Semantic Retrieval of Video

Ziyou Xiong[1], Xiang Zhou[2], Qi Tian[3], Rui Yong[4], and Thomas S. Huang[5]

Abstract: In this article we will review different research works in 3 types of video, i.e., video of meetings, movies and broadcast news, and sports video. We will then put them into a general framework of video summarization, browsing, and retrieval. We will also review different video representation techniques for these three types of video content within this general framework. At last we will present the challenges facing the video retrieval research community.

1.INTRODUCTION

Video content can be accessed by using either a top-down approach or a bottom-up approach [1, 2, 3, 4]. The top-down approach, i.e. video browsing, is useful when we need to get an "essence" of the content. The bottom-up approach, i.e. video retrieval, is useful when we know exactly what we are looking for in the content, as shown in Fig. 1 When we do not exactly know what we are looking for in the content, human-computer interaction, such as relevance feedback and active learning, can be used to better match the intentions and needs of the user with the video content.

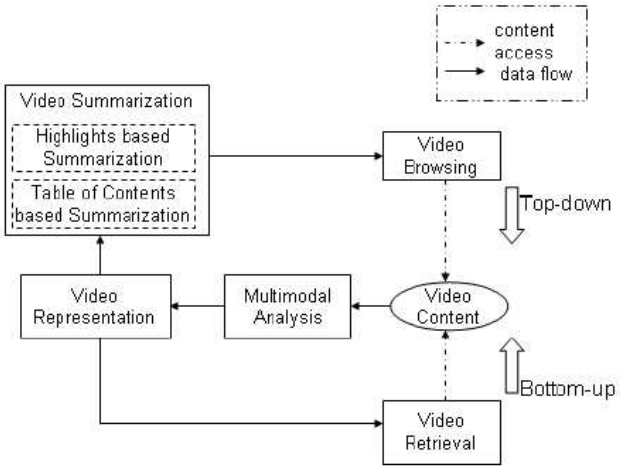

Figure 1.Relationship between video retrieval and browsing

In the following, we give an overview of the research work on 3 major types of video content, i.e., video of meetings, movies and broadcast news, and sports.

1.1Video of Meetings

Meetings are an important part of everyday life for many workgroups. Often, due to scheduling conflicts or travel constraints, people cannot attend all of their scheduled meetings. In addition, people are often only peripherally interested in a meeting such that they want to know what happened during the meeting without actually attending; being able to browse and skim these types of meetings could be quite valuable. Initial work on summarization and retrieval of video of meetings has been reported in [33].

1.2Movies and Broadcast News

Recently, movies and news videos have received great attention by the research community basically motivated by the interest of movie makers and broadcasters in building large digital archives of their assets for reuse of archive materials for TV programs or on-line availability to other companies [5]. Movies and news have a rather definite structure and do not offer a wide variety of edit effects, which are mainly cuts, or shooting conditions (e.g., illumination). This definite structure is suitable for content analysis and has been exploited for automatic classification, for example, in [6], [7], [8], [9], [10], [11]. In all of these systems a two stage scene classification scheme is employed. First, the video stream is parsed and video shots are extracted. Each shot is then classified according to content classes such as *news report*, *weather forecast etc*. The general approach to this type of classification relies on the definition of one or more image

[1] United Technologies Research Center, East Hartford, CT, Email: xiongz@utrc.utr.com

[2] Siemens Corporate Research, Princeton , NJ 08540, Email: xzhou@scr.siemens.com

[3] Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, Email: qitian@cs.utsa.edu

[4] Microsoft Research, One Microsoft Way, Redmond, WA, E-mail:yongrui@microsoft.com

[5] Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, E-mail: huang@ifp.uiuc.edu

templates for each content class. To classify a generic shot, a key frame is extracted and matched against the image template of every content class. Other works deal with the problem of video indexing using information sources like the text of captions and the audio track [7], [12]. This is due to the fact that movie and news images have an ancillary function with respect to words and video content is strongly related to textual and audio information that is contained in the audio. Also, speaker identity is important information that can effectively index the movie content [13].

1.3Sports Video

Event indexing and highlight extraction in sports video have been actively studied recently due to their tremendous commercial applications. Many researchers have studied the respective role of visual, audio and textual mode in this domain. For example, for the visual mode, Kawashima et al. [14] have extracted bat-swing features based on the video signal. Xie et al. [15] and Xu et al. [16] have segmented soccer videos into play and break segments using dominant color and motion information. Gong et al. [17] have targeted parsing TV soccer programs. By detecting and tracking the soccer court, ball, players, and motion vectors, they were able to distinguish nine different positions of the play (e.g., mid-field, top-right corner of the court, etc.). Ekin and Tekalp [18] have analyzed soccer video based on video shot detection and classification; for the audio mode, Rui et al. [19] have detected the announcer's excited speech and ball-bat impact sound in baseball games using directional audio template matching; for the textual mode, Babaguchi et al. [20] have looked for time spans in which events are likely to take place through extraction of keywords from the closed captioning stream. Their method has been applied to index events in American football video.

Since the content of sports video is intrinsically multimodal, many researchers have also proposed different information fusion schemes to combine different modes. A good reference is Snoek and Worring [21].

1.4An Analogy

How does a reader efficiently access the content of a 1000-page book? Without reading the whole book, he can first go to the ToC to find which chapters or sections suit his needs. If he has specific questions (queries) in mind, such as finding a term, he can go to the Index at the end of the book and find the corresponding sections addressing that question. On the other hand, how does a reader efficiently access the content of a 100-page magazine? Without reading the whole magazine, he can either directly go to the featured articles listed on the front page or use the ToC to find which article suits his needs. In short, the book's ToC helps a reader *browse*, and the book's index helps a reader *retrieve*. Similarly, the magazine's featured articles also help the reader *browse* through the highlights. All these three aspects are equally important in helping users access the content of the book or the magazine. For today's video content, techniques are urgently needed for automatically (or semi-automatically) constructing video ToC, video Highlights and video Indices to facilitate summarization, browsing and retrieval.

