Semantic Video Indexing: Approach and Issues

Arun Hampapur
IBM TJ Watson Research Center
30 Sawmill River Rd, Hawthorne, NY 10532
arunh@us.ibm.com

Abstract

Providing concept level access to video data requires, video management systems tailored to the domain of the data. Effective indexing and retrieval for high-level access mandates the use of domain knowledge. This paper proposes an approach based on the use of knowledge models to building domain specific video information systems. The key issues in such systems are identified and discussed.

1 Introduction

The digital age along with the world wide web provides access to data on a scale that is unprecedented. Most databases provide users with data access at a very low-level. A user will have to deal with a significant *data overload* and spend considerable effort to get the information needed from the data available. One approach to solving this problem is to provide semantic search [13] and filtering capabilities. An additional strategy to dealing with the problem of *data overload* is to build domain specific information systems, where the prior knowledge from the domain is effectively utilized to provide higher-level semantic access to the information.

The problem of *data overload* is exasperated in the case of temporal multimedia streams like video. Multimedia data management is an active area of research which addresses the video management problem at the base level. The next challenge in video management is to build domain specific video management systems, which can provide concept level access to video data. A key component of such systems is *semantic video indexing.*

This paper, examines the issues in semantic video indexing and proposes a knowledge model to support semantic indexing and retrieval techniques. Sports video is the domain of choice and a basketball game is used as an example. Section 2 presents the example and provides the context for the rest of the paper. Section 3 presents some of the issues in the design of semantic video indexing systems and relates the issues to existing research. Section 4 presents a knowledge model for representing the various components of the domain knowledge. Section 5 discusses some of the inter-operability issues between multiple domain specific video systems. Section 6 summarizes the discussion presented in the paper.

The following are the definitions of a few terms used in the remainder of the paper.

Domain is defined as *a sphere of knowledge, influence, or activity* [11]. In the context of this paper a domain is defined as a *set of activities* which are related to each other and have a meaning attached to them. Examples include Sports, News, Education, Surveillance and Security etc. Videos from basketball, football, soccer, tennis etc, would be considered as **domain instances** within the sports domain. The term video is used to include, the image-sequence, synchronized audio and text transcript (closed captioning).

Indexing/Annotation: This is the process of parsing the video and associating different segments of the video to events in the given domain. This process could be completely automatic (using a software algorithm) or manual or human-assisted.

Semantic Video Indexing (SVI) is the use of prior knowledge about the domain of the video data in the indexing process. A system that uses SVI is called *Domain Specific Video Indexing System.*

2 Illustrative Example

A **3.00 minute** segment of a basketball game between the **New York Knicks** and the **Miami Heat** is used as the example. The segment is taken from the television broad cast of the game. The video data is stored as an *MPEG 1* video file. In this paper it is represented as an annotated time line of events in Figure 2 (top). This *annotated representation* uses *five annotation tracks* namely, *possession, event, player, speech and audio*. Each of these annotations are discussed below.

Possession: indicates which team is in control of the ball at each point in the game. In this case it is either the *Miami Heat* or *NY Knicks* or the ball is dead **(D)**.

Event: These are few of the possible activities in a basketball game, that occurred during the chosen 3 minute example clip. The events in the example include, *2 Missed-Shots, 2 Fouls, 2 Ball-Inbounds, 2 Scored-Shots* and a *free throw*. In general there are many more events that occur in a basketball game.

Player: This annotation records the name of the player who was primarily responsible for the event at the given time. For example, at 1:38 seconds into the clip, Hardaway scored a successful basket. At 1:24, the ball was *in-bounded* by an *Unknown* player.

Speech: This layer provides an index into the speech table (table 1 (top)), which is a transcript derived from the closed captioning that accompanies the game. One of the key observations to be made here is that, the speech (commentary) is almost totally unrelated to the visual activity going on at the time on the court, except at points where events of interest occur in the game.

Audio: This track captures the audio description of the clip, this includes the other audio events in progress in addition to the speech by the commentators.

Shot Description: This track describes the shots (continuous camera takes) [4] in the example clip. The descriptions in table 1 (bottom) provide a visual description of the activity in the shot. These descriptions were derived by the author. The key observation here is that the actual shots which carry the plays are much longer than the ones which capture other visual content.

