Automatic Detection of 'Goal' Segments in Basketball Videos

Surya Nepal
CSIRO Mathematical and Information Sciences
Locked Bag 17, North Ryde
NSW 1670 Australia
61 2 93253148
sutya.nepal@cmis.csiro.au

Uma Srinivasan
CSIRO Mathematical and Information Sciences
Locked Bag 17, North Ryde
NSW 1670 Australia
61 2 93253150
uma.srinivasan@cmis.csiro.au

Graham Reynolds
CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra
ACT 2601 Australia
61 262167009
graham.reynolds@cmis.csiro.au

ABSTRACT

Advances in the media and entertainment industries, for example streaming audio and digital TV, present new challenges for managing large audio-visual collections. Efficient and effective retrieval from large content collections forms an important component of the business models for content holders and this is driving a need for research in audio-visual search and retrieval. Current content management systems support retrieval using low-level features, such as motion, colour, texture, beat and loudness. However, low-level features often have little meaning for the human users of these systems, who much prefer to identify content using high-level semantic descriptions or concepts. This creates a gap between the system and the user that must be bridged for these systems to be used effectively. The research presented in this paper describes our approach to bridging this gap in a specific content domain, sports video. Our approach is based on a number of automatic techniques for feature detection used in combination with heuristic **rules** determined through manual observations of sports footage. This has led to a set of models for interesting sporting events -goal segments- that have been implemented as part of an information retrieval system. The paper also presents results comparing output of the system against manually identified goals.

Keywords

Content-based retrieval, temporal models, sports video analysis.

1. INTRODUCTION

Organizations with large digital assets have a need to retrieve meaningful information from their collection. Applications such as digital libraries, video-on-demand systems and interactive video applications introduce new challenges in managing large collections of audio-visual content. Content-based retrieval systems have tried to address this need by supporting search and retrieval of specific features [5,7,8,9]. Most of these systems support searches on low-level features such as colour, texture, shape of objects and images. Similarly, video retrieval applications are based on analysis of audio-visual content and retrieve camera-operations such as pan and zoom, track moving objects, and so on [1,2,3,4]. Retrieving content using low level features alone is not adequate to capture the inherent semantics of a visually and acoustically rich medium. We need to develop applications that allow users to interact with the content at a semantic level that is more natural in a given context. The difficulty in supporting semantics lies in the gap between low-level media features and high-level concepts. While this is extremely difficult for videos in general, limiting analysis to a specific domain can help in identifying high-level semantics relevant to that domain. The goal of this work has been to raise the semantic level of interaction with the content, by identifying segments that are meaningful in a given context. In this paper we explore the sports domain, in particular basketball games.

A meaningful segment in a basketball game is a segment that shows a shooter scoring a goal. Our approach for identifying such meaningful segments in videos of basketball games is shown in Figure 1. The first step consists of manually observing television broadcasts and videos of basketball games, and identifying certain common repetitive events that are present throughout the game. We refer to these as key events. Secondly, we develop temporal models based on the pattern of occurrence of the key events observed during the course of the game. Next, we use these key events to direct or motivate our automatic feature analysis techniques. The automatically extracted features are used to verify and validate the temporal models, which in turn are used to automatically identify meaningful segments from a video broadcast of a basketball game. Our focus is on automatically identifying goal segments in a basketball video using the temporal models we have developed. We have conducted a number of experiments on a set of basketball videos to demonstrate the relative merits of each temporal model using both manual observation and automatic detection. The results are quite encouraging. We believe that the same technology can be used to determine other basketball specific high-level concepts such as "fouls", "free throws", "field goals" and "successful goals". We also believe that many other sports could use similar technology. For example, the appearance of a scoreboard is used to calculate the likelihood of a "goal" segment, and the same technology could be used in other sports such as rugby, football and soccer.

MM'01, **Sept.** 30-Oct. 5, 2001, **Ottawa, Canada.**

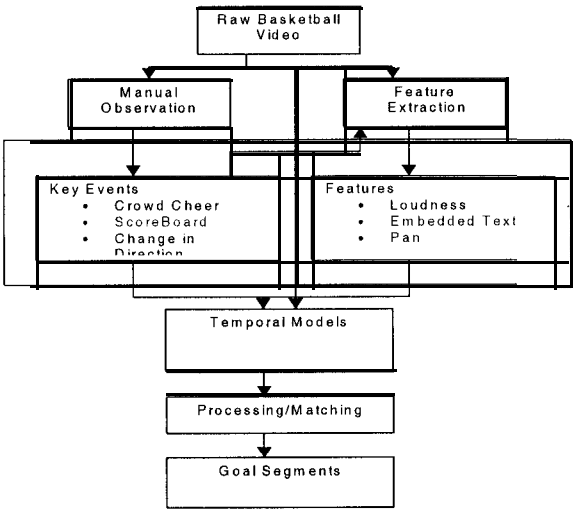


Figure 1: A flowchart to identify goal segments.

