A PROBABILISTIC-BASED MECHANISM FOR VIDEO DATABASE MANAGEMENT SYSTEMS

Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124

Shu-Ching Chen
School of Computer Science
Florida International University
Miami, FL 33199

R. L. Kashyap[*]
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907

ABSTRACT

As more information sources become available in multimedia systems, the development of *multimedia database management systems (MDBMSs)* to efficiently model and search multimedia data, especially video data, becomes very crucial for multimedia applications. In response to such a demand, a probabilistic-based mechanism called the *Markov Model Mediator (MMM)* to facilitate an *MDBMS* for video database systems is presented. In our previous studies, the spatial relations of the semantic objects in the image/video data modeled by the *MMM* mechanism are assumed given by image processing/computer vision techniques or by human annotations. In this paper, an unsupervised video segmentation method that can identify objects with their corresponding spatial relations automatically is incorporated into the *MMM* mechanism. Based on the information obtained, users can retrieve video materials from video databases via database queries. Hence, both multimedia data modeling and image processing capabilities are integrated into the *MMM* mechanism.

1. INTRODUCTION

Multimedia applications require the development of *multimedia database management systems (MDBMSs)* to support the efficient organization, storage and retrieval of multimedia data, especially for the video data. For images and video frames, the *MDBMS* needs to keep the relative spatial positions of semantic objects (building, car, etc.) so that users can issue queries.

Many data models have been proposed for video data [1, 4, 7, 9]. However, most of them focus on either data modeling or browsing/retrieval. For example, in our previous studies [9], we have demonstrated that a probabilistic-based *Markov Model Mediator (MMM)* mechanism has capabilities for both video data modeling and information retrieval, but the spatial relations of the semantic objects in the image/video data are assumed given by image processing/computer vision techniques or by human annotations. In order to meet the needs for a variety of video applications, particularly with respect to semantic video data modeling, searching, and retrieval, a video database management system which incorporates both the multimedia data modeling and the image processing techniques is more than desirable. Toward this end, a video segmentation method – *simultaneous partition and class parameter estimation (SPCPE)* algorithm [10] – is incorporated into the *MMM* mechanism to automatically identify the spatial relations of the semantic objects. Hence, both data modeling and image processing capabilities are integrated into the *MMM* mechanism to facilitate the functionality of a video database management system.

A media object such as a video clip, an image, a text file, or a complex entity of these simpler entities is represented as a node in an MMM and is associated with an *augmented transition network (ATN)*. An ATN is a model for multimedia presentations, multimedia database searching, and multimedia browsing [3]. Multimedia input strings are the inputs for ATNs. The basic twenty-seven spatial relations introduced in [3] are used in the multimedia input strings to indicate the objects' spatial relations that are captured by the unsupervised video segmentation method [10]. We apply the video segmentation method to a small portion of a soccer game video and use the information obtained to illustrate how the *MMM* mechanism facilitates the functionality of a video database management system. Under our design, the spatial relations of the semantic objects in the video are captured and modeled in the proposed mechanism, which allows users to retrieve video materials by database queries.

The organization of this paper is as follows. Section 2 introduces the proposed mechanism. How information retrieval can be performed via a stochastic process along with an example soccer game video is presented in Section 3. Section 4 concludes this paper.

[*] This work has been partially supported by National Science Foundation under contract IRI 9619812.

2. THE INTEGRATED PROBABILISTIC-BASED MECHANISM

A probabilistic-based mechanism called *Markov Model Mediators (MMMs)* which integrates both data modeling and image processing capabilities is proposed in this paper. The *MMM* mechanism adopts the *Markov Model* framework and the *mediator* concept. A Markov model is a well-researched mathematical construct which consists of a number of states connected by transitions; while a mediator is defined as a program that collects information from one or more sources, processes and combines it, and exports the resulting information [11]. Many applications use Markov models as a framework such as Hidden Markov Models (HMMs) [8] and Markov Random Field Models [5]. However, no existing research uses Markov models as a framework in designing a database management system.

Each MMM is represented by a 6-tuple λ=($\mathcal{S}$, $\mathcal{F}$, $\mathcal{A}$, $\mathcal{B}$,Π,Ψ) where $\mathcal{S}$ is a set of media objects called states; $\mathcal{F}$ is a set of attributes/features; $\mathcal{A}$ is the state transition probability distribution; $\mathcal{B}$ is the observation symbol probability distribution; Π is the initial state probability distribution; and Ψ is a set of augmented transition networks (ATNs). The elements in $\mathcal{S}$ and $\mathcal{F}$ determine the dimensions of $\mathcal{A}$ and $\mathcal{B}$. The structure of the member media objects is modeled by the sequence of the MMM states connected by transitions. When an ATN consists of images, video, or texts, the corresponding subnetworks are constructed. Subnetworks are developed to allow the users to issue queries relative to the spatio-temporal relations of the video or image contents or to specify the criteria based on a keyword or a combination of keywords. The inputs for ATNs and subnetworks are modeled by multimedia input strings that are used to represent the spatio-temporal relations of semantic objects and keyword compositions.

