Volume 2, Issue 3, March 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Video Content Representation with Grammar for Semantic Retrieval

M.Ravinder*
Associate Professor in CSE
Sree Chaitanya College of Engineering
Karimnagar, India.
email: mravinder_cse@yahoo.com

Dr.T.Venu Gopal
Associate Professor in CSE
JNTUCEH, Jagtial,
Karimnagar, India.

Abstract— **Video database modeling and representation is important for well-organized storage and retrieval of video. Modeling of semantic video content so as to enable spatiotemporal queries is one of the challenging tasks. In this paper, 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 our model.**

Keywords— **Semantic retrieval, video modeling, indexing, spatial queries, temporal queries.**

I. INTRODUCTION

This 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 [1], Google videos [2] 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 [3], 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 [4], 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 [5], the fuzzy information i.e. imprecise information is being modeled and managed in multimedia databases. A video management and application processing framework is been proposed in [6]. The main component of this framework is query based video retrieval mechanism that allows querying in its language, CAROL/ST [6]. 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 [8], 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. There has been a significant research in the extraction of low level features of sports videos and mapping it to high level concepts [9] [10] [11].

The grammar-based representation simplifies maintaining the semantics of a video especially if the video content follows some rules as in sports games. These types of videos have three types of contents: objects, events and locations. The grammar helps us determine the connections between these components. The grammar can even further help us embed the rules of a game as part of the database. Since each game has a different set of rules, this requires differences especially when creating queries or applying semantic data reduction. In terms of modeling and representation, different types of videos can be represented with different grammars. For a new type of game, our system just requires a new grammar and a set of rules based on this grammar. A visual query language [12] is developed to express SQL queries that include joins with no emphasis on spatio-temporal content. BilVideo [13] extends SQL for spatio-temporal queries. Nevertheless, SQL-like query languages or extensions to SQL complicate querying [14]. The visualization of many video query results at the same time is actually cumbersome, if the video clips are not short. The view of many clips also requires the streaming of multiple clips from the video database server.

In this paper we propose a grammar based method for representing the content of a news video, we discuss about related work in section II, section III talks about modeling of video data, in section IV we focus on modelling and

representation of cricket video, we end with conclusion and future work in section V.

II. RELATED WORK

In this section, we provide an overview of approaches on video representation, modeling and retrieval. These approaches mainly vary upon the aspect of video to be used for modeling.

A. Semantic content

A video can be represented in terms of alphabets. The sequence of alphabets represents the features of video. In [15], the video data is transformed and represented as a stream of alphabets. A video is represented in terms of four streams: court field, camera motion, scoreboard and audio events. A mapping table has been suggested where the symbols are mapped to video data. For example, court stream can be court or non-court. Here, symbol 'A' can represent court and symbol 'B' can represent non-court. A clip can be represented with the help of these symbols such as "B, A, D, G……..". This representation provides an alphabetical representation to the video clip. However, there is no representation for the main objects of the video such as the players or the ball in sports video. Representation of the events like 'kicking ball' and 'defending goal' is missing. They have also not provided the queries that can help in retrieval of information, since they target video data mining.

B. Affective perception of video

In [16], the videos are divided with respect to video content perception as affective and cognitive. Cognitive perception represents the facts that are present in the videos such as news and events, whereas affective perception represents the effects the video creates on the viewers such as thrilling scenes. A graph-based model has been proposed for the modeling of affective video content. A 2D emotion space with arousal and valence as dimensions is defined. Valence is the intensity of emotion; and arousal is the type of emotion. The low level features of the video that are extracted by processing tools are mapped to 2D emotion space.

C. Modeling fuzzy information

Fuzzy information basically represents the uncertain and imprecise values where exact values are not available. In [17], a conceptual model called ExIFO2 is used to model the fuzzy information. In [18], ExIFO2 is utilized for multimedia database applications. This conceptual model is then mapped to logical object-oriented model called fuzzy object oriented data (FOOD). Various mapping algorithms are proposed to map from conceptual model to logical model. Experiments were carried out on soccer videos. The querying interface also permits fuzzy values in setting up the queries.