2.Terminology

Before we go into the details of the discussion, it will be beneficial to first introduce some important terms used in the digital video research field.

Scripted/Unscripted Content: A video that is carefully produced according to a script or plan that is later edited, compiled and distributed for consumption is referred to as *scripted content*. News videos, dramas and movies are examples of scripted content. Video content that is not scripted is then referred to as *unscripted*. In unscripted content, such as sports video and meetings video, the events happen spontaneously. One can think of varying degrees of “scripted-ness” and “unscripted-ness” from movie content to surveillance content.

Video shot: is a consecutive sequence of frames recorded from a single camera. It is the building block of video streams.

Key frame: is the frame which represents the salient visual content of a shot. Depending on the complexity of the content of the shot, one or more key frames can be extracted.

Video scene: is defined as a collection of semantically related and temporally adjacent shots, depicting and conveying a high-level concept or story. While shots are marked by physical boundaries, scenes are marked by semantic boundaries

Video group: is an intermediate entity between the physical shots and semantic scenes and serves as the bridge between the two. Examples of groups are temporally adjacent shots [22] or visually similar shots [3].

Play and Break: is the first level of semantic segmentation in sports video and surveillance video. In sports video (e.g soccer, baseball, golf), a game is in *play* when the ball is in the field and the game is going on; *break*, or out of play, is the complement set, i.e., whenever "the ball has completely crossed the goal line or touch line, whether on the ground or in the air" or "the game has been halted by the referee" [23]. In surveillance video, a play is a period in which there is some activity in the scene.

Audio Marker: is a contiguous sequence of audio frames representing a key audio class that is indicative of the events of interest in the video. An example of an audio marker for sports video can be the audience reaction sound (cheering and applause) or commentator's excited speech.

Video Marker: is a contiguous sequence of video frames containing a key video object that is indicative of the events of interest in the video. An example of a video marker for baseball videos is the video segment containing the squatting catcher at the beginning of every pitch.

Highlight Candidate: is a video segment that is likely to be remarkable and can be identified using the video and audio markers.

Highlight Group: is a cluster of highlight candidates.

In summary, scripted video data can be structured into a hierarchy consisting of five levels: video, scene, group, shot, and key frame, which increase in granularity from top to bottom [4] (see Fig. 2). Similarly, the unscripted video data can be structured into a hierarchy of four levels: play/break, audio-visual markers, highlight candidates, highlight groups, which increase in semantic level from bottom to top (see Fig. 3).

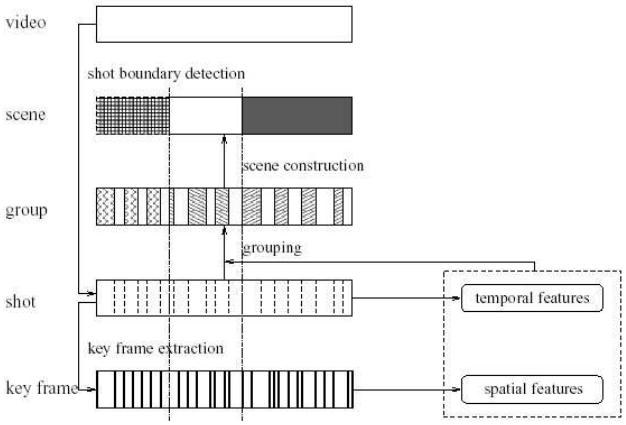

Figure 2.A hierarchical video representation for scripted content

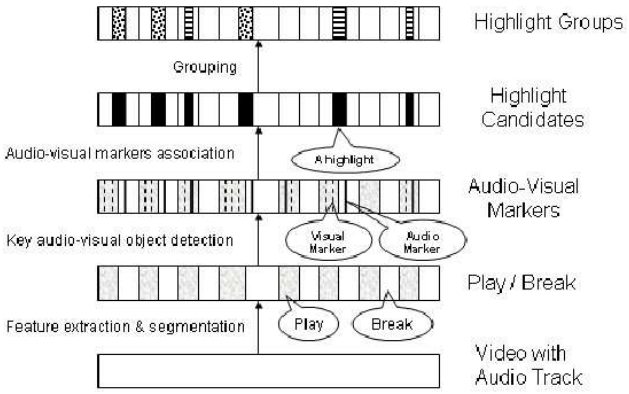

Figure 3.A hierarchical video representation for unscripted content

3.Video Representation and Indexing

Considering that each video frame is a 2D object and the temporal axis makes up the third dimension, a video stream spans a 3D space. Video representation is the mapping from the 3D space to the 2D view screen. Different mapping functions characterize different video representation techniques.

3.1Video of Meetings

We use the following user interface as an example to show video representation and indexing for meeting videos in Fig. 4. A low-resolution version of the RingCam panorama image is shown in the lower part of the client. A high resolution image of the speaker is shown in the upper left, which can either be automatically selected by the virtual director or manually selected by the user (by clicking within the panoramic image).