The rest of the paper is organized as follows. We describe related work from a semantic analysis perspective in Section 2. In Section 3, we present our manual observations and identification of key events that occur during a basketball game. We then present various temporal models developed based on the key events. In section 4, we describe audio-visual features that capture key events identified during our observations. The experimental results are presented in Section 6. Finally, we draw some conclusions.

Table 1: A list of data set used for our experiments and the result of our manual observations.

Games	Description	Length	Number of Cheers	Number of Scoreboard displays	Number of Goals	Number of changes in direction of players' movements
A	Australia Vs Cuba 1996 (Women)	00:42:00	31	49	52	74
B	Australia Vs USA 1994 (Women)	00:30:00	27	46	30	46
C	Australia Vs Cuba 1996 (Women)	00:14:51	16	16	17	51
D	Australia Vs USA 1997 (Men)	00:09:37	13	16	18	48

2. BACKGROUND

There have been some recent attempts to exploit domain knowledge and inherent properties of low-level features for automatic detection of high-level concepts. Most of them attempt to relate or map the low-level information measured from video data to high-level concepts [11,12,13,14,15]. Adams et al. [II] exploit the existing cinematic conventions known as film grammar to define and detect tempo of a movie. Their approach is based on the notion that film sections of differing *pace/tempo* will leave distinct marks on the features of shot length and motion, and hence may be detected and classified based on the primitive media features shot length and motion. Yoshitaka et al. [12] use spatio-temporal correlations of objects to detect a certain semantic content, which can be perceived by considering the combined motion of multiple objects. They use a set of relational descriptors to define several techniques that are commonly performed by soccer players, namely "wall pass", "overlap", "through pass", and "zone pass". Rui et al. [23] present techniques to automatically generate highlights for TV baseball programs using audio features. Naphade et al. [13] use hierarchical hidden Markov models for modeling audio-visual signatures of events such as "explosions". They use a *multiject*, which is a probabilistic multimedia object that has a semantic label and that summarizes a time sequence of a set of features extracted from single or multiple media. Saur et al. [14] present a method for automatic analysis and annotation of basketball video. They use low-level information available directly from MPEG compressed video of basketball, as well as nrior knowledge of basketball video structure, to provide high-level content analysis such as "close-up views", "fast breaks" and "steals". Sudhir et al. [15] exploit the available domain knowledge about tennis video and demonstrate that it is possible to generate semantic annotations applicable to that domain. They generate useful high-level semantics such as "baseline-rallies", "passing-shots" "net-games", and "serve-and-volley games" using automatically extracted information such as "tennis court lines" and "players' position".

3. OBSERVATIONS AND ALGORITHMS

3.1 Manual Observations

The manual observation of television broadcasts of basketball games gives some insights into commonly occurring patterns of events perceived during the course of a basketball game. We have considered a subset of these as key events that occur repeatedly throughout the video of the game, as follows:

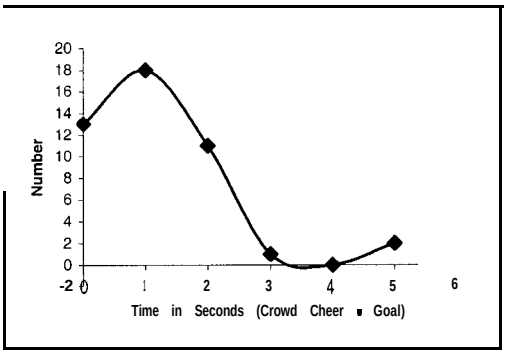
I. **Crowd cheer:** In a basketball game, the most exciting moments occur when a player shoots a successful goal, thereby increasing the score of the shooter's team. The game becomes particularly interesting and gains momentum when there is a successful field goal which

increases the score by two or three points, depending on the position from where the shooter scores the goal. Such exciting field goals are marked by high excitement from the audience, which in turn manifests itself in the form of loud cheers from the crowd after each successful field goal. It is also noticeable that while field goals elicit a loud response from the crowd, a successful free throw has a milder response, which is understandable as the score increases by just one point in such cases. Table I column 4 shows the number of manually recorded crowd cheers for the basketball videos in our data set.

