ClassMiner: Mining medical video for scalable skimming and summarization

Xingquan Zhu[1], Jianping Fan[2], Mohand-Said Hacid[3], Ahmed K. Elmagarmid[1]

[1]Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907, USA

[2]Dept. of Computer Science
University of North Carolina
Charlotte, NC 28223, USA

[3]University Claude Bernard
Lyon 1, Lyon
France

Keywords
Video data mining, scene detection, event detection, scalable skimming, video summarization

1. SYSTEM TECHNICAL DESCRIPTION
The ClassMiner system demonstrates a fully implemented tool for scalable video skimming and summarization. The key technology in the system is the integrated medical video content structure and events mining process, which was presented in a paper at the SIGMOD workshop on Data Mining and Knowledge Discovery [1]. As the system architecture in Fig. 1 indicates, we first apply a general video shot segmentation and key-frame selection scheme to parse the video stream into physical units. Then, the video group detection, scene detection and clustering strategies are executed to mine the video content structure. Various visual and audio feature processing techniques are utilized to detect some semantic cues, such as slides, face and speaker changes, etc. within the video, and these detection results are joined together to mine three types of events (presentation, dialog, clinical operation) from the detected video scenes. Finally, a scalable video skimming and summarization tool is constructed based on the mined video content structure and event information to help the user visualize and access video content.

1.1 Video Content Structure Mining
In general, most videos from daily life can be represented using a hierarchy of five levels (video, scene, group, shot and frame), increasing in granularity from top to bottom. Hence, the most efficient way to address video content is to construct a video content hierarchy. As shown in Fig. 1, our video content structure mining is executed in three steps: (1) group detection, (2) scene detection, and (3) scene clustering.

Shots belonging to one group generally share a similar background or have a high correlation in time series. Therefore, to segment the spatially or temporally related video shots into groups, a given shot is compared with shots that precede and succeed it (using no more than 2 shots) to determine the correlation between them. Given video shot S_i, if it is the first shot of a new group, it will have a higher correlation with shots on its right side than shots on its left side. Accordingly, a *separation factor* for shot S_i is defined by the ratio of the correlation between S_i and shots on both sides to evaluate a potential group boundary. With a threshold selection scheme among all *separation factors*, video group boundaries can be detected. With this strategy, two kinds of shots are absorbed into a given group: (1) shots related in temporal series, where similar shots are shown back and forth. Shots in this group are referred to as *temporally related*, and (2) shots similar in visual perception, where all shots in the group are similar in visual features. Shots in this group are referred to as *spatially related.*

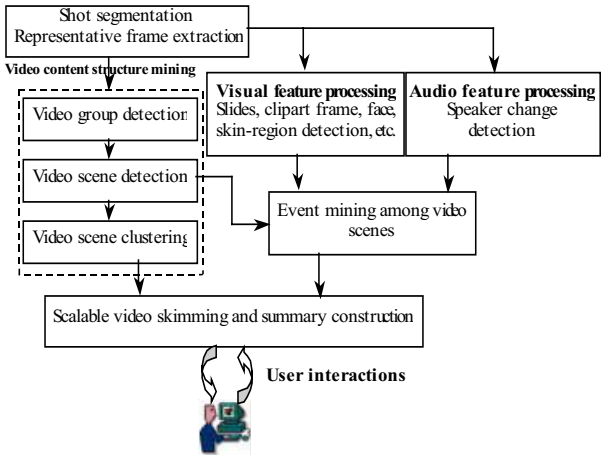

Figure 1. System architecture

With this strategy, there is no doubt that one scene may be parsed into several groups. However, groups in the same scene usually have a higher correlation with each other than with groups from different scenes. Hence, a group merging method is adopted to merge temporally related adjacent neighboring groups into scenes.

In most cases, many scenes are shown several times throughout the video. Clustering similar scenes into one unit eliminates redundancy and produces a more concise video content summary. We therefore introduce a seedless *Pairwise Cluster Scheme* (*PCS*) for video scene clustering, which iteratively merges the scenes with the highest similarity into one unit, until a certain number of clusters has been obtained. To determine the optimal cluster number for the *PCS*, we employ cluster validity analysis. The intuitive approach is to find clusters that minimize intra-cluster distance while maximizing inter-cluster distance.

Multimedia'02, December 1-6, 2002, Juan-les-Pins, France.

By applying these strategies, a given a video sequence can be parsed into a hierarchical structure of increasing granularity, consisting of the video, clustered scenes, scenes, groups and shots.