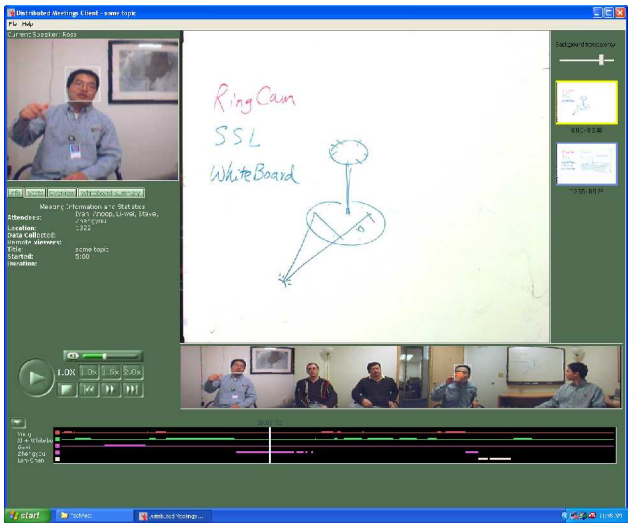

Figure 4.Meeting browsing client: Panorama window (bottom), speaker window (upper left), whiteboard (upper right), timeline (bottom).

The timeline is shown in the bottom of the window, which shows the results of speaker segmentation. The speakers are automatically segmented and assigned a unique color. The person IDs have been manually assigned, though this process could be automated by voice identification. The remote viewer can select which person to view by clicking on that person's color. The speakers can also be filtered, so that playback will skip past all speakers not selected.

The playback control section to the left of the panorama allows the remote view to seek to the next or previous speaker during playback.

In the following, we describe the major modules of the system in more detail.

Figure 5.RingCam: an inexpensive omnidirectional camera and microphone array designed for capturing meetings.

3.1.1Sound Source Localization

At the base of the RingCam (see Figure 5) is an 8-element microphone array used for beamforming and sound source localization. The microphone array has an integrated preamplifier and uses an external 1394 A/D converter (Motu828) to transmit 8 audio channels at 16-bit 44.1KHz to the meeting room server via a 1394 bus. The goal for sound source localization (SSL) is to detect which meeting participant is talking. For more reference on the RingCam, please see [24].

3.1.2Person Detection and Tracking

Although audio-based SSL can detect who is talking, its spatial resolution is not high enough to finely steer a virtual camera view. In addition, occasionally it can lose track due to room noise, reverberation, or multiple people speaking at the same time. Vision-based person tracking is a natural complement to SSL. Though it does not know who is talking, it has higher spatial resolution and tracks multiple people at the same time.

However, robust vision-based multi-person tracking is a challenging task, even after years of research in the computer vision community. The difficulties come from the requirement of being fully automatic and being robust to many potential uncertainties. After careful evaluation of existing techniques, we implemented a fully automatic tracking system by integrating three important modules: auto-initialization, multi-cue tracking and hierarchical verification [25].

Working together, these three modules achieve good tracking performance in real-world environment. A tracking example is shown in Fig. 4 with white boxes around the person's face.

3.1.3Sensor Fusion and Virtual Director

The responsibilities of the virtual director (VD) module are two folds. At a lower level, it conducts sensor fusion to gather and analyze reports from the SSL and multi-person tracker and make intelligent decisions on what the speaker window (the top left window in Fig. 4) should show. We use particle filters (PF) to obtain the speaker location. The proposal function of the PF is obtained from individual audio/video sensors, e.g., the person tracking module and SSL module. Particles are then drawn from the

proposal function that is then weighted and propagated through time. For a detailed description of the PF-based sensor fusion algorithm, readers are referred to [26].

At a higher level, just like video directors in real life, the VD module observes the rules of the cinematography and video editing in order to make the recording more informative and entertaining [24]. For example, when a person is talking, the VD should promptly show that person. If two people are talking back and forth, instead of switching between these two speakers, the VD may decide to show them together side by side (note that our system captures the entire 360º view). Another important rule to follow is that the camera should not switch too often; otherwise it may distract viewers.

3.1.4Speaker Segmentation and Clustering

For archived meetings, an important value-added feature is speaker clustering. If a timeline can be generated showing when each person talked during the meeting, it can allow users to jump between interesting points, listen to a particular participant, and better understand the dynamics of the meeting. The input to this preprocessing module is the output from the SSL, and the output from this module is the timeline clusters. There are two components in this system: pre-filtering and clustering. During pre-filtering, the noisy SSL output will be filtered and outliers thrown away. During clustering, K-mean's clustering is used during the first a few iterations to bootstrap, and a mixture of Gaussians clustering is then used to refine the result. An example timeline cluster is shown in the lower portion of Fig. 4.

3.2Movies and Broadcast News

Using an analysis framework that can detect shots, key frames and scenes, it is possible to come-up with the following representations for scripted content such as movies and broadcast news [8]. An example is shown in Fig. 6.

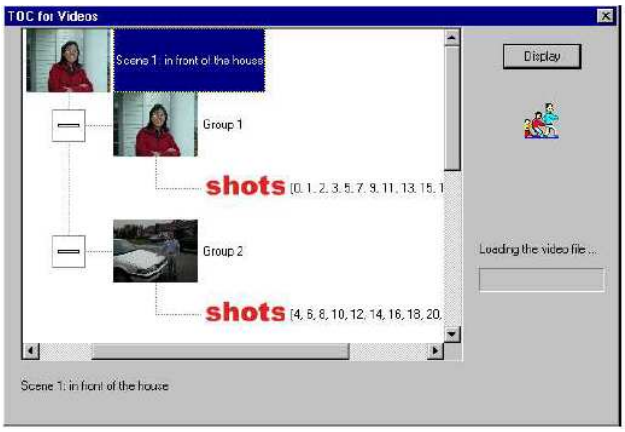

Figure 6. An example of video ToC

3.2.1Representation based on Sequential Key Frames

After obtaining shots and key frames, an obvious and simple video representation is to sequentially lay out the key frames of the video, from top to bottom and from left to right. This simple technique works well when there are few key frames. When the video clip is long, this technique does not scale, since it does not capture the embedded information within the video clip, except for time.

