SMART: A Grammar - Based Semantic Video Modeling and Representation

Vani Jain
Computer Science Department
University of Alabama in Huntsville
Huntsville, AL 35899, USA
vjain@cs.uah.edu

Ramazan Aygun
Computer Science Department
University of Alabama in Huntsville
Huntsville, AL 35899, USA
raygun@cs.uah.edu

Abstract

Multimedia database modeling and representation is significant for efficient storage and retrieval of multimedia. Modeling of semantic video content that enables spatiotemporal queries is one of the challenging tasks. In this paper, we provide a semantic modeling and retrieval system, termed as SMART. We provide a method that helps represent the semantic contents of video such as objects, events, and locations as a grammar based string. This linear string representation enables both the spatial and temporal description of the video. Various types of queries such as event-object-location, event-location, object-location, and event-object are supported by SMART. We show our results on the tennis video database.

Index Terms: - Video databases, video modeling, semantic video retrieval, spatiotemporal queries

1. Introduction

High bandwidth internet, high speed processors, and large storage devices have made videos more popular and easily available to everyone. Video conferencing, online videos, advertisements, sports, movies, and news have aimed at the users of all ages and professions. The increase in use of digital videos has brought a challenge of providing a competent scheme for modeling the video databases. The most successful video retrieval tools such as YouTube [5], Google videos [6] are keyword-based. For example, when a user uploads a file in YouTube, the user needs to enter the metadata description of the video.

So far, many video representations and extraction techniques have been proposed. In [1], the video data is transformed and represented as a stream of alphabets. It represents court, scoreboard, camera motion and audio of the sports videos. However, there is no representation for the objects such as the players and the events such as kicking ball. They have also not provided the queries that can help in retrieval of information, since they target video data mining. In [3], the effect created by the videos on the viewers such as thrilling scenes, are modeled as a graph. Another approach, is modeling of fuzzy information. In [4], the fuzzy information i.e. imprecise information is being modeled and managed in multimedia databases whereas our system, SMART, provides a linear string representation. A video management and application processing framework is been proposed in [10]. The main component of this framework is query based video retrieval mechanism that allows querying in its language, CAROL/ST [10] where as, SMART uses simple regular expressions provided by Structured Query Language (SQL). There has also been research on the representation of an image as a string [8].

A video can be modeled as spatial objects over time. In [7], the trajectory of moving object is modeled. The spatial representation for each object is provided with the help of minimum bounding rectangles and temporal interval algebra is used for representing the temporal relationships. Another model [9], proposes a Topological -Directional Model for the spatiotemporal contents of the video. A key frame is described with the help of relative positions of objects. The temporal contents are described by the set of these mutual spatial relationships. These representations do not model the semantic contents of the videos such as the events and does not provide a very efficient retrieval method. SMART represents the semantic contents and enables queries in SQL, the most popular and widely used query language. Although there has been research where the SQL is extended to describe various spatiotemporal queries [2] but the queries where events are one after the other are missing and also the representation of data is not very efficient.

There has been a significant research in the extraction of low level features of sports videos and mapping it to high level concepts [11] [12] [13] . Neil et al. [11] propose human behavior recognition method. They conducted experiments on the tennis videos where the events such as ‘service’, ’backhand at net’, and ‘walking at net’ were recognized. By using method in [13], player actions such as back-hand shot and fore-hand shot in tennis game can be predicted. The key events of the ball such as bounce and hit in a specific region of the tennis court can be found out using the tennis ball tracking algorithm [12]. The focus of

our research is modeling of this information in a grammar based manner that enables various types of queries such as event-object-location, event-location, object-location, and event-object.

This paper is organized as follows. The following section describes the representation of videos and its application on tennis videos. Spatiotemporal queries are explained in Section 3. The last section concludes the paper.