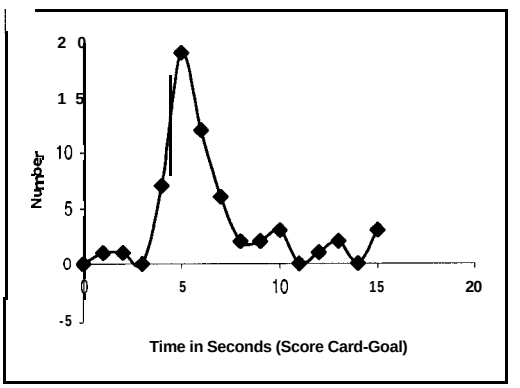
2. **Scoreboard display:** In television broadcasts, the latest score tally of the two teams are displayed at frequent intervals in order to keep the viewers up to date with the current status of the game. As can be expected, the scoreboards are updated after each valid goal. The scoreboard is normally displayed as separate embedded text after each legal goal. Hence, we use the displayed scoreboard as a second key event in television broadcasts of basketball games. Table I column 5 shows the number of scoreboard displays shown throughout the game. Most of them show the change in score after goals. Embedded text is normally inserted in the video during the post-processing time. Hence, this event depends on the production styles. However, our observations of several basketball videos indicate that each valid goal is reflected in embedded text in most of the existing production styles. In earlier production styles, embedded text appears at the center of the screen after each valid goal. In recent production styles, a scoreboard is displayed at the left-top or right-bottom corner of the screen throughout the game and the colour of the text is changed for a few seconds after each valid goal. We may need to apply different algorithms to identify such events for different production styles, but the event as such is general for basketball videos.

3. **Change in direction:** Typically, a field goal attempt starts when a player begins the motion that precedes the actual shot. It continues until the shooting effort ceases and the shooter returns to a normal floor position. Soon after a field goal attempt, most of the players will start to move away from the basket where the field goal was attempted. That is, there is a change of direction of most players in the field. Change in direction of motion also occurs after other game events such as fouls, free throws and throw-ins. Hence, we consider a change in direction of motion as a key event in basketball videos. This change is normally reflected in camera motion as the camera focuses on players. Table 1 shows the number of change in direction by camera in our experimental videos.

Using the patterns occurrence of these key events, we have developed temporal models that aid the classification of a basketball video into meaningful segments. Our objective is to determine models that can facilitate the automatic identification of segments that show the most exciting parts of a basketball game, for example, segments where a shooter scores a legal goal. If we are able to achieve this type of semantic segmentation, a broadcaster/webcaster can automatically generate segments that show the highlights of the basketball game, and use these segments for news broadcasts, headline news flashes, and so on.



(a)



(b)

Figure 2: (a) Temporal Model H1. X-axis: Number of crowd cheers, Y-axis: Interval between start of crowd cheer and goal time (b) Temporal Model H2. X-axis: Number of scoreboard displays that accompanied the goal, Y-axis: Interval between scoreboard start time and goal time for games A and B.

3.2 Temporal Models

Model

This model is based on the first key event ▪ crowd cheer (refer to Section 5). Our observation shows that there is a loud cheer within 3 seconds of scoring a legal goal. Figure 2(a) illustrates that more than 95% crowd cheers that accompanied goals fall within this observation. Hence, in this model, the basic assumption is that a loud cheer follows every legal goal, and a loud cheer only occurs after a legal goal. The model H1 is represented by

H1: Goal → [3 sec] → crowd cheer

Intuitively, one can see that the converse may not always be true, as there may be other cases where a loud cheer occurs, e.g., when a streaker runs across the field. Such limitations are addressed to some extent in the subsequent models.

Model II

This model is based on the second key event ▪ scoreboard display. Our observation shows that the scoreboard display is updated

after each goal. Our assumption in this model is that a scoreboard display appears (usually as embedded text) within 10 seconds of scoring a legal goal. Again, we observed that more than 95% of scoreboard displays followed this rule, as shown in Figure 2(b). This is represented by the model H2.

H2: Goal → [10 sec] → Scoreboard

The limitation of this model is that the converse here may not always be true, i.e., a scoreboard display may not always be preceded by a legal goal.