3.2.2Representation based on Groups

To obtain a more meaningful video representation when the video is long, related shots are merged into groups [22], [3]. In [22], Zhang et al. divide the entire video stream into multiple video segments, each of which contains an equal number of consecutive shots. Each segment is further divided into sub-segments; thus constructing a tree structured video representation. In [3], Zhong et al. proposed a cluster-based video hierarchy, in which the shots are clustered based on their visual content. This method again constructs a tree structured video representation.

3.2.3Representation based on Scenes

To provide the user with better access to the video, the construction of a video representation at the semantic level is needed [4], [2]. It is not uncommon for a modern movie to contain a few thousand shots and key frames. This is evidenced in [27] -- there are 300 shots in a 15-minute video segment of the movie "Terminator 2 - Judgment Day" and the movie lasts 139 minutes. Because of the large

number of key frames, a simple 1D sequential presentation of key frames for the underlying video (or even a tree structured layout at the group level) is almost meaningless. More importantly, people watch the video by its semantic scenes rather than the physical shots or key frames. While a *shot* is the building block of a video, it is a *scene* that conveys the semantic meaning of the video to the viewers. The discontinuity of shots is overwhelmed by the continuity of a scene [2]. Video ToC construction at the scene level is thus of fundamental importance to video browsing and retrieval. In [2], a scene transition graph (STG) of video representation is proposed and constructed. The video sequence is first segmented into shots. Shots are then clustered by using *time-constrained clustering*. The STG is then constructed based on the time flow of the clusters.

3.2.4Representation based on Video Mosaics

Instead of representing the video structure based on the video-scene-group-shot-frame hierarchy as discussed above, this approach takes a different perspective [28]. The mixed information within a shot is decomposed into three components:

- *Extended spatial information*: this captures the appearance of the entire background imaged in the shot, and is represented in the form of a few mosaic images.
- *Extended temporal information*: this captures the motion of independently moving objects in the form of their trajectories.
- *Geometric information*: this captures the geometric transformations that are induced by the motion of the camera.

3.3Sports Video

Highlights extraction from unscripted content requires a different representation from the one that supports browsing of scripted content. This is because shot detection is known to be unreliable for unscripted content. For example, in soccer video, visual features are so similar over a long period of time that almost all the frames within it, may be grouped as a single shot. However, there might be multiple semantic units within the same period such as attacks on the goal, counter attacks in the mid-field, etc. Furthermore, the representation of unscripted content should emphasize detection of remarkable events to support highlights extraction while the representation for scripted content does not fully support the notion of an event being remarkable compared to others. An example of the detected baseball highlights from a baseball video is shown in Fig. 7.

For unscripted content, using an analysis framework that can detect plays and specific audio and visual markers, it is possible to come up with the following representations.

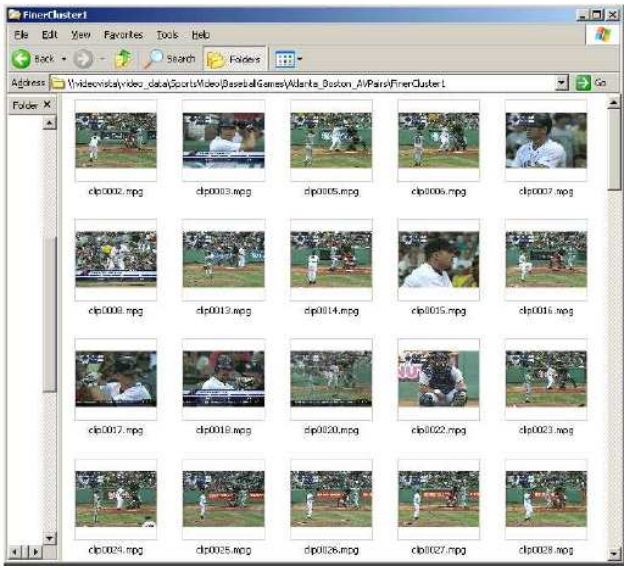

Figure 7. Highlight segments in a baseball video

3.3.1Representation based on Play/Break Segmentation

As mentioned earlier, play/break segmentation using low-level features gives a segmentation of the content at the lowest semantic level. By representing a key frame from each of the detected play segments, one can enable the end user to select just the play segments.

3.3.2Representation based on Audio-Visual Markers

The detection of audio and visual markers enables a representation that is at a higher semantic level than play/break representation is. Since the detected markers are indicative of the events of interest, the user can use both of them to browse the content based on this representation.

3.3.3Representation based on Highlight Candidates

Association of an audio marker with a video marker enables detection of highlight candidates that are at a higher semantic level. Such a fusion of complementary cues from audio and video helps eliminate false alarms in either of the marker detectors. Segments in the vicinity of a video marker and an associated audio marker give access to the highlight candidates for the end-user. For instance, if the baseball catcher pose (visual marker) is associated with an audience reaction segment (audio marker) that follows it closely, the corresponding segment is highly likely to be remarkable or interesting.

3.3.4Representation based on Highlight Groups

Grouping of highlight candidates would give a finer resolution representation of the highlight candidates. For example, golf swings and putts share the same audio markers (audience applause and cheering) and visual markers (golfers bending to hit the ball). A representation based on highlight groups, supports the task of retrieving finer events such as "golf swings only" or "golf putts only".

4.Sports Video Highlights Extraction

In this section, we describe our proposed approach for highlights extraction from "unscripted" content. We show the framework's effectiveness in three different sports namely soccer, baseball and golf. Our proposed framework can be summarized in Fig. 8. There are 4 major components in Fig. 8. We describe them one by one in the following.