2. Video modeling and representation

Videos can be classified into two types with respect to its semantic content [1]. The first type covers videos having content structure such as movies, and the second one includes events such as sports videos. Our method provides a new approach for modeling semantic features of the sports videos. These semantic features are represented as string data that allows spatiotemporal queries using query language such as SQL.

2.1 Modeling of video data

In our model, we identify the main objects, events, locations and cameras (or camera views) in the videos and the temporal information, τ. These can be briefly explained as follows:

Objects. An object in a video is a region that has a semantic meaning and its spatial properties change over time. Object represents the main entity that performs some action. Objects are of prime interest to viewer of the clip. For example, 'players' in a sports video are objects. Each object *O* in a video is represented with an alphabet from domain $\Sigma_O = \{O_1, O_2, O_3, \ldots O_n\}$.

Events. Events represent the main occurrence in the video. Event is happening of some action by the object. When an object does something at a given a location and time, and attracts the viewer then, it can be termed as an event. For example, 'serving' is an event in a tennis game. Each event E is represented with an alphabet from the event domain $\Sigma_E = \{E_1, E_2, E_3 \ldots E_m\}$.

Locations. Spatial information is represented with locations (or positions). It represents the space occupied by the objects. For example, soccer field is the location for players and ball. The location can be determined semantically with respect to the regulations of the sports. Each region L can be represented from the location domain $\Sigma_L = \{L_1, L_2, L_3, \ldots, L_p\}$.

Cameras. In each video, various cameras (or camera views) provide different footage. For example, in a video first camera provides court view, second camera provides audience view, and another camera provides zoom coverage of players. Each camera C is represented with an alphabet from the camera domain $\Sigma_C = \{C_1, C_2, C_3, \ldots, C_q\}$.

In the videos each event occurs one after the other, thus we can represent the videos as a sequence string $S \in \{O_n, E_m, L_p, Cq\}^*$.The length of $|S|$ provides information in the temporal dimension, τ.

Grammar for spatiotemporal representation. Since we represent spatiotemporal content of a video as a string, the grammar is required to parse and extract the spatiotemporal information from the string. We can define the grammar for the representation of the sports videos as:

<video> :: = <sequence of clips>
<sequence of clips> :: =<clip>| <sequence of clips>
<clip> ::= <camera> "["<sequence of spt>"]"
<sequence of spt> ::= <spt>|<sequence of spt>
<spt> :: = [<event>] <obj> <loc>

where *video* is a sequence of clips; *clip* has a camera view and a sequence of spatiotemporal instances; and a spatiotemporal instance (*spt*) is represented with an object (*obj*), location (*loc*), and an optional *event*.

For example consider the subsequence $S_t = \{SG_2 10\}$. In S_t,
S represents the event saving goal, G_2 represents the goal keeper object, and 10 represents the location of the goal keeper.

2.2 Modeling and representation of tennis video

We explain the application of the above approach on the tennis videos. The representation and modeling of tennis videos are as follows:

2.2.1 Representation of tennis video

Here, we give brief information on tennis videos while explaining the representation of the tennis video content.

Objects. There are 3 main objects identified in the tennis game. The objects are identified as $\Sigma_O = \{U, V, b\}$ where U is the player who serves, V is the other player, and b is the ball.

Events. The main events are identified in the alphabet $\Sigma_E = \{F, B\}$ where F is the forehand shot and B is the backhand shot

Location. The tennis court is divided into regions by line segments to apply the rules of tennis game as in Figure 1.

For representation of locations and for semantic retrieval, we divide the court into partitions in the same way and apply the numbering in Figure1.