D. Spatiotemporal data modeling

There has been a lot of research in the area of modeling spatial and temporal contents of video. In [19], the trajectory of a 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. A trajectory matching algorithm is provided for matching the moving objects. The model has been integrated with Object Oriented Database Management System (OODMBS) and uses Object Query Language (OQL). Pissinou et al. propose a Topological-Directional Model for the spatiotemporal contents of the video [20]. The model represents an object by a minimum bounding rectangular parallelepiped (mbrp). Mbrp is used to enclose an object. A key frame is described with the help of relative positions of objects. The relative positions are specified using the relationships between the projections of mbrp 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 do not provide a very efficient retrieval method. In [21], SQL is extended to describe various spatiotemporal queries. However, the representation of data is not very efficient and the queries where the events are one after the other are also missing. In [22], the authors suggest a model called Content Based Retrieval (COBRA) for mapping low level features to high level features. This model supports stochastic techniques such as Hidden Markov Models (HMMs) for mapping purposes. They provide an object grammar, an event grammar and algebraic operators. However, their grammar defines the components and domain of objects and events rather than how they should be represented in the database. There is also no hierarchical representation of video content.

III. MODELING OF VIDEO DATA

Videos can be classified into two types with respect to its semantic content [23]. The first type covers videos having content organization 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 by means of query language such as SQL. In our model, we classify 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:

A. Objects

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

B. Locations

Spatial information is represented by 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 rules of the sports. Each region L can be represented from the location domain $\Sigma_L = \{L_1, L_2, L_3 \ldots Lp\}$.

C. Cameras

In every video, various cameras (or camera views) give different footage. For example, inside 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 $\Sigma_C = \{C_1, C_2, C_3 \ldots Cq\}$.

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

D. Grammar for spatiotemporal representation

As we represent spatiotemporal content of a video as a string, the grammar is necessary to parse and extract the spatiotemporal information from the string. We can describe the grammar used 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 series of spatiotemporal instances; and a spatiotemporal instance (spt) is represented by means of an object (obj), location (loc), and an optional event [24].

IV. MODELING AND REPRESENTATION OF CRICKET VIDEO

We describe the application of the above approach on the cricket videos. The representation, modeling and data reduction of cricket video are described in the following subsections.

A. Architecture used in our model

Fig. 1 provides an architecture that maintains and stores the videos, their representation and summaries. It besides facilitates the querying of these videos with the help of user interface.

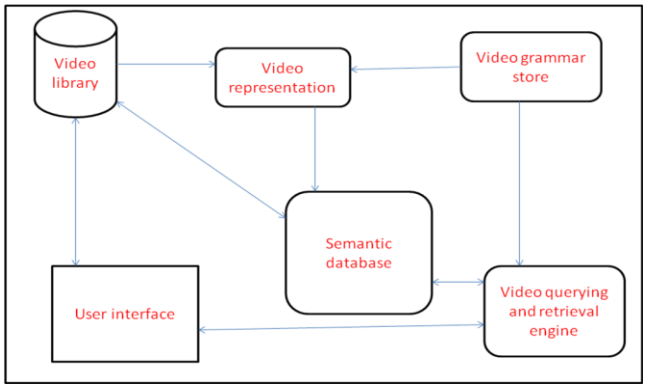

Fig. 1 Architecture used in our model

The major components of the architecture are: video library, video grammar store, video representation, Semantic database, video querying and retrieval engine, and user interface. By using the grammar store, the video content is converted to string representation, and these are maintained in the semantic database. The video querying engine responds to user queries using the grammar from the repository and applying pattern matching on the semantic database. The video summaries are primary presented to the user after querying. We now explain the components of this architecture.

- **Video library:** Video library maintains the collection of all videos. It stores the data in a raw format without any processing. Video library provides the unprocessed clips to video representation. The semantic database maintains the pointers to videos in the video library. The videos are retrieved from the library for displaying to the user.
- **Video grammar store:** The grammar repository maintains the syntax of data representation and is used to represent the contents of the video and then storing the semantic strings in the semantic database.
- **Video representation:** Video representation enables representation of semantic features of the sports videos as string data that are compliant with the proposed grammar. This allows spatiotemporal queries using query language, SQL.
- **Semantic database:** Semantic database stores all the semantic features of the video clips that are maintained by means of the video representation module.
- **Video querying and retrieval engine:** Video querying and retrieval module constructs the query from the information provided by the user interface, executes the query and retrieves the result. The query engine performs the pattern matching according to the existing grammar from the video grammar store.
- **User interface:** User interface accepts the queries and give them to video querying and retrieval engine. After the query is executed, the user interface displays the results of the query to the user.