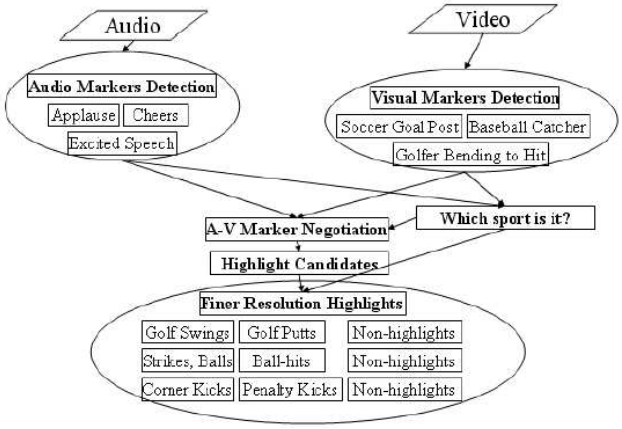

Figure 8.Proposed approach for sports highlights extraction

4.1Audio Marker Detection

Broadcast sports content usually includes audience reactions to the interesting moments of the games. Audience reaction classes including applause, cheering, and commentator's excited speech can serve as audio markers. Fig. 9 shows a unified audio marker detection framework for sports highlights extraction.

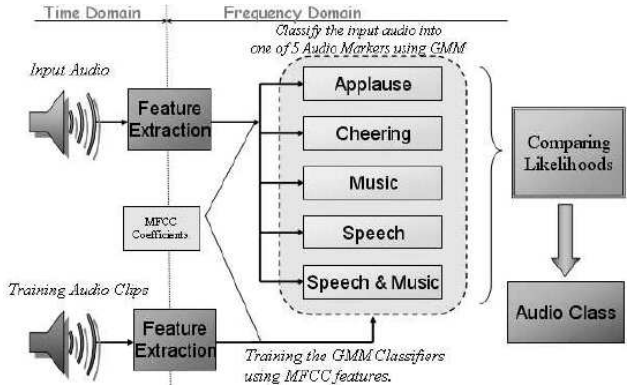

Figure 9.Audio markers for sports highlights extraction

Fig. 10 shows examples of some visual markers for three different games. For baseball games, we want to detect the pattern in which the catcher squats waiting for the pitcher to pitch the ball; for golf games, we want to detect the players bending to hit the golf ball; for soccer, we want to detect the appearance of the goal post. Correct detection of these key visual objects can eliminate the majority of the video content that is not in the vicinity of the interesting segments. For the goal of one general framework for all three sports, we use the following processing strategy: for the unknown sports content, we detect whether there are baseball catchers, or golfers bending to hit the ball, or soccer goal posts. The detection results can enable us to decide which sport (baseball, golf or soccer) it is.

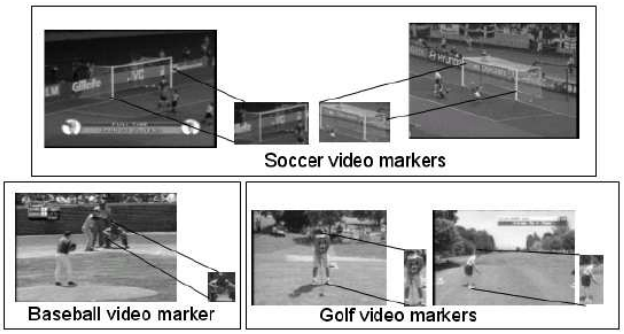

Figure 10.Examples of visual markers for different sports

An object detection algorithm such as Viola and Jones's [29] has been used to detect these visual markers. We can compare the number of detections with those in the ground truth set (marked by human viewers). We show the precision-recall curve for baseball catcher detection in Fig. 11. We have achieved, for example, 80% precision for a recall of 70%.

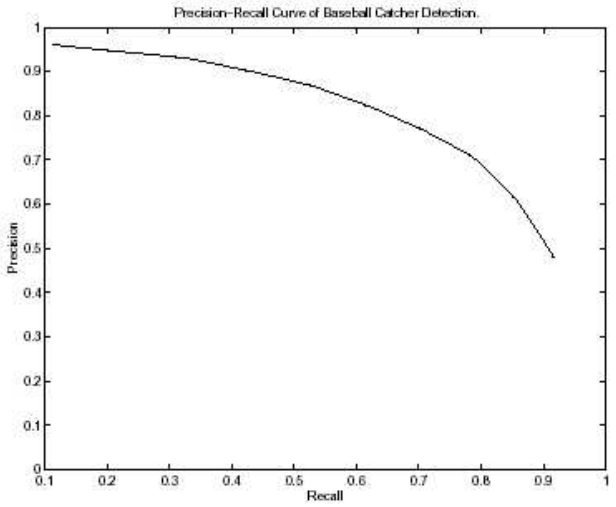

Figure 11.The precision-recall curve of baseball catcher detection

4.2Audio-Visual Markers Negotiation for Highlights Candidates Generation

Ideally each visual marker can be associated with one and only one audio marker and vice versa. Thus they make a pair of audio-visual markers indicating the occurrence of a highlight event in their vicinity. But, since many pairs might be wrongly grouped due to false detections and misses, some post-processing is needed to keep the error to a minimum. An algorithm might be as follows:

- If a contiguous sequence of visual markers overlaps with a contiguous sequence of audio markers by a large margin (e.g., the percentage of overlapping is greater than 50\%), then we form a "highlight" segment spanning from the beginning of the visual marker sequence to the end of the audio visual marker sequence.
- Otherwise, associate a visual marker sequence with the nearest audio marker sequence that follows it if the duration between the two is less than a duration threshold (e.g., the average duration of a set of training "highlights" clips from baseball games).