Consider the segment of the game between minutes, 1:30 and 1:40in figure 2 (top). There was a shot scored by the Hardaway of the Miami Heat. A possession change is indicated on the *possession* track. The *event* track shows that a *shot scored* event occurred and the *player* track indicates that it was scored by Hardaway. The *speech track* (table 1 (top)), shows the commentator **(Mike)** saying: *"Hardaway off the dribble. Tim Hardaway creating his own shot, knocks it down, hardaway now five of eight from the field"*. The shot table (table 1 (bottom) shot numbers 10 and 11) show clips of Hardaway scoring the shot and a close-up of Hardaway.

Consider the following scenario, while surfing the web for basket ball related information a user hits the NBA site maintained by ESPN [1]. Being a Miami Heat fan, the user looks at the statistics of different players. Impressed by Hardaway's scores, the user would like to access all game videos where *"Hardaway created his own shot off a dribble"*.

A number of important video domains like surgical training, assembly manuals, surveillance etc require such high-level access to the data. A video database which could support these types of a query would require semantic video indexing. The general problem of video data management [4] and video indexing [6] have been an active area of research for some time. There are a few systems that are commercially available which address the general problem of video management [8]. However, semantic video indexing and domain specific video indexing systems are a very new research area. This paper examines the critical issues in sematic video indexing, proposes a knowledge model and list some of the future directions for research in the area.

3 Design Issues

Given a specific domain of interest (like basketball), there is a significant amount of prior knowledge about the domain that can be used in the design of the the indexing system that handles data from this domain. Using domain specific knowledge to support fine grained queries (like the example in section 2) gives rise to the following issues.

Automatic Indexing: How will the information be captured? Dealing with video manually is an extremely cumbersome and tedious process even with coarse grain indexing. As the granularity of indexing

becomes finer, the cost of manual indexing becomes prohibitive. Automatic data indexing algorithms become extremely important in the design of a feasible system.

Knowledge Representation: How will the domain knowledge be represented and used in the system ? Examining a domain like basketball games, reveals that it incorporates different kinds of knowledge, suggesting the use of multiple representation for different components of the domain knowledge model. The domain knowledge can be used in both the indexing and the querying aspects of the system.

In the remainder of the section we focus on the automatic indexing issue. The goal is to to be able to propose a knowledge model which can support this class of indexing algorithms.

3.1 Automatic Indexing

Automatic Indexing is critical to the success of any domain specific video indexing system. There are several automatic indexing algorithms [2, 3, 9, 10, 14, 16] which have been developed to extract low-level features from video streams. All these indexing algorithms use *image-pixel data* for feature extraction. There has been other research efforts which use audio processing, text and speech analysis to index video, but so far none of these techniques have been used for domain specific indexing.

Most of the indexing algorithms cited above use the following key image processing algorithms in the automatic indexing process.

Camera Motion Estimation and Compensation: These algorithms recover parametric estimates of the camera motion between consecutive frames in the video. These parameters are used either to *locate independently moving objects* or to label the sequence in terms of camera motion operations like *pan-left, pan-right, zoom-in, zoom-out etc.*

Object Segmentation and Tracking Techniques: Here objects of interest like *ball, player* etc are segmented from the background based on color or motion attributes. These objects are then tracked from across the different frames of the video, to obtain an estimate of the trajectory of the objects.

Line Detection: is used to detect structures of interest in the scene like *goal posts, markings on the field* etc. The algorithms incorporate operations like edge-detection, edge-linking and edge-thinning.

Once these basic features have been extracted they are filtered using domain knowledge to generate a set of automatic-index-events. For example, if a the **basketball board detector** has a positive response between 1:30 and 1:40 in the example, and the track of the basketball is within the vicinity of the board, then a *basket scored* event can be annotated.

The automatic indexing algorithms use three types of knowledge, namely, *physical, cinematic* and *semantic*. Each of these types of knowledge is briefly discussed below.

Physical Knowledge: This type of knowledge includes constraints derived from the physical environment in which the sport is played. For example, in basketball this includes color of the ball, structure of the backboard and basket.