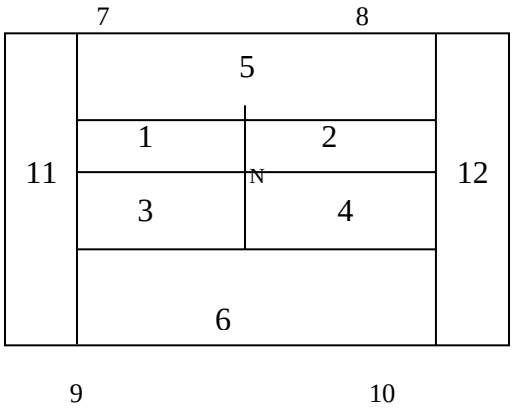

Figure 1. Tennis court segmentation

Cameras Views. We identify six types of camera views in a tennis game:

A – Gives a close view of the player at location 7 & 8 in Fig 1
B - Gives a close view of the player at location 9 &10 in Fig 1
C - Court view
D - Action Replay
R - Rest time
Com – Commentators

Grammar for Tennis Video Database
We now extend the grammar, mentioned before in the paper, for tennis videos.
<obj > ::= U|V|b
<event>::= F|B
<location> ::= 1|2|3|4|5|6|7|8|9|10|11|12|N
<camera> ::= <close-view camera>|C|D
<close-view camera> ::= A|B
<spt> :: = [<event>] <obj> <loc>
<sequence of spt> ::= <spt>|<sequence of spt>
<clip> ::= < close-view camera >"["<obj>"]" | C "["< sequence of spt > "]" |D"[""]"
<sequence of clips> :: =<clip>| <sequence of clips>
<video> :: = <sequence of clips>

2.2.2 Modeling tennis videos

Consider the following video sequence in Figure 2. We can represent this video sequence using the grammar for tennis videos as:

T_1={A [U] C [U7b7V10 b4 F_v10 b8]}

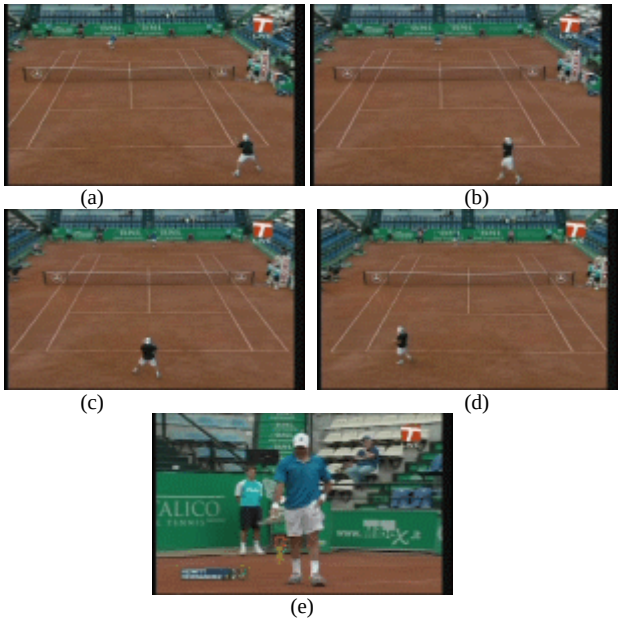

Figure 2. Different spatial locations and camera views

The above sequence explains that the camera A captures the close view of the player who is serving and then court view captures that first player and ball is at location 7 and player 2 is at location 10.The first player serves and ball goes in location 4 and then the player from location 10 gives a forehand shot and ball goes out of the court at location 8. The following is another sequence from a tennis video:

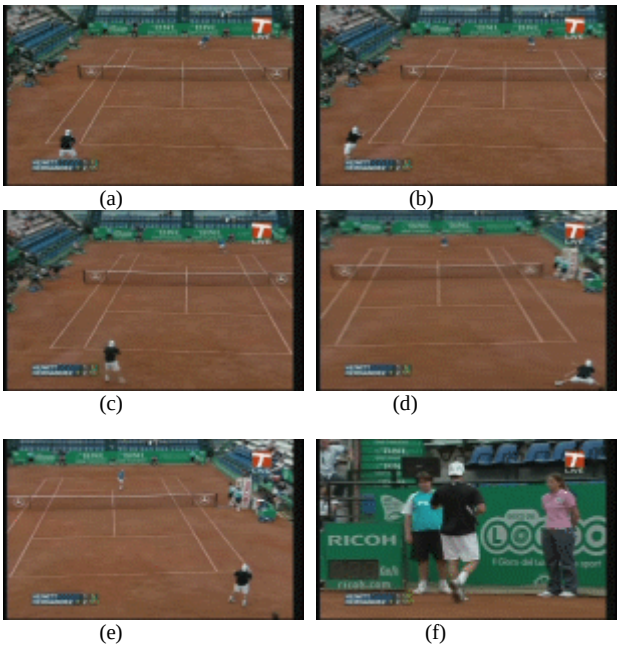

Figure 3. Sample sequences from a tennis video

T_2={A [U] C [U8 b8 V9 b3 Bv9 b5 B_U8 b4 F_V10 b5] D[ ]}

3. Spatiotemporal queries and examples

Representing the video content as a string helps the user describe many spatiotemporal queries. Most of the queries can be expressed by using SQL and few complex queries can be written by extending SQL through macro facility provided by *SMART*.
The following types of queries are allowed in SMART.

1) Event-Object-Location
2) Object-Location
3) Event- Location
4) Event-Object

We explain each of these with the examples on Tennis videos as follows:

3.1 Event-object-location

These queries retrieve the clips by specifying an event and location of an object in a video. Event-Object-Location describes the action and spatial information for the specified object. For example, a player takes a free-kick near penalty area in a soccer game. In this example 'player' is an object, 'free-kick' is the event and 'near penalty area' is the location. Retrieval of such a clip is called event-object-location query. These queries can be written in the following way:

SELECT clip
FROM *database*
WHERE *sequence like '%event object location %'*

The above query does the string matching and looks for the pattern of "event object location" anywhere in the sequence.

Example 1. List all the clips where the player who serves gives a backhand shot from area outside the court.

SELECT clip
FROM MainTable
WHERE sequence like '%BU7%' or sequence like '%BU8%';

This query finds out the pattern of 'BU7' and 'BU8' anywhere in the string.'BU7' represents the backhand shot by player U at location 7. Similarly, 'BU8' represents the backhand shot by player U at location 8.

3.2 Object-location

These queries retrieve the clips by specifying the location and object of the video. Object-Location describes the spatial information for the given object. For example, in soccer game defender stands near the goal post. Here, the 'defender' is the object and 'near goal post' is location .Retrieval of such a clip is called object-location query. These queries can be written in the following way:

SELECT clip
FROM *database*
WHERE *sequence like '%object location %'*

The above query does the string matching and looks for the pattern of "object location" anywhere in the sequence.

Example 2. List all the clips where the ball goes outside the court

SELECT clip
FROM MainTable
WHERE sequence like '%b9%' or sequence like '%b10%' or sequence like '%b7%' or sequence like '%b8%';

This query finds out the pattern of 'b9', 'b10', 'b7' and 'b8' anywhere in the string. 'b9' represents the ball and location 9. Similarly, 'b10', 'b7' and 'b8' represent the ball at location 10,7 and 8.

3.3 Event-location

These queries are formed by specifying the event and the location of the occurrence of event in the video. Event-Location describes the spatial information of the action. For example, in soccer game throw-in from outside the field. Here, 'throw-in' is the event and 'outside the field' is the location for the event. Retrieval of clips on this criterion is called event-location query.
These queries can be written in the following way:

SELECT clip
FROM *database*
WHERE *sequence like '%event_ location %'*

The above query does the string matching and looks for the pattern of 'event_location' anywhere in the sequence. Here, '_' can take any single value (character).

Example 3. List all the clips of forward shots from area inside the court.