4.3Finer-Resolution Highlights Recognition and Verification

Highlight candidates, delimited by the audio markers and visual markers, are quite diverse. For example, golf swings and putts share the same audio markers (audience applause and cheering) and visual markers (golfers bending to hit the ball). Both of these two kinds of golf highlight events can be found by the aforementioned audio-visual markers detection based method. To support the task of retrieving finer events such as "golf swings only" or "golf putts only", we have developed techniques that model these events using low level audio-visual features.

Furthermore, some of these candidates might not be true highlights. We eliminate these false candidates using a finer-level highlight classification method. For example, for golf, we build models for golf swings, golf putts and non-highlights (neither swings nor putts) and use these models for highlights classification (swings or putts) and verification (highlights or non-highlights).

As an example, let us look at finer level highlight classification for a baseball game using low-level color features. The diverse baseball highlight candidates found after the audio markers and visual markers negotiation step are further separated using the techniques described here. For baseball, there are two major categories of highlight candidates, the first being "balls or strikes" in which the batter does not hit the ball, the second being "ball-hits" in which

the ball is hit to the field or audience. These two categories have different color patterns. In the first category, the camera is fixed at the pitch scene, so the variance of color distribution over time is low. In the second category, in contrast, the camera first shoots at the pitch scene, then it follows the ball to the field or the audience, so the variance of color distribution over time is higher.

5.Video Summarization and Browsing

In video summarization, what "essence" the summary should capture depends on whether the content is scripted or not. Since scripted content, such as news, drama, and movie, is carefully structured as a sequence of semantic units, one can get its essence by enabling a traversal through representative items from these semantic units. Hence, Table of Contents (ToC) based video browsing caters to summarization of scripted content. For instance, a news video composed of a sequence of stories can be summarized/browsed using a key-frame representation for each of the shots in a story. However, summarization of unscripted content, such as meetings and sports, requires a "highlights" extraction framework that only captures remarkable events that constitute the summary.

5.1ToC Based Browsing

For "Representation based on Sequential Key Frames", browsing is sequential, scanning from the top-left key frame to the bottom-right key frame. For "Representation based on Groups", a hierarchical browsing is supported [22], [3]. At the coarse level, only the main themes are displayed. Once the user determines which theme he is interested in, he can then go to the finer level of the theme. This refinement process can go on until the leaf level. For the STG representation, a major characteristic is its indication of time flow embedded within the representation. By following the time flow, the viewer can browse through the video clip.

5.2Highlights Based Browsing

For "Representation based on Play/Break Segmentation", browsing is also sequential, enabling a scan of all the play segments from the beginning of the video to the end. "Representation based on Audio-Visual Markers" supports queries such as "find me video segments that contain the soccer goal post in the left-half field", "find me video segments that have the audience applause sound" or "find me video segments that contain the squatting baseball catcher". "Representation based on Highlight Candidates" supports queries such as "find me video segments where a golfer has a good hit" or "find me video segments where there is a soccer goal attempt". Note that "a golfer has a good hit" is represented by the detection of the golfer hitting the ball followed by the detection of applause from the audience. Similarly, that "there is a soccer goal attempt" is represented by the detection of the soccer goal post followed by the detection of long and loud audience cheering. "Representation based on Highlight Groups" supports more detailed queries than the previous representation. These queries include "find me video segments where a golfer has a good swing", "find me video segments where a golfer has a good putt", or "find me video segments where there is a good soccer corner kick" etc.

6.Challenging Problems in Video Retrieval

In this paper we have reviewed some of the important issues in semantic video (esp. video) retrieval. Now, we step back to look at the broad field of content-based video retrieval, and try to ascertain: what are the most challenging research problems facing us?

1) *Bridging the semantic gap*

Perhaps the most desirable mode of image/video retrieval is still by keywords or phrases. However, manually annotate the images and video data is extremely tedious. To do annotation automatically or semi-automatically, we need to bridge the "semantic gap", i.e., to find algorithms that will infer high-level semantic concepts (sites, objects, events) from low-level image/video features that can be easily extracted from the data (color, texture, shape and structure, layout; motion; audio - pitch, energy, etc.).

One sub-problem is Audio Scene Analysis. Researchers have worked on Visual Scene Analysis (Computer Vision) for many years, but Audio Scene Analysis is still in its infancy, and an under-explored field.

Another sub-problem is multimodal fusion, esp. how to combine visual and audio cues to bridge the semantic gap in video.

2) *How to best combine human intelligence and machine intelligence*

One advantage of information retrieval is that in most scenarios there is a human (or humans) in the loop. One prominent example of human-computer interaction is Relevance Feedback.

3) *New Query Paradigms*

For image/video retrieval, people have tried query by keywords, similarity, sketching an object, sketching a trajectory, painting a rough image, etc. Can we think of useful new paradigms?

4) *Video Data Mining*

Searching for interesting/unusual patterns and correlations in video has many important applications, including Web Search Engines and dealing with intelligence data. Work to date on Data Mining has been mainly in Text data.

5) *How To Use Unlabeled Data?*

We can consider Relevance Feedback as a two-category classification problem (relevant or irrelevant). However, the number of training samples is very small. Can we use the large number of unlabeled samples in the database to help? Also, how about active learning (to choose the best samples to return to the user to get most information about the user's intention through feedback)?

Another problem related to image/video data annotation is Label Propagation. Can we label a small set of data and let the labels propagate to the unlabeled samples?

6) *Incremental Learning*

In most applications, we keep adding new data to the database. We should be able to change the parameters of the retrieval algorithms incrementally, not needing to start from scratch every time we have new data.

7) *Using Virtual Reality Visualization To Help*

Can we use 3D audio/visual visualization techniques to help a user to navigate through the data space to browse and to retrieve?