Cinematic Knowledge includes the details on how the particular sporing event is filmed and produced. This includes the camera-motion, editing and camera locations used in producing the video. Further details on using cinematographic constraints can be found in [4].

Semantic Knowledge includes the knowledge about the actual sport, the temporal structure of the game, the rules of the games and other high level information about the sport. For example, a basketball game has 4 quarters, with a 130 sec break between Q1,Q2 and Q3,Q4 etc (figure 2 (bottom)).

A study of the various indexing techniques [2, 3, 9, 10, 14, 16] reveals that most of the techniques rely very much on the *physical and cinematic* knowledge to extract the features and event labels. These labels are not related to the *semantic* knowledge of the sport. Swanberg et al [15] and Zhang et al [17] have proposed the use of a temporal domain model for news. However, in the case of news videos the *unconstrained nature* of the content *precludes* the use of *strong semantic* knowledge models. In the case of sports, the use of such knowledge models is very appropriate and can be used effectively to index the video at a much higher level. The knowledge model proposed in section 4 provides a structure for representing domain knowledge.

4 The Knowledge Model

The core of a *domain specific video indexing system* is the *knowledge model*, analogous to the *data-model* in a generic database system. This section presents a structure for representing the knowledge model and examines the different components with reference to the basketball domain. Explicit knowledge representation has been recognized as the key to dealing with domain specific problems in the artificial intelligence community [12]. Figure 1 shows the overall structure of the knowledge model. The knowledge model includes the *semantic, cinematic* and *physical world* model. Each of these are discussed below.

Semantic Model: The semantic model represents the high level knowledge in the given domain. The semantic model includes several *sub-models*. The choice of the sub-models and the exact knowledge representation scheme used is dependent on the domain and the nature of queries that need to be supported by the system. For example in basketball [7] we could represent, the game model using a temporal representation scheme, the violations and foul rules can be represented using a rule-base and the structure of the team can be represented using a hierarchical object oriented model. In the case of video, the *temporal game-model* is the most important sub-model and is discussed below.

Finite State Game Model: Team sports like basketball, soccer, baseball etc [7] are structured activities which lend themselves well to finite state modeling (FSM). Figure 2 (bottom) represents the top-level FSM of a basketball game. *Q1,Q2,Q3* and *Q4* represent the 4 quarters of the game. *OT* represents *over time*. The labels on the arrows indicate the events that cause transitions. Figure 2 (bottom inset) also shows the level 2 FSM of Q3. This represents the game at a finer level of detail. Depending on the system requirements, the model can be expanded represent finer details of the game.

Instancing the Game Model: The game model represents the game at an abstract level. In the case of a particular video, the FSM has to be instantiated by *linking* each state of the model to *one or more* time intervals within the instance of the video. This is shown in figure 2 (bottom) by the arrow.

Using the Game Model: The FS game model can be used for the purpose of *browsing* a video and for the purpose of *prioritizing operators* during the indexing process. Swanberg et al [15] have discussed the use of a FSM but only for the purpose of matching a given state to a potential video segment.

Cinematic Model: The cinematic model [4] represents the prior knowledge about how the game is captured on video. The sub-models shown in figure 1 include the *composition* and *imaging* models. The imaging model includes details of the *number* and *location* of the camera and microphones used to record the game as part of the *camera geometry model.* The *camera motion* model includes, the details of the camera operations used to film the game. For example, Saur et al [14] use the *camera panning patterns* to infer events in the video stream. The *composition* model deals with the switching pattern between camera's to generate the final version of the video. For example, in basketball games the duration of *camera-takes* of *normal play* states tends to be much longer than the duration of other *camera-takes*. This can be seen by comparing the duration of shots 10, 12, 19, 20 with other shots in table 1 (bottom).

Physical World Model: This model incorporates the knowledge of the physical world. The *environment* model includes knowledge about the *court dimensions, court color, backboard and basket structure* and several such aspects. The environment model includes only the static (more permanent) aspects of the physical world. The *object model* includes details about the *players* (uniform colors, motion patterns, approximate physical dimensions, etc), *ball* (color,shape, speeds, motion patterns). This knowledge is used in performing low level operations like *player segmentation, ball tracking* etc [3, 2].

