions in the keyframes and output the MPEG-7 XML description file. The following four subsections describe these components in further detail.

2.1 Video Shot Segmentation
A short video clip can be simply annotated by describing its content in its entirety. However when the video is longer, annotation of its content can benefit from segmenting the video into smaller units. A video shot is defined as a continuous camera-captured segment of a scene, and is usually well defined for most video content. Given the shot boundaries, the annotations are assigned for each video shot.

The VideoAnnEx Shot Segmentation component is based on the multiple timescale differencing of the color histogram. This algorithm uses sampled RGB color histograms in the I- and motion histograms in the P-frames of video sequences. Heuristic rules are designed to make the algorithms robust to flashes and

Given the shot boundaries, the annotations are assigned for each video shot.

Figure 1: Four Major Components of the VideoAnnEx
2.2 Ontology Editor and Controlled Item List

Given the segmentation of video content into video shots, the second step is to define the semantic lexicon in which to label the shots. A video shot can fundamentally be described by three attributes. The first is the background surrounding of where the shot was captured by the camera, which is referred to as the static scene. The second attribute is the collection of significant subjects involved in the shot sequence, which is referred to as the key object. Lastly, the third attribute is the corresponding action taken by some of the key objects, which is referred to as the event. These three types of lexicon define the vocabulary for our video content.

Using the defined vocabulary for static scenes, key objects, and events, the lexicon is imported into VideoAnnEx. Note that the set of lexicon as well as the category attributes are dependent on the application, and can be easily generated and modified using VideoAnnEx. Details of this ontology-editing component can be seen in [2].

2.3 Annotation Learning

Annotation Learning is a characteristic that helps speed up the annotation speed. Right before the user annotates a video shot, predicted labels would have been shown on the “keyword” field of the VideoAnnEx. The prediction functionality on the current public-release version of VideoAnnEx v. 1.5 propagates labels from the visually most similar annotated shot. While the VideoAnnEx opens a video, a background thread calculates the feature-space distances between shots in the video. A distance combining both the feature space distance and the temporal space difference of shots are calculated to decide the visually closest shot. This propagation mechanism has been shown quite effective and helpful in speeding up the annotation task. A new mechanism of incorporating pre-trained models is under developing.

3.4 MPEG-7 Video Segment Description

The ISO standardized MPEG-7 defines the compatible scheme and language to represent semantic meaning of multimedia content. Our MPEG-7 output is the Video Segment Description Scheme. In MPEG-7, each video shot is defined as a Video Segment. Furthermore, the embedded <SpatioTemporal Description Scheme> tag allows us to specify the region location and the corresponding text annotation in a key frame. An example of the output XML file can be found at [6].

3. Graphical User Interface

The VideoAnnEx is divided into four graphical sections as illustrated in Figure 2. On the upper right-hand corner of the tool is the Video Playback window with shot information. On the upper left-hand corner of the tool is the Shot Annotation with a key frame image display. On the bottom portion of the tool is two different Views Panel of the annotation preview. A fourth component, not shown in Figure 2, is the Region Annotation pop-up window for specifying annotated regions. These four sections provide interactivity to assist authors of the annotation tool.

The Video Playback window displays the opened MPEG video sequence. As the video is played back in the display window, the current shot information is given as well. The Shot Annotation module displays the defined semantic lexicons and the key frame window. The key frame is a representative image of the video shot segment, and thus offer an instantaneous recap of the whole video shot. This is the region where the annotator selects the descriptions for the video segment. The Views Panel displays two different previews of representative images of the video. The Frames in the Shot shows all the I-frames as representative images of the current video shot, while the Shots in the Video view (as in the bottom of Figure 2) shows all the key frames of each shot as representative images over the entire video. As the annotator labels each shot, the descriptions are displayed below the corresponding key frames in the Shots in the Video view. Furthermore after the MPEG-7 descriptions are saved into an XML file, anyone can load and review these files at a later time by previewing the annotations at this views panel. The Region Annotation window allows the author to associate a rectangular region with a labeled text annotation. After the text annotations are identified on the Shot Annotation window, each description can be associated with a corresponding region on the selected key frame of that shot. More details are shown in [2][6].

4. SUMMARY

We built an MPEG-7 Annotation Tool to facilitate multimedia annotation tasks for general users. Use of MPEG-7 is transparent to users so that no prior knowledge on MPEG-7 is required. Various features, such as shot segmentation, ontology editing, storyboard generation, etc., are provided. In the next phase, we are developing a new version for incorporating multimedia annotation task in a distributed environment.

5. REFERENCES