B. Representation of cricket Video

We provide short information on cricket videos while explaining the representation of the cricket video content.

- **Objects**: There are 5 main objects identified in the cricket game. The objects are identified as $\Sigma_O = \{X_1, X_2, Y, x, y, z, K, F, b\}$ where X_1 is batsman1, X_2 is batsman2, Y is bowler, x, y, z are umpire1, umpire2, umpire3, F is the fielder and b is the ball.
- **Events:** The main events are identified in the alphabet $\Sigma_E=\{f_w,b_w,w,n,R,L,C,S,D,0,1,2,3,4,5,6,7\}$ where f_w is forward shot, b_w is backward shot w is wide ball, n is no ball, R is run out, L is LBW, C is catch out, S is stump out, D direct to wicket out and runs represented by 0, 1, 2, 3, 4, 5, 6, and 7.
- **Locations:** The cricket ground is divided into regions by line segments to apply the rules of cricket game as in Fig 2. For representation of locations and for semantic retrieval, we divide the ground into

partitions in the same way and apply the numbering in Figure 2.

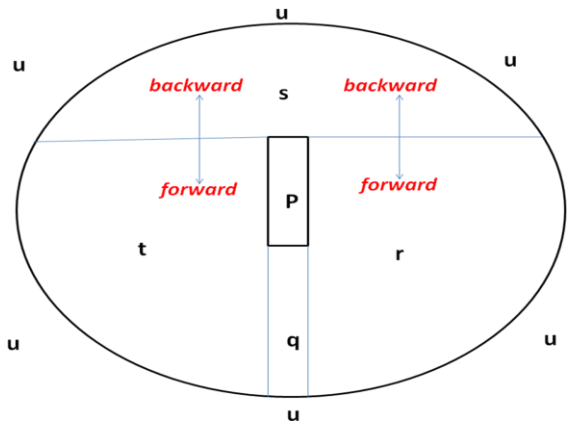

Fig. 2 Cricket ground segmentation

- **Cameras Views:** We identify six types of camera views in a cricket game:
  A – Gives a close view of the player at location p, q, and r in Fig 2
  B – Gives a close view of the player at location s, t, and u in Fig 2
  G – Ground view
  P – Pitch view
  H - Action Replay
  R - Rest time
  Com – Commentators
- **Grammar for Cricket Video Database:** We now extend the grammar for cricket videos.
  <obj>:: = <player>|b|<umpire>
  <umpire>::=x|y|z
  <player>:: =X_1|X_2|Y|K|F
  <event>:: = <shot><runs>|<out>|<extra><runs>
  <shot>::=f_w|b_w
  <extra>::=w|n
  <runs>::=0|1|2|3|4|5|6|7
  <out>::=R|L|C|S|D
  <location>:: = p|q|r|s|t|u
  <camera>:: = <close view camera>|G|H
  <close view camera>::=A|B
  <spt>:: = <obj><loc>|<event><player> <loc>|<event>
  <spatiotemporal>::= <spt> | <spatiotemporal>
  <sequence of clips>:: =<clip>|<sequence of clips>
  <clip>::= < close-view camera > "[" <player> "]"
  | G"[" <spatiotemporal> "]" |H"[""]"
  <video> :: = <sequence of clips>

Where different main objects involved in cricket game are player, ball (b), umpire. As in cricket three umpires are involved we have represented first umpire with x, second umpire with y, and third umpire with z. Players involved in cricket can be classified and represented as batsman1(X_1), batsman2(X_2), bowler(Y), keeper(K), fielder(F).