H3: Goal-+ [3 sec] → Audio Cheer+ [7 sec] → Score Board

Model IV

This model addresses the strict constraints imposed in Model 3. Our observations show that while the pattern shown in 3 is valid most of the times, there are cases where the field goals are accompanied by loud cheers and no scoreboard display, and similarly there are cases where there are goals followed by scoreboard displays but no crowd cheer, as in the case of free throws. In order to capture such scenarios, we have used a combination of models I and 11 and proposed a model IV.

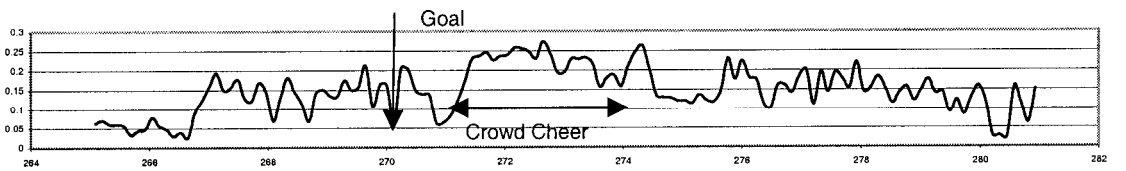


Figure 3: The figure shows the audio segments of a basketball game. The Audio segment shows the loudness curve. The "goal" occurs at 270 (seconds) as shown in video segment and the crowd cheers occurs from 271 to 274.

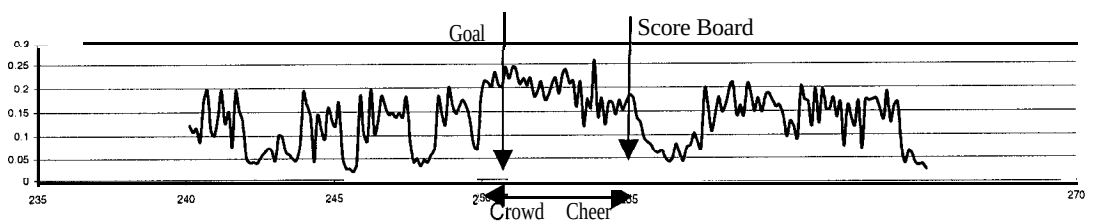


Figure 4: The figure shows the audio segments of a basketball games. The audio segment shows the loudness curve. The "goal" occurs at 251 (seconds), the crowd cheer occurs from 250 to 254 and the scoreboard appears at 255.

Model III

This model uses a combination of two key events with a view to address the limitations of HI and H2. As pointed out earlier, all crowd cheers and scoreboard displays may not always indicate a legal goal. Ideally when we classify segments that show a shooter scoring goals, we need to avoid inclusion of events that do not show a goal, even though there may be a loud cheer. In this model this is achieved by temporally combining the scoreboard display with crowd cheer. Here, we assume that every goal is followed by crowd cheer within 3 seconds, and by a scoreboard display within 7 seconds after the crowd cheer. This discards events that have crowd cheer, but no scoreboard and events that that have scoreboards. but no crowd cheer

H4: H1 u H2

where u is the union of results from models HI and H2.

Model

While model IV covers most cases of legal goals, due to the inherent limitations pointed out in models I and II, model IV could potentially classify segments where there are no goals. Our observations show that Model IV captures the maximum number of goals, but it also identifies many non-goal segments. In order to retain the number of goal segments and still remove the non-goal segments, we introduce the third key event ▪ change in direction of players. In this model if a crowd cheer appears within IO

seconds of a change in direction or a scoreboard appears within 10 seconds of a change in direction, there is likely to be a goal within 10 seconds of the change in direction. This is represented as follows.

H5: Goal → [10 secs] → Change in direction+ [10secs] → Crowd cheer
OR
Goal → [10 secs] → Change in direction+ [10secs] → Scoreboard

Although, in all models the time interval between two key-events is hardwired, the main idea is to provide a temporal link between key-events. We believe that these temporal models can be used in other applications by setting the suitable time interval.

4. AUDIO-VISUAL FEATURES USED

In this section, we identify the media features that correspond to the key events manually observed, and give an overview of how they are automatically derived from MPEG videos. The features identified include energy levels of audio signals, embedded text regions, and change in direction of motion. Since the focus of this paper is on temporal models, we will not elaborate the feature extraction algorithms in detail. For details, readers are referred to [21, 22, 25].