1.2 Video Events Mining

After video shots have been parsed into scenes, the event mining strategy is applied to detect event information within the scenes. A successful result would satisfy a query such as "Show me all patient-doctor dialogs within the video." Since medical videos are mainly used for educational purposes, the video content is usually recorded or edited using the style formats described below:

1. Using *presentations* by doctors or experts to express general topics in the video.
2. Using *clinical operations* (such as diagnosis, surgery, organ pictures, etc.) to present details of diseases, etc.
3. Using *dialogs* between doctors and patients to acquire other knowledge about medical conditions.

Accordingly, three types of events, Presentation, Clinical Operation and Dialog, within medical videos are mined by utilizing visual/audio features and rule information.

Visual feature processing is executed among all representative frames to extract semantically related visual cues. Currently, five types of special frame and regions are detected: slides or clip art frame, black frame, frame with face, frame with large skin area and frame with blood-red regions. Algorithm details can be found in [1-4].

Audio signals are a rich source of information in the video. They can be used to separate different speakers, detect various audio events, etc [1][5]. In our system, the objective is to verify whether speakers in different shots are the same person. The classification scheme can be separated into two steps: (1) select the representative audio clip for each shot, and (2) compare whether representative clips of different shots belong to the same speaker.

Given any mined scene, our objective is to verify whether it belongs to one of the following event categories:

1. A "*Presentation*" scene is defined as a group of shots that contain slides or clip art frames. At least one group in the scene should consist of temporally related shots. Moreover, at least one shot should contain a face close-up (human face with size larger than 10% of the total frame size), and there should be no speaker change between adjacent shots.
2. A "*Dialog*" scene is a group of shots containing both face and speaker changes. Moreover, at least one group in the scene should consist of spatially related shots. The speaker change should take place in adjacent shots, which both contain the face. At least one speaker should be duplicated more than once.
3. The "*Clinical Operation*" scene includes medical events, such as surgery, diagnosis, symptoms, etc. In this paper, we define the "Clinical Operation" as a group of shots without speaker change, where at east one shot in the scene contains blood-red or a close-up of a skin region (skin region with size larger than 20% of the total frame size) or where more than half of the shots in the scene contain skin regions.

Based on the above definitions, an integrated rule and visual/audio feature mining strategy has been developed to mine events information within given medical video scenes.

2. SYSTEM DEMONSTRATION

Based on mined video content structure and event information, a scalable video skimming and summarization tool, ClassMiner, has been constructed to visualize the video overview and help users access video content efficiently, as shown in Fig. 2. Currently, the system utilizes a four layer video skimming, where levels 4 through 1 consist of representative shots of clustered scenes, all scenes, all groups, and all shots of the video, respectively. Hence, the granularity of video skimming increases from level 4 through level 1. The user can switch to a different level of video skimming by clicking the upward or downward button (skimming level switcher). While the video skimming is playing, only the selected skimming shots are shown, and all other video is skipped. A progress bar indicates the position of the current skimming shot among all shots of the video. Users can drag the tag of the progress bar to fast access an interesting video unit.

To help the user visualize the mined events information within the video, a color bar is used to represent the content structure of the video so that the scenes can be accessed efficiently using event categorization. As shown in Fig. 2, the color of the bar for a given region indicates the event category to which the scene belongs. Using this strategy, a user can access the video content directly.

Figure 2. Scalable video skimming and summarization tool

3. REFERENCES

[1] Zhu, X.Q., Fan, J.P., Aref, W.G., Elmagarmid, K.A."ClassMiner: Mining medical video content structure and events towards efficient access and scalable skimming", *SIGMOD workshop on Data Mining and Knowledge Discovery, pp.9-16, Madison, June, 2002.*

[2] Zhu, X.Q, Fan, J.P. Elmagarmid, K.A., "Towards facial feature localization and verification for omni-face detection in video/images", *Prof. IEEE ICIP, 2002.*

[3] Fan, J.P., Zhu, X.Q., Wu, L.D., "Automatic model-based semantic object extraction algorithm", *IEEE CSVT, 11(10), pp.1073-1084, Oct., 2000.*

[4] Zhu, X.Q., Fan, J.P., Elmagarmid, A.K., Aref, W.G., "Hierarchical video summarization for medical data", SPIE *storage and retrieval for media database, pp.395-406, 2002.*

[5] Delacourt, P., Wellekens, C.J., "DISTBIC: A speaker-based segmentation for audio data indexing", *Speech communication, vol.32, p.111-126, 2000.*