Different events that are possible in cricket are shot followed by runs, out, and extra followed by runs. Runs can be represented with 0, 1, 2, 3, 4, 5, 6, and 7 depending on number of runs taken by batsman and given by bowler. The variable shot can be defined as forward (f_w) shot and backward (b_w) shot. The variable extra can be defined as wide ball (W), no ball (N). The variable out can be represented using R (run out), L (LBW), C (catch out), S (stump out), and D (direct to wickets). We have segmented the cricket ground in to six main parts. They are p(pitch location), q, r, s, t, and u.

C. Modeling: We now apply the above grammar for the following video clip. Consider the video sequences in Fig 3. We can represent this video sequence using the grammar for cricket videos as:

T_1 = {A[X_1]P[YqbqX$_1$pbpX$_1$b$_w$1bs]A[X_1]}

The above sequence indicates that camera A captures the close view of batsman X_1 then Pitch view captures the sequence bowler Y at location q and ball at location q, batsman X_1 at location p, followed by ball at location p, followed by the event backward (b_w) shot followed by runs 1, followed by ball in location s, followed by close view of batsman X_1

Fig. (a): close view of batsman X_1
Fig. (b): Pitch view and bowler at location q and batsman at location p.
Fig. (c): Pitch view and bowler moves from location q to location p.
Fig. (d): Pitch view ball moves from location q to location p
Fig. (e): ball in location p
Fig. (f): backward (b_w) shot by X_1
Fig. (g): ball moves from location p to s
Fig. (h): event runs 1
Fig. (i): close view of batsman X_1

V. VIDEO RETRIEVAL USING QUERY

Our architecture supports different types of queries. These queries can be written using user interface. The user interface abstracts the lower level representation of videos and allows the user to write queries. Representing the video content as a string helps the user describe many spatiotemporal queries. Most of the queries can be expressed by using SQL. The following types of queries are allowed.

1. Event-object-location
2. Event
3. Object-location
4. Event- location
5. Event-object
6. Current and next event

A. Event-object-location:

Event-object-location 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. These queries can be written in the following way:

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

For example we can write the query for our cricket database as follows:

SELECT clip
FROM cricketdatabase
WHERE sequence like '%f_wX_1p%' or sequence like '%b_wX_1p%'

B. Event:

Event queries retrieve the clips by specifying an event. These queries can be written in the following way:

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

For example we can write the query for our cricket database as follows:

SELECT clip
FROM cricketdatabase
WHERE sequence like '%R%' or sequence like '%L%' or sequence like '%C%' or sequence like '%S%' or sequence like '%D%'

C. Object-location:

Object-location queries retrieve the clips by specifying the location and object of the video. Object-location describes the spatial information for the given object. These queries can be written in the following way:

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

For example we can write the query for our cricket database as follows:

SELECT clip
FROM cricketdatabase
WHERE sequence like '%bp%' or sequence like '%bq%' or sequence like '%br%' or sequence like '%bs%' or sequence like '%bt%' or sequence like '%bu%'

D. Event-Location:

Event-location 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. These queries can be written in the following way:

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

For example we can write the query for our cricket database as follows:

SELECT clip
FROM cricketdatabase
WHERE sequence like '%Cp%' or sequence like '%Cq%' or sequence like '%Cr%' or sequence like '%Cs%' or sequence like '%Ct%'

E. Event-Object:

Event-object queries retrieve the clips by specifying the event and object of the video. These queries can be written in the following way:

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

For example we can write the query for our cricket database as follows:

SELECT clip
FROM cricketdatabase
WHERE sequence like '%f_wX_1%' or sequence like '%b_wX_1%'

F. Current and next event:

Current-next event queries are formed by specifying the events that happen one after the other. Current- next event describes the occurrence of an action followed by another occurrence of action anywhere on the time axis. These queries can be written in the following two ways:

1) SELECT clip
FROM database
WHERE sequence like '%event_ _object location %'
2) SELECT clip
FROM database
WHERE sequence like '%event_ _ event %'

The first query checks cases where an event leads to a specific spatial instance whereas the second query explicitly represents back-to-back events. Two '_' characters are used, since it is not important who performs the event and what the location is. Here, event represents the current event and next event can be either represented by an object or an event itself.