Audio Analysis of High Energy segments

The Audio-MPEG- I coding standard encodes the characteristics of 32 frequency bands of the audio signal. For each of these sub-bands it calculates the value of its amplitude 1500 times per second. Each group of 12 values forms a frame. The amplitudes for each frame and each subband are divided by a ScaleFactor such that the largest amplitude is unity. These scaled values are encoded along with the ScaleFactor. The ScaleFactor is therefore the magnitude of the maximum value for this sub-band and this frame. The Sum of the 32 Scale Factors (SSF) corresponding to each of the 32 subbands provides a reasonably good approximation to the maximum value of each acoustic signal in this frame. We use an audio analysis toolkit called MPEG Maaate [20] developed in our lab to identify all such segments where the SSF value is high. Figures 3 and 4 show the loudness curve obtained from our analysis, along with the corresponding key frames of the video segment of a basketball game.

As can be seen, the high-energy points closely follow the occurrence of goals, and fit in with our temporal model H1. We then use the SSF value as the basis to identify our key event, crowd cheer from the audio signal. We do this by applying appropriate filters and threshold values to the SSF values. We refer readers to [25] for further details on the crowd cheer detection algorithm.

Video Analysis: Text Detection

The MPEG video stream has three types of frames: I-, P- and B-frames [6]. An I-frame is completely intra coded, i.e., only information from the current frame is used. Each frame is divided into 16x16 macroblocks and each macroblock composed of four 8x8 luminance blocks and two chrominance blocks. Each block is Discrete Cosine Transformation (DCT) transformed and the coefficients then quantised and entropy encoded. A P-frame is predictively coded with motion compensation from a past I- or P-frame. Each macroblock in the current P-frame is matched to the most similar macroblock in the reference frame. The displacement between the two matched macroblocks is the motion vector of the current macroblock. If no match is found within the specified search region for a macroblock, it will be intra-coded. A B-frame is similar to a P-frame except that it uses a future reference frame as well as a past reference.

Typically, text could appear anywhere in a video. There are two types of text that can be seen in videos: scene text, i.e. text that is an inherent part of the scene, and artificial text, i.e., text that is inserted externally. We are interested in the artificial text, as scoreboards belong to this category. Artificial text is external to the video production process, and is overlaid over a scene in the post-processing stage. Since a text region is characterised by sharp edges, it should have a strong high spatial frequency. We therefore use the AC components, instead of the DC component, to retain the sharp edges that are invariant over time. The AC coefficients of each 8x8 luminance block in an I-frame represent the high frequency response of the block. We use the sum of moduli of all the 63 AC coefficients of each luminance block and check for temporal stability to extract text regions. We refer readers to [2 1] for details.

Figure 3 shows a sample set of frames of a basketball video segment, and corresponding loudness curve for the segment. As can be seen a scoreboard display starts at frame 255 soon after a high-energy segment that occurs between frames 250 and 255. This fits in well with our models H2 and H3.

Video Analysis: Change in Motion Direction

In basketball videos, the camera normally focuses to the players and the camera moves or pans along with movements of the players. We use the camera panning operation to determine the change in direction of motion of the players. The motion vector field induced by the camera panning operation exhibits a single strong modal vector value, which corresponds to the direction of camera movement. Therefore the majority of the motion vectors will have the same direction. Our method calculates the motion vector angles of all macroblocks with valid motion vectors in each P-frame and thereafter the angle histogram. If this angle histogram reaches a peak in one of the directions, the related frame will be marked as a potential panning point. Usually a camera panning operation lasts for several frames and the direction of movement during this period remains constant. Therefore, only if there is a set of such consecutive potential points and motion vector directions in each frame are similar, a camera panning can be declared. The panning direction is determined from the location where the peak occurs. We use the panning direction to determine the change in direction by players. We refer readers to [22] for further details.

5. EXPERIMENTAL RESULTS

The data set used to evaluate the proposed algorithms for detecting goal segments in basketball videos is shown in Table I. The first two clips A and B are used for evaluating manual observations and the last two clips are used for evaluating automatic analysis. For our manual observations, we used additional data, broadcast by different TV channels not shown in Table 1.

5.1 EVALUATION FRAMEWORK

Two standard methods of evaluation: precision-recall and false positive-negative were used to compare our automatically generated goal segments with the ones manually judged by humans. Precision and recall is based on the notion that for any given query, the collection of images in the database can be partitioned into two sets. The first, the relevant set, consists of clips relevant to the query, where relevance is decided by human judgment. All other clips are then irrelevant. Ideally, the system retrieved set (in response to a query) should be the same as the manually retrieved set. Usually this is not so. Recall/precision defined below are essentially measures of the relative number (expressed as a percentage) of the relevant and irrelevant sets between human judgment and system generated results.