The knowledge model can be instantiated for every new video that is indexed. The population of various sub-models in the knowledge model can be performed both manually and automatically depending on the types of indexing algorithms. Architectures which facilitates such annotation activity have been explored in literature [5]. Depending on the nature of the application, sub-models included in each model and the knowledge represented in each sub-model can be varied. For example, in the case of domain specific indexing system, that is meant to support coaching, the semantic model can incorporate the *player skill* model and the *player position* model [7].

The knowledge model presented above addresses

the indexing issues presented in section 3. The use of the FS game model allows for *queries* which are regular expressions in terms of the states of the model which translates to higher level user queries. The knowledge represented in the knowledge model can be used in the *automatic indexing* algorithms. The use of explicit representation of the knowledge separates out the details of the particular domain from the automatic indexing algorithms. For example, if the uniform colors of a team change, this can be changed in the knowledge model while keeping the rest of the system intact.

5 Inter-operability Issues

The previous sections introduced and discussed various issues involved in semantic video indexing. This section discusses the issues involved in interoperability of *domain instances*. For example, consider the domain of **Sports Video** with two *domain instances* namely, *basketball* and *soccer*. The inter-operability issues across domain instances include *portability of indexing techniques*, *portability of knowledge models* etc. The most important challenge is that **'the domain-specific nature of the underlying information management systems should be transparent to the user '** *(human or software agent)*. In other words, different domain instances within a given domain should present a standardized *common user interface*. The rest of this section focussed on this issue.

5.1 Common User Interface

Given two domain instances, (like basketball and soccer), the the goal of the common user interface is to allow an application to query and retrieve data from both the domain instances using a single language. This requires that the two information management systems use a *common frame work* for representing their domain events. Section 4 proposed the use of a Finite State Model to capture the events and the relationships within a domain instance. The same finite state model frame work can also be used to ensure that two domain instances present a common user interface.

Basketball and Soccer are used as examples to explore the issues involved with designing common user interfaces across domain events. Figure 2 (middle) illustrates how *domain events* from both basket ball and soccer can be grouped under *abstract event categories*. The abstract categories used in figure 2 (middle) are *temporal structure, fouls and violations, play strategies, play starting rules, scoring patterns* and *other events*. These categories do not cover the entire set of events in the two domain-instances and are intended only to illustrate the use of *abstract event categories (AEC)*. An application which uses these AECs will be shielded to a large extent from the specific nature of the events in each of the individual domain-events. The idea of using AECs to design a common user interface also raises several issues about measuring the efficacy with which the AEC covers domain-events in the two domain-instances and other issues which are beyond the scope of this presentation.

6 Conclusion

Semantic access to video requires the use of domain knowledge both in the indexing and retrieval processes. This paper examined the issues involved in the design of domain specific video management systems and identified automatic indexing and knowledge modeling as the key components of such systems. The paper proposed a knowledge model for domain specific indexing and a FSM based temporal game model for basketball. The paper has also identified *common user interfaces* as a key requirement for ensuring inter-operability between domain instances and proposed the use of *Abstract Event Categories* as an approach.

Although the discussion in this paper has been focussed on the domain of Sports Videos, the approach to the problem and the ideas presented here can be extend to other data types and to other domains.

References

[1] espn.sportszone.com/nba. *ESPN Basketball web site*. ESPN, 1998.

[2] Sudhir G, John Lee, and Anil Jain. Automatic classification of tennis video for high level content based retrieval. Technical report, The Hong

Kong University of Science and Technology, 1997.

[3] Yihong Gong, Chua Hock-Chuan, and Lim Teck Sin. An automatic video parser for tv soccer games. In *Proceedings of the Second Asian Conference on Computer Vision*, Singapore, December 1995. IEEE.

[4] Arun Hampapur. *Designing Video Data Management Systems*. PhD thesis, The University of Michigan, 1994.

[5] Arun Hampapur and Ruud Bolle. Design of Real Time Video Annotation Systems. Technical report, IBM T J Watson Research Center, P.O Box 218, Yorktown Heights, NY 10598, 1997.

[6] Arun Hampapur and Ramesh Jain. Video data management systems: Metadata and architecture. In *Multimedia Data Management*. McGraw-Hill Series on Data Warehousing and Data Management, 1998.