8) *Structuring Very Large Databases*

Researchers in audio/visual scene analysis and those in Databases and Information Retrieval should really collaborate CLOSELY to find good ways of structuring very large video databases for efficient retrieval and search.

9) *Performance Evaluation*

How do we compare the performances of different retrieval algorithms?

10) *What Are the Killer Applications of Video Retrieval?*

Few real applications of video retrieval have been accepted by the general public so far. Is web video search engine going to be the next killer application? It remains to be seen. With no clear answer to this question, it is still a challenge to do research that is appropriate for real applications.

7.Acknowledgement

We thank Dr. Regunathan Radhakrishnan and Dr. Ajay Divakaran of Mitsubishi Electric Research Labs, Cambridge, MA for their contributions and helpful discussions.

8.REFERENCES

[1] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," ACM Multimedia Sys., vol. 1, no. 1, pp. 1-12, 1993.

[2] R. M. Bolle, B.-L. Yeo, and M. M. Yeung, "Video query: Beyond the keywords," Technical Report, IBM Research, Oct. 17, 1996.

[3] D. Zhong, H. Zhang, and S.-F. Chang, "Clustering methods for video browsing and annotation," Tech. Rep., Columbia University, 1997.

[4] Y. Rui, T. S. Huang, and S. Mehrotra, "Exploring video structures beyond the shots," in Proc. of IEEE Conf. Multimedia Computing and Systems, 1998.

[5] M. Bertini, A. D. Bimbo, and P. Pala, “Content based indexing and retrieval of TV news,” Pattern Recognition Letters, 22, pp. 503-516, 2001.

[6] D. Swanberg, C. Shu, and R. Jain, “Knowledge guided parsing in video databases,” SPIE 1908 (13), pp. 13-24, 1993.

[7] K. Ohtsuki, K. Bessho, Y. Matsuo, S. Matsunaga, and Y. Hayashi, “Automatic Indexing of Broadcast News by Combining Audio, Speech and Visual Information”, in this issue.

[8] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video Shot Boundary Detection and Condensed Representation: A Review", in this issue.

[9] A. Hanjalic, "Extracting Moods from Pictures and Sounds: Towards Truly Personalized TV", in this issue.

[10] X. Wu, C.-W. Ngo, and Q. Li, "Threading and Auto-Documentary in News Videos", in this issue.

[11] A. Kokaram, N. Rea, R. Dahyot, M. Tekalp, P. Bouthemy, P. Gros, and I. Sezan, "Browsing sports video", in this issue.

[12] A. G. Hauptmann, and M. J. Witbrock, “Informedia: news-on-demand multimedia information acquisition and retrieval,” Intelligent Multimedia Information Retrieval, pp. 213-239, 1997.

[13] Y. Li, S. Narayanan, and C.-C J. Kuo, “Content-based movie analysis and indexing based on audio visual cues,” IEEE Trans. Circuits and Systems for Video Technology, Vol. 14, No. 8, August 2004.

[14] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki, "Indexing of baseball telecast for content-based video retrieval," in Proceedings of the International Conference on Image Processing, 1998, pp. 871-874.

[15] L. Xie, S. F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in Proceedings of the International Conference on Acoustic, Speech, and Signal Processing, vol. 4, May 2002, pp. 4096-4099.

[16] P. Xu et al., "Algorithms and system for segmentation and structure analysis in soccer video," in Proceedings of IEEE Conference on Multimedia and Expo, Aug. 2001, pp. 928-931.

[17] Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in IEEE International Conference on Multimedia Computing and Systems, 1995, pp. 167-174.

[18] A. Ekin and A. M. Tekalp, "Automatic soccer video analysis and summarization," in Proceedings of the International Conference on Electronic Imaging: Storage and Retrieval for Media Databases, 2003, pp. 339-350.

[19] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in Proceedings of the Eighth ACM International Conference on Multimedia, 2000, pp. 105-115.

[20] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event-based indexing of broadcasted sports video by intermodal collaboration," IEEE Transactions on Multimedia, vol. 4, no. 1, pp. 68-75, March 2002.

[21] C. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," Tech. Rep., Intelligent Sensory Information Systems Group, University of Amsterdam, Technical Report 2001-20, 2001.

[22] Zhang, S. W. Smoliar, and J. J. Wu, "Content-based video browsing tools," in Proc. ISandT/SPIE Conf. on Multimedia Computing and Networking, 1995.

[23] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," Pattern Recognition Letters, 2004. to appear.

[24] R. Cutler, Y. Rui, et. al. Distributed meetings: a meeting capture and broadcasting system, Proc. of ACM Multimedia 2002.

[25] Y. Chen, Y. Rui and T. Huang, JPDAF Based HMM for Real-Time Contour Tracking, , Proc. of IEEE CVPR 2001, pp.I-543 to 550, Kauai, Hawaii, December 11-13, 2001

[26] Y. Chen and Y. Rui Real-time Speaker Tracking Using Particle Filter Sensor Fusion, Proceedings of the IEEE , vol. 92, no. 3, pp. 485-494, Mar. 2004

[27] M. Yeung, B.-L. Yeo, and B. Liu, "Extracting story units from long programs for video browsing and navigation," in Proc. IEEE Conf. on Multimedia Comput. and Syss, 1996.

[28] M. Irani and P. Anandan, "Video indexing based on mosaic representations," Proceedings of IEEE, vol. 86, pp. 905{921, May 1998.

[29] P. Viola and M. Jones, "Robust real-time object detection," Second International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing and Sampling, July 2001. Vancouver, Canada.