For example we can write the query for our cricket database as follows:

SELECT clip
FROM cricketdatabase
WHERE sequence like '%f_w__X_1p%' or sequence like '%b_w_ _X_1p%'

SELECT clip
FROM database
WHERE sequence like '%f_w1_ _ R %'

VI. CONCLUSION AND FUTURE WORK

In this paper, we have provided a technique of video modeling, representation, and retrieval based on a grammar.. The strongest points of our architecture is the power of making spatiotemporal queries simple by using SQL (without extending SQL) and allowing multimodal presentation by returning images and videos. Actually, this originates from the powerful representation of semantic video content as a linear string. As future work, we plan to develop indexing strategies for fast retrieval of data.

REFERENCES

[1] YouTube, http://www.youtube.com
[2] Google Video, http://video.google.com.
[3] 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.
[4] Alan Hanjalic, Li-Qun Xu,"Affective Video Content Representation and Modeling", IEEE Transactions on Multimedia, Vol. 7, No 1,P.P. 143-154, 2005.
[5] 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.
[6] 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.
[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] 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.
[9] 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 .
[10] F.Yan, W.Christmas and J.Kittler," A Tennis Ball Tracking Algorithm for Automatic Annotation of Tennis Match", The British Machine Vision Association, 2005.
[11] 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 .
[12] D.A. Keim and V. Lum. "Visual query specification in a multimedia database system". In Proceedings of the 3rd Conference on Visualization '92 (Boston, Massachusetts, October 19 - 23, 1992). A. Kaufman and G. Nielson, Eds. IEEE Visualization. pp. 194-201, 1992.
[13] O. Kucuktunc, U. Gudukbay and O. Ulusoy, "A natural language-based interface for querying a video database," IEEE MultiMedia, vol. 14, no. 1, pp. 83-89, January-March, 2007.
[14] G. Erozel, N. K. Cicekli and I. Cicekli. "Natural language querying for video databases". Volume 178, Issue 12, pp. 2534-2552. June 2008.
[15] X. Zhu, X. Wu, A. K. Elmagarmid, Z. Feng and L. W, "Video data mining: semantic indexing and event detection from association perspective", IEEE Transactions on Knowledge and Data engineering, Vol. 17, No 5, pp. 665- 677, 2005.
[16] A. Hanjalic and L-Q. Xu, "Affective video content representation and modeling", IEEE Transactions on Multimedia, Vol. 7, No 1, pp. 143-154, 2005.
[17] A. Yazici and A. Cinar,"Conceptual modeling for the design of fuzzy OO databases", in Knowledge Management in Fuzzy Databases, O. Pons, A. Vila, and A. Vila and J Kackrzyk (Eds), Physica-Verlag:Heidlberg, New York, Vol 39, 2000, pp.12-35.
[18] R. S. Aygun and A. Yazici, "Modeling and management of fuzzy information in multimedia database application", Multimedia Tools and Applications, Vol. 24,No 1, pp. 29-56, 2004
[19] J.Z. Li, M.T. Ozsu and D. Szafron, "Modeling of moving objects in a video database", 1997 International Conference on Multimedia Computing and Systems, pp. 336, 1997.
[20] N. Pissinou, I. Radev, K. Makki and W. 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, pp. 1033-1040, 2001.
[21] M. Erwig and M. Schneider, "Developments in spatio- temporal query languages", Database and Expert systems Applications, 1999. Proceedings. Tenth International workshop, pp. 441-449, 1999 .
[22] M. Petkovic and W. Jonker, "Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events", Detection and Recognition of Events in Video, 2001, Proceedings, IEEE Workshop on Volume, Issue, pp. 75 – 82, 2001.
[23] H. Agius and M. C. Angelides,"MPEG-7 In Action: End user experiences with Cosmos-7 front end systems", Proceedings of the 2006 ACM Symposium on Applied Computing SAC '06, pp. 1348-1355,2006.
[24] Vani Jain and Ramazan Aygun, "Spatio-Temporal Querying of Video Content Using SQL for Quantizable Video Databases" JOURNAL OF MULTIMEDIA, VOL. 4, NO. 4, AUGUST 2009.