Precision is the ratio of the number of relevant clips retrieved to the total number of clips retrieved. The ideal situation corresponds to 100% precision, when all retrieved clips are relevant.

$$precision = \frac{relevant \cap retrieved}{retrieved}$$

Recall is the ratio of the number of relevant clips retrieved to the total number of relevant clips. We can achieve ideal recall (100%) by retrieving all clips from the data set, but the corresponding precision will be poor.

$$recall = \frac{|relevant \cap retrieved|}{|relevant|}$$

False positive and negative are based on the following two hypotheses.

Let R be a result set and TS a temporal segment

Consider hypotheses HO and H1, where

HO = TS is a "goal" segment and

H1 = TS is not a "goal" segment

If TS $\in$ R then decide HO is true

If TS $\notin$ R then decide H1 is true.

A decision based on the result set R returned by the algorithm leads to three possible outcomes:

1. a correct decision (CD)- leads to HO when HO is true, or H1 when H1 is true.
2. a false alarm (or false positive) (FP)- leads to HO when H1 is true
3. a false negative (FN)- leads to H1 when HO is true.

Table 4: A summary of results that show the performance of our audio-visual features capturing corresponding events. The match column shown the percentage of manually detected segments that match with automatically detected segments.

Games	Crowd Cheer			Scoreboard		
	Manual	Automatic	Match %	Manual	Automatic	Match %
C	16	15	86.66	16	20	80.00
D	13	16	81.25	16	19	80.00

Table 2: A summary of our manual evaluation of various algorithms.

Clips	Total number of baskets (*relevant*)	Algorithms	Total (*retrieved*)	Correct Decision (*relevant* $\cap$ *retrieved*)	Precision (%)	Recall (%)	FP	FN
A	52	H1	31	29	93.54	55.76	2	23
		H2	49	30	61.22	57.69	19	22
		H3	17	17	**100**	32.69	0	35
		H4	84	42	50.00	**80.76**	42	10
		H5	37	31	83.78	59.16	6	21
B	30	H1	27	16	59.25	53.33	11	14
		H2	46	24	52.17	**80.00**	22	6
		H3	10	10	**100**	33.33	0	20
		H4	63	30	47.61	**100**	33	0
		H5	31	20	64.51	66.66	11	10

Let N_w be the set of manually determined goal segments and R_w be the automatically derived set of goal segments. We then define:

$CD = count(R_w \cap N_w)$

$FP = count(R_w) - count(R_w \cap N_w)$

$FN = count(N_w) - count(R_w \cap N_w)$

where $count(X)$ returns the number of events in the set X.

The ideal situation corresponds to 100% CD and $FP = FN = 0$.

Next, we compare the various models to test their results against this criterion.

5.2 RELATIVE PERFORMANCE AND DISCUSSION

We used two methods to compare the temporal models proposed in this paper: manual and automatic. The manually identified "legal goals" (which include field goals and free throws) of the videos in our data set are shown in column 2 in Tables 2 and 3, indicated using start and end time. Both manual and automatic evaluation of the models are then performed against the manually recorded legal goals.

Manual evaluation involved two human subjects viewing the video games and recording the timing for key events: crowd cheer, scoreboard and change in direction by players. While it would be interesting to have both start and end time for each key event, this was not manually possible. Hence, only the start time of each key event was recorded. For example, we recorded the time when the players start to move from left to right as a time of change in direction. We then matched such recorded timings against the manually recorded legal goals for each model. The results of manual evaluation are shown in Table 2. The model based on crowd cheer followed by scoreboard (H3) gives very high precision (100%) and low recall (-33%). That is, the result set contains only goal segments (FP = 0), and about one third of total legal goals (high FN). This model might be useful for applications that need to identify the sports segments for the evening news or creating highlights for TV programs. The model based on crowd cheer union scoreboard (H4) gives very high recall (-80%). That is, the result set contains most of the legal goal segments. This model may be useful in application areas such as interactive video playback, where the result set acts as a guide for watching the game. A user who does not like a particular segment could simply skip to the next segment of interest, or watch a whole segment if they like it. In such situations, the low precision value can be compensated for by the user. The model based on all three key events (H5) yields average precision and recall values. We can thus conclude that we can use different models in different application areas based on their needs and requirements.

We next evaluate our models using automatically detected goal segments. The aim of automatic analysis is two-fold. The first is to evaluate the choice of audio-visual features used to represent key events and the second is to evaluate the model automatically using those extracted features against the set of legal goals judged by humans (column 2 Table 3). The result of this analysis is shown in Table 3.