Ziyou Xiong received his BS degree from Wuhan University, Hubei Province, China, in July 1997. He received his MS degree in electrical and computer engineering from University of Wisconsin, Madison in December 1999 and PhD degree in electrical and computer engineering from University of Illinois at Urbana Champaign (UIUC) in October 2004. He has also been a research assistant with the Image Formation and Processing Group of the Beckman Institute for Advanced Science and Technology at UIUC from January 2000 to August 2004. In the summers of 2003 and 2004, he worked on sports audio-visual analysis at Mitsubishi Electric Research Labs, Cambridge, Massachusetts. Since September 2004, he has been with the Dynamic Modeling and Analysis group at the United Technologies Research Center as a senior researcher/scientist in East Hartford, Connecticut. His current research interests include image and video analysis, video surveillance, computational audio-visual scene analysis, pattern recognition, machine learning and related applications. He has published several journal and conference papers, invited book chapters on audio-visual person recognition, image retrieval, and sports video indexing, retrieval, and highlight extraction. He has also coauthored a book titled Facial Analysis from Continuous Video with Application to Human-Computer Interface (Kluwer, 2004).

Xiang Zhou received the PhD degree in electrical engineering from the University of Illinois at Urbana Champaign (UIUC) in 2002. Previously, he received the bachelor's degree in automation and in economics and management (minor) and studied economics for two years in a PhD program at Tsinghua University, China. Since 2002, he has been with Siemens Corporate Research in Princeton, New Jersey. His research interests include computer vision and machine learning, object detection, tracking, and recognition, multimedia analysis, representation, understanding, and retrieval. He was the program cochair for the 2003 International Conference on Image and Video Retrieval. He is the coauthor of the book Exploration of Visual Data (Kluwer 2003). He was the recipient of eight scholarships and awards from Tsinghua University from 1988 to 1995. In 2001, he received the M.E. Van Valkenburg Fellowship Award, an award given to one or two PhD students in the ECE department of UIUC each year "for demonstrated excellence in research in the areas of circuits, systems, or computers. " He is a member of the IEEE and the IEEE Computer Society.

Qi Tian received his Ph.D. degree in Electrical and Computer Engineering from University of Illinois at Urbana-Champaign (UIUC), Illinois in 2002. He received his M.S. degree in Electrical and Computer Engineering from Drexel University, Philadelphia, Pennsylvania, in 1996, and B.E. degree in Electronic Engineering from Tsinghua University, China, in 1992, respectively. He has been an Assistant Professor in the Department of Computer Science at the University of Texas at San Antonio (UTSA) since 2002 and an Adjunct Assistant Professor in the Department of Radiation Oncology at the University of Texas Health Science Center at San Antonio (UTHSCSA) since 2003. He has published over 50 referred book chapters, journal and conference papers in these fields. He has served as co-chairs of ACM Multimedia Workshop Multimedia Information Retrieval (2005) and SPIE Internet Multimedia Management Systems (2005), session chairs and PC members of a number of international conferences in computer vision and multimedia such as ICME, ICPR, CIVR, MIR, and VCIP. He is a Senior Member of IEEE, and a Member of ACM.

Yong Rui is a Researcher in the Communication and Collaboration Systems (CCS) group in Microsoft Research, where he leads the Multimedia Collaboration team. Dr. Rui is a Senior Member of IEEE and a Member of ACM. He is on the editorial board of International Journal of Multimedia Tools and Applications. He received his PhD from University of Illinois at Urbana-Champaign (UIUC). Dr. Rui's research interests include computer vision, signal processing, machine learning, and their applications in communication, collaboration, and multimedia systems. He has published one book (*Exploration of Visual Data*, Kluwer Academic Publishers), six book chapters, and over sixty referred journal and conference papers in the above areas. Dr. Rui was on Program Committees of ACM Multimedia, IEEE CVPR, IEEE ECCV, IEEE

ACCV, IEEE ICIP, IEEE ICASSP, IEEE ICME, SPIE ITCom, ICPR, CIVR, among others. He was a Co-Chair of IEEE International Workshop on Multimedia Technologies in E-Learning and Collaboration (WOMTEC) 2003, the Demo Chair of ACM Multimedia 2003, and a Co-Tutorial Chair of ACM Multimedia 2004. He was on NSF review panel and National Academy of Engineering's Symposium on Frontiers of Engineering for outstanding researchers.

Thomas S. Huang received his B.S. Degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, China; and his M.S. and Sc.D. Degrees in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge, Massachusetts. He was on the Faculty of the Department of Electrical Engineering at MIT from 1963 to 1973; and on the Faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing at Purdue University from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, and Research Professor at the Coordinated Science Laboratory, and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology and Co-Chair of the Institute's major research theme Human Computer Intelligent Interaction. Dr. Huang's professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 14 books, and over 500 papers in Network Theory, Digital Filtering, Image Processing, and Computer Vision. He is a Member of the National Academy of Engineering; a Foreign Member of the Chinese Academies of Engineering and Sciences; and a Fellow of the International Association of Pattern Recognition, IEEE, and the Optical Society of American; and has received a Guggenheim Fellowship , an A.V. Humboldt Foundation Senior U.S. Scientist Award , and a Fellowship from the Japan Association for the Promotion of Science . He received the IEEE Signal Processing Society's Technical Achievement Award in 1987, and the Society Award in 1991. He was awarded the IEEE Third Millennium Medal in 2000. Also in 2000, he received the Honda Lifetime Achievement Award for "contributions to motion analysis". In 2001, he received the IEEE Jack S. Kilby Medal. In 2002, he received the King-Sun Fu Prize, International Association of Pattern Recognition; and the Pan Wen-Yuan Outstanding Research Award. He is a Founding Editor of the International Journal Computer Vision, Graphics, and Image Processing; and Editor of the Springer Series in Information Sciences, published by Springer Verlag.