SELECT clip
FROM MainTable

WHERE sequence like '%F_1%' or sequence like '%F_2%' or sequence like '%F_3%' or sequence like '%F_4%';

This query finds out the pattern of 'F one character 1', 'F one character 2', 'F one character 3' and 'F one character 4' anywhere in the string. 'F_1' represents the forehand shot by any player at location 1. Similarly, 'F_2' represents the forehand shot by any player at location 2.

3.4 Event-object

These queries retrieve the clips by specifying the event and object of the video. For example, in soccer game saving the goal by goal keeper, 'goal keeper' is the object and 'saving the goal' is event. Retrieval of such a clip is called event-object query. These queries can be written in the following way:

SELECT clip
FROM *database*
WHERE *sequence like '%event object %'*

The above query does the string matching and looks for the pattern of 'event object' anywhere in the sequence

Example 4. List all the clips where the player U gives a backhand shot
SELECT clip
FROM MainTable
WHERE sequence like '%BU%' ;

This query finds out the pattern of 'BU' anywhere in the string. B represents the backhand shot and U represents player who serves.

4. Conclusion

In this paper, we have provided a method of video modeling and representation. SMART defines a grammar for string based linear representation of the semantic contents of the video. Experiments are carried out on tennis videos and we have got promising results. SMART also provides various spatiotemporal queries. As future work, we plan to test SMART on other sports videos such as soccer and develop indexing strategies for fast retrieval of data.

References

[1] Xingquan Zhu, Xindong Wu, Ahmed K. Elmagarmid, Zhe Feng and Lide W, "Video Data Mining: Semantic Indexing and Event Detection from Association Perspective", *IEEE Transaction on Knowledge and data engineering*, Vol. 17, No 5,P.P. 665-677, 2005

[2] Martin Erwig & Markus Schneider,"Developments in Saptio- Temporal Query Languages", *Database and Expert systems Applications,1999. Proceedings. Tenth International workshop*,P.P.441-449,1999

[3] Alan Hanjalic, Li-Qun Xu, "Affective Video Content Representation and Modeling", *IEEE Transactions on Multimedia*, Vol. 7, No 1,P.P. 143-154, 2005

[4] Ramazan Savas Aygun and Adnan Yazici, "Modeling and management of fuzzy Information in Multimedia Database Application", *Multimedia Tools and Applications*, Vol. 24,No 1, P.P. 29-56 ,2004

[5] YouTube, http://www.youtube.com

[6] Google Video, http://video.google.com

[7] John Z.Li, m.Tamer Ozsu and Duane Szafron, "Modeling of moving objects in a video database", *1997 International Conference on Multimedia Computing and Systems*, P.P. 336**,** 1997

[8] Chang S. , Shi Q., Yan C., "Iconic indexing by 2-D strings", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9,No. 3, P.P. 413-428 ,1987

[9] Niki Pissinou, Ivan Radev, Kia Makki and William J. Campbell, "Spatio-Temporal Composition of Video Objects: Representation and Querying in Video Database Systems", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, No 16, P.P. 1033-1040, 2001.

[10] Shermann S.M. Chan, Qing Li, Yi Wu and Yueting Zhuang, "Accommodating Hybrid Retrieval in a Comprehensive Video Database Management System", *IEEE Transactions on Multimedia*, Vol. 4, No 2, P.P.,146-159, June 2002.

[11] Neil Robertson, I an Reid,"A general method for human activity recognition in video",*Computer Vision and Image Understanding*, Vol. 104 ,No. 2, P.P.232-248, 2006

[12] F.Yan, W.Christmas and J.Kittler," A Tennis Ball Tracking Algorithm for Automatic Annotation of Tennis Match", *The British Machine Vision Association*, 2005

[13] Guangyu Zhu, Changsheng Xu, Qingming Huang, Wen Gao and Liyuan Xing," Player Action Recognition in Broadcast Tennis Video with Applications to Semantic Analysis of Sports Game", International Multimedia Conference Proceedings of the 14th annual ACM international conference on Multimedia, P.P. 431-440,2006