We see that models H3 and H4 correspond to the manual observation in Table 2 - H3 has high precision (100%) and H4 has high recall (94.1%). The result of automatic analysis shows that the combination of all three key events performs much better with high recall and precision values (-88%). Overall the results of automatic analysis follow the manual observations (see Tables 2 and 3). That is, the audio-visual features used in our experiments capture most of the manually identified key events (-80% see Table 4). In model HI, the model correctly identifies 12 goal

Table 3: A summary of results of automatic evaluation of various algorithms.

Clips	Total number of baskets (*relevant*)	Algorithms	Total (*retrieved*)	Correct Decision (*relevant* ∩ *retrieved*)	Precision (%)	Recall (%)	FP	FN
C	17	H1	15	12	80	70.50	3	5
		H2	20	15	75	88.23	5	2
		H3	11	11	**100**	64.70	0	6
		H4	24	16	66.66	94.1	8	1
		H5	17	15	**88.23**	**88.23**	2	2
$\overline{D}$	18	H1	16	11	68.75	61.11	5	7
		H2	19	17	89.47	94.44	2	1
		H3	10	10	**100**	55.55	0	8
		H4	25	18	72.0	**100**	7	0
		H5	22	16	**72.72**	**88.88**	6	2

segments out of 17 for video C and I 1 out of 18 for video D. However, the total number of crowd segments detected by our crowd cheer detection algorithm is15 and16 for videos C and D, respectively. That is, there are few legal goal segments that do not have crowd cheer and vise versa.

Further analysis shows that crowd cheer events resulting from other interesting events such as "fast break", "clever steal" and "great check or screen" gives false positive results. We observed that most of the field goals are accompanied by crowd cheer. However, many goals scored by free throws are not accompanied by crowd cheer. We also observed that in certain cases lack of supporters among the spectators for a team yield false negative results. In model H2, the model correctly identifies 15 goals out of 17 in video C and 17 out of 18 in video D. Our further analysis confirmed that legal goals due to free throws are not often accompanied by scoreboard displays, particularly when the scoreboard is updated at the end of free throws rather than after each free throw. Similarly, our feature extraction algorithm used for scoreboard not only detect scoreboards but also other textual displays such as team coach names and the number of fouls committed by a player. Such textual features increase the number of false positive results. We plan to use the heuristics developed here to improve our scoreboard detection algorithm in future. The above discussion is valid for algorithms H3, H4 and H5 as well. We refer readers to Section 3 for further theoretical limitations and merits of each of these models.

6. CONCLUDING REMARKS

Adaptation of early content-based systems has been slow, as users of content collections prefer to query video content at a conceptual level rather than at a feature level that is supported by most of the existing systems. Our aim in this work has been to bridge this gap by raising the semantic level of interaction with the content. We have proposed a range of temporal models that combine the results of feature extraction techniques with domain specific knowledge to automatically identify meaningful segments ▪ goal segments ▪ in basketball videos. Domain knowledge was gained through manual observations where we identified key events that occur around an important game event. For instance, when a shooter scores a goal, we found three key events: crowd cheer, scoreboard display and change in direction of the players. Having identified the key events, we then used feature extraction techniques that correspond to these manually observed events. While we have used only three key events in this case, other domains may have more key events that could contribute to a meaningful segment. The point to note here is that we have used these key events to guide the feature extraction work. We then developed different temporal models that help to identify 'goal' segments based on these key events. Working with different models gives us a way to gently guide the process of extracting meaningful segments in a given context. While the models presented here are specific to this domain, the idea of developing appropriate models for the identified key events is the message we wish to convey here. Our results with identifying 'goal' segments in basketball videos have been encouraging. In order to extract meaningful segments, we believe that we need to understand the domain and develop such models to bridge the gap between features and semantics. More work needs to be done to generalise the ideas presented in this paper for different domains.

7. REFERENCES

[1] Myron Flickner , Harpreet Sawhney ,Wayne Niblack, Jonathan Ashley, Qian Huang , Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic and David Steele. Query by Image and Video Content: The QBIC System. Computer. Vo;l 28, No. 9, 1995, pages 23-32.

[2] Shih-Fu Chang, William Chen, Horace J. Meng, Hari Sundaram and Di Zhong. VideoQ: An Automated Content Based Video Search System Using Visual Cues. ACM Multimedia 97, Seattle WA, Nov 1997.