[7] P J Harari and Dave Ominsky. *Basketball Made Simple: A Spectators Guide*. First Base Sports Inc, 1994.

[8] Virage Inc. *www.virage.com*. 1997.

[9] Stephen Intille. Tracking using a local closed-world assumption: Tracking in the football domain. Technical report, M I T Media Lab, Perceptual Computing Group, 1994.

[10] Gopal Pingali, Yves Jean, and Ingrid Carlbom. Real time tracking for enhanced tennis broadcasts. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 1998.

[11] Zane Publishing. *The Merriam-Webster Dictionary*. Zane Publishing, 1998.

[12] Rich and Knight. *Artificial Intelligence*. McGrawHill, 1991.

[13] Deerwester S, Dumais S T, Furnas G W, Landauer T K, and Harshman R. Indexing by latent semantic analysis. *Journal of the Society of Information Science*, 41(6):391–407, 1990.

[14] Drew Saur, Yap-Peng Tan, Sanjeev Kulkarni, and Peter Ramadge. Automated analysis and annotation of basketball video. In *Image and Video Databases*, San Jose, California. SPIE.

[15] Deborah Swanberg, Chiao-Fe Shu, and Ramesh Jain. Knowledge guided parsing in video databases. In *Electronic Imaging: Science and Technology*, San Jose, California, February 1993. IST/SPIE.

[16] Dennis Yow, Book-Lock Yeo, Minerva Young, and Bede Lui. Analysis and presentation of soccer highlights from digital video. In *Proceedings of the Second Asian Conference on Computer Vision*, pages 499–503, Singapore, December 1995. IEEE.

[17] HongJiang Zhang, Yihong Gong, Stephen W Smoliar, and Shuang Yeo Tan. Automatic parsing of news video. In *Proceedings of the IEEE Conference on Multimedia Computing Systems*, May 1994.

Knowledge Model	Semantic Model	Game Model	Temporal Model
		Violations & Fouls	Rule Base
		Team Structure	Hierarchical Model
	Cinematic Model	Composition Model	Audio/Video Mixing
			Editing Models
		Imaging Model	Camera Geometry Model
			Camera Motion Model
	Physical World Model	Environment Model	Court/Field Models
		Object Model	Player, Ball Models

Figure 1: The Structure of the Knowledge Model showing the semantic, cinematic and physical world model components.

	Time	Speech Transcript from Closed Captioning
2	0:30 - 1:00	**>> John:** Starks wide open, of course nobody saw him, he was downcourt quickly. **>> Mike:** Johns on has mourning on him, mourning with two fouls, crowd urging johnson to take mourning. Kicks it out, childs thebaum movement, houston a three. Too far, hardaway getting it on the rebound. Murdock all the time,that three-pointer will not go, alonzo mourning and starks called for the reach-in foul. Two fouls now on starks.
3	1:00 - 1:30	**>> John:** Yes, it has that look, that's for sure. Mourning gets the good bounce and starks trying to stop him commits the foul. **>> Mike:** Cummings will sit down, after playing eight minutes, didn't score, two rebounds, oakley back in. Approaching the midway point of the second, knicks trailing by five, they've been down by as many as eight. Hardaway off the dribble. Tim hardaway creating his own shot, knocks it down hardaway now five of eight from the field.
4	1:30 - 2:00	**>> John:** Can't play good defense on a player like hardaway by having to run at him when ehe gets the ball. Just going to go by you and get a jump shot as he just did. You've got to stay closer to him. **>> Mike:** Johns on gets it and he's fouled. Childs threading the needle on that pass.