[3] J. R. Smith and S.-F. Chang. Querying by color regions using the VisualSEEk content-based visual query system. In M. T. Maybury, editor, Intelligent Multimedia Information Retrieval. IJCAI, 1997.

[4] Arun Hampapur and Ramesh Jain. Virage Video Engine. SPIE Vol. 3022. pp 188-197.

[5] Atsuo Yoshitaka, and Tadao Ichikawa. A Survey on Content-Based Retrieval for Multimedia Databases. IEEE Transactions on Knowledge and Data Engineering, Vol 11, No I, Jan/Feb 1999.

[6] Barry G. Haskell, Atul Puri and Arun N. Netravali. Digital Video: An Introduction to MPEG-2. Chapter 4: Audio. Chapman & Hall International Thomson Publishing. 1996.

[7] Rainer Leinhart, Wolfgang Effelsberg and Ramesh Jain.VisualGREP: A Systematic Method to Compare and Retrieve Video Sequences. Multimedia Tools and Applications, Vol 10. pp 47-72.

[8] Hari Sundaram and Shih-Fu Chang. Efficient Video Sequence Retrieval in Large Repositories. SPIE '99 Storage and Retrieval of Image and Video Databases VII, San Jose CA, Jan 24-29, 1999.

[9] N.Bryan-Kinns. VCMF: A Framework for Video Content Modeling. Multimedia tools and Applications, Vol 10, pp 23-45, 2000.

[10] C.Carson and V.E. Ogle. Storage and retrieval of feature data for a very large online image collection. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 19(4):19-27, December 1996.

[11] Brett Adams, Chitra Dorai and Svetha Venkatesh. Study of Shot Length and Motion as Contributing Factors to Movie Tempo. ACM Multimedia 2000. pages 353-355.

[12] Atsuo Yoshitaka, Yuichi Hosoda, Masahito Hirakawa, and Tadao Ichikawa. Content-Based Retrieval of Video Data Based on Spatiotemporal Correlation of Objects. In Proc. IEEE Multimedia Computing and Systems, 1998. pp. 208-213

[13] M. Naphade, T. Kristjansson, B. Frey, and T.S. Huang. Probabilistic multimedia objects (multijects): A novel approach to indexing and retrieval in multimedia systems. In Proceedings of the fifth IEEE International Conference on Image Processing, Volume 3, pages 536-540, Chicago, IL, Oct 1998.

[14] Drew D. Saur, Yap-Peng Tan, Sanjeev R. Kulkarni, and Peter J. Ramadge. Automatic analysis and Annotation of Basketball Video. In Storage and Retrieval for Image and Video Databases V, volume SPIE-3022, pages 176-187, Feb. 1997.

[15] G. Sudhir, John CM. Lee, and Anil K. Jain. Automatic Classification of Tennis Video for High-level Content-based Retrieval. Technical Report HKUST-CS97-2, The Hong Kong University of Science and Technology, Hong Kong, August 7, 1997.

[16] Davis Pan. A Tutorial on MPEG/Audio Compression. IEEE Multimedia. Vol 2, No. 2, 1995. Pages 60-74.

[17] Peter Noll. MPEG Digital Audio Coding. IEEE Signal Processing Magazine. Sept. 1997. Pages 59-8

[18] Surya Nepal, Uma Srinivasan and Graham Reynolds. Semantic-Based Retrieval Model for Digital Audio/Video. CSIRO Mathematical and Information Sciences. Technical Report No 2000/174. October 2000.

[19] Mediaware solutions. http://www.mediaware.com.au/

[20] MPEGMaaate: http://www.cmis.csiro.au/dmis/Maaate/

[21] Lifang Gu. Scene Analysis of Video Sequences in the MPEG Domain. Proceedings of the IASTED International Conference Signal and Image Processing October 28-3 I, Las Vegas, Nevada, U.S.A.

[22] Hongjiang Zhang, Chien Yong Low and Stephen W. Smoliar. Video Parsing and Browsing Using Compressed Data. Multimedia Tools and Applications, Vol 1, No I, March 1995.

[23] Yong Rui, Anoop Gupta and Alex Acero. Automatically Extracting Highlights for TV Baseball Programs. ACM Multimedia 2000, pages 105-l 15.

[24] Basketball Rules http://www.basketball.com/

[25] Uma Srinivasan, Jordi Robert-Ribes and Graham Reynolds. Querying Video Content Using Multi-Modal Features. CSIRO Mathematical and Information Sciences. Technical Report 1997.