	Time	Visual Content Description
6	0:55 - 1:02	Player who committed foul (John Starks)
7	1:02 - 1:13	Replay of Foul
8	1:13 - 1:14	Substitution (Player Walking Out)
9	1:14 - 1:18	Substitution (New Player Walking In)
10	1:18 - 1:33	**Play**,Knicks Basket in view, shot scored by Hardway
11	1:33 - 1:35	Player who scored the shot (Hardaway)
12	1:35 - 1:46	**Play** Ball being moved by Knicks towards the Heat Basket ends with foul committed by Heat
13	1:46 - 1:51	Player who was fouled (Johnson)
14	1:51 - 1:54	Player (Johnson) getting ready for free throw
15	1:54 - 2:01	Heat player talking
16	2:01 - 2:07	Player (Johnson)getting ready for free throw (bouncing the ball)
17	2:07 - 2:11	Free throw being attempted by Johnson
18	2:11 - 2:19	Player (Johnson) getting ready for next free throw
19	2:19 - 2:33	**Play** Free throw complete, ball in-bounded and moved towards Knicks Basket
20	2:33 - 2:51	**Play** Heat advance toward Knicks Basket. Basket scored by Mashburn
21	2:51 - 2:55	Player (Mashburn) walking back
22	2:55 - 3:00	**Play** Knicks moving the ball towards the Heat Basket

Table 1: The Speech and Video Shot annotation tracks for the 3 minute segment of the basketball game between the **NY Knicks** and the **Miami Heat**. The *Time* column is the offset (in minutes) into the video. **Top Table:** This is the closed caption transcript of the speech during the 3 minute video clip. **Bottom Table:** Visual Content Annotation of the scenes

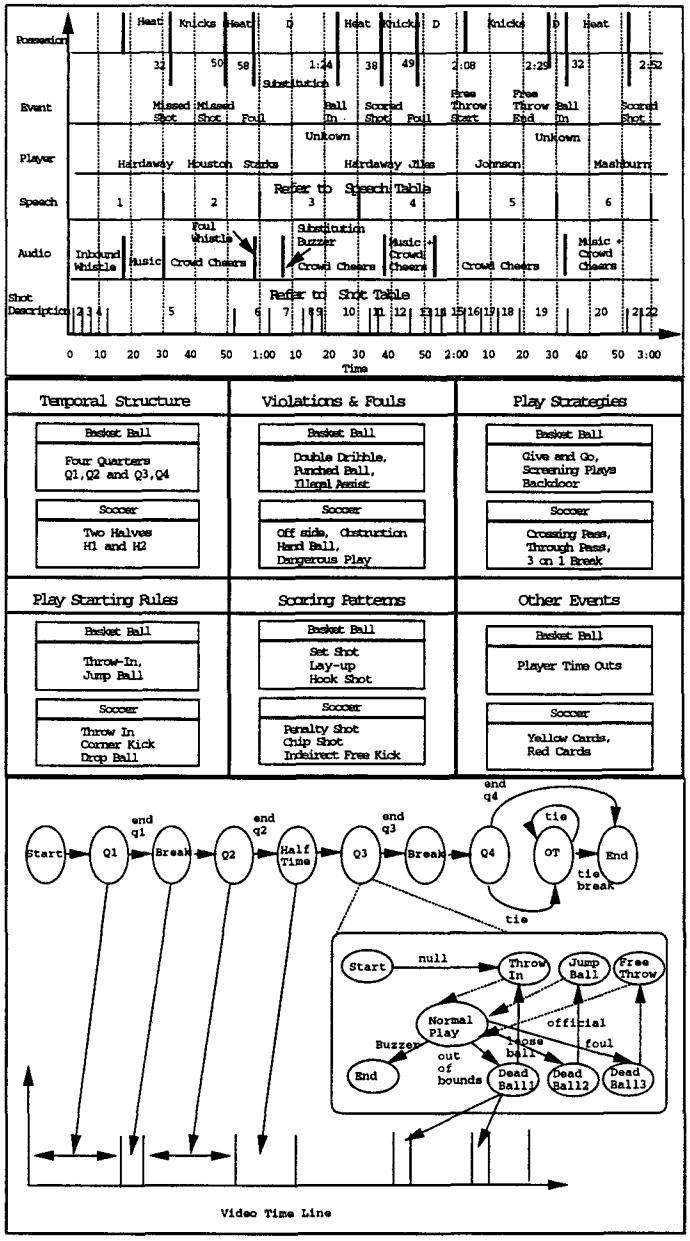

Figure 2: **Top:**The annotated representation of a 3 minute basketball video segment. **Middle:** Abstract Event Categories across two domain-instances. **Bottom** The semantic model of a basketball game.