In our previous studies, the spatial relations of the semantic objects in the video frames are given as *a priori* in the *MMM* mechanism. In this paper, we have incorporated an unsupervised video segmentation method that captures the spatial relations into the *MMM* mechanism. The method for partitioning a video frame starts with an arbitrary partition and employs an iterative algorithm to estimate the partition and the class description parameters jointly. The key idea is to use the *SPCPE* algorithm successively on each video frame and incorporate the partition information of the previous frame as the initial condition while partitioning the current frame. With appropriate assumptions, the joint estimation can be simplified to the following form:

$$J(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{y_{ij} \in \boldsymbol{c}_1} -\ln p_1(y_{ij}; \boldsymbol{\theta}_1) + \sum_{y_{ij} \in \boldsymbol{c}_2} -\ln p_2(y_{ij}; \boldsymbol{\theta}_2).$$

where y_{ij} is a pixel of the image in each frame, $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are the partition variables, and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the parameters. Please see [10] for the details of this method.

The minimal bounding rectangle (MBR) concept in R-tree [6] is adopted so that each semantic object is covered by a rectangle. One semantic object is selected as the target semantic object in each video frame. The centroid point of each semantic object is used for space reasoning so that any semantic object is mapped to a point object. Therefore, the relative position between the target semantic object and a semantic object can be derived from these centroid points. Twenty-seven numbers representing three dimensional relative positions for semantic objects [3] are used to distinguish the relative positions of each semantic object relative to the target semantic object and are represented by subscripted numbers in the multimedia input strings.

3. STOCHASTIC PROCESS FOR INFORMATION RETRIEVAL

The need for efficient information retrieval for video database management systems is strong because searching databases one by one is very time-consuming and expensive. The cost for query processing usually is very high and the complexity of a query depends heavily on the order in which the databases are searched for a successful path. With the help of probabilistic models, information retrieval can be performed by a stochastic process that predicts the most likely path (state sequence) for a specific query. A dynamic programming algorithm which conducts the stochastic process is proposed to provide a systematic way to compute the edge weights and the cumulative edge weights for path ranking.

3.1. The Stochastic Process

Define $W_t(i,j)$ and $D_t(i,j)$ to be the edge weight and the cumulative edge weight of the edge $S_i \rightarrow S_j$ at time t, where $1 \leq i,j \leq N$, $1 \leq t \leq T-1$.

$$\begin{aligned}
W_1(i,j) &= \begin{cases} \pi_{S_i} b_{S_i}(o_1) & \text{i=j} \\ 0 & \text{otherwise} \end{cases} \\
D_1(i,j) &= W_1(i,j) \\
W_{t+1}(i,j) &= \max_k (D_t(k,i) a_{S_i,S_j}) b_{S_j}(o_{t+1}) \quad (1) \\
D_{t+1}(i,j) &= \max_k (D_t(k,i) + W_{t+1}(i,j)) \quad (2)
\end{aligned}$$

$\mathcal{A}$=$\{a_{S_i,S_j}\}$ is the state transition probability distribution for the MMM, where a_{S_i,S_j}=Pr(S_j at t+1 | S_i at t).
$\mathcal{B}$=$\{b_{S_j}(o_k)\}$ is the observation symbol probability distribution for the MMM, where $b_{S_j}(o_k)$=Pr(o_k at t | S_j at t).
$\Pi = \{\pi_{S_i}\}$ is the initial state probability distribution, where π_{S_i}=Pr(S_i at t=1).

The steps for the stochastic process are:

1. At time t=1, $W_1(i,j)$ is assigned the value of the joint probability of the state S_i with probability π_{S_i} and the attribute or feature o_1 with probability $b_{S_i}(o_1)$ when

$i = j$; $W_1(i,j) = 0$ if $i \neq j$. The cumulative edge weight $D_1(i,j)$ equals $W_1(i,j)$.

2. As time goes from t=1 to t=2, a transition goes from state S_i to state S_j with the probability $a_{S_i,Sj}$ and the attribute/feature o_2 is generated with probability $b_{S_j}(o_2)$.

3. Since $D_t(i,j)$ indicates the cumulative edge weight for the joint event that $o_1 \ldots o_t$ are observed and the state stops at S_i at time t, the product $D_t(i,j)a_{S_i,S_j}$ represents the joint event that $o_1 \ldots o_t$ are observed and the state S_j is reached at time $t+1$ via state S_i at time t.

4. Finding the maximal cumulative edge weight $D_t(i,j)$ over all the cumulative edge weights $D_t(k,j)$ (where $1 \leq k \leq N$) results in the most likely edge from S_i at time t to S_j at time $t+1$ with all the accompanying previous partial observations.

5. $W_{t+1}(i,j)$ is obtained by augmenting multiplicatively the maximum quantity of $D_t(i,j)a_{S_i,S_j}$ with $b_{S_j}(o_{t+1})$ and $D_{t+1}(i,j)$ is the addition of current edge weight $W_{t+1}(i,j)$ and the maximal $D_t(k,j)$ at time t.

6. At each time slot, $\sum_i D_t(i,j)$ which sums up all the incoming cumulative edge weights is calculated for each node S_j.

7. Sort the $\sum_i D_t(i,j)$ values for all the nodes at each time slot and the list of possible state sequences is ranked by the values of $\sum_i D_t(i,j)$ and to suggest the paths to retrieve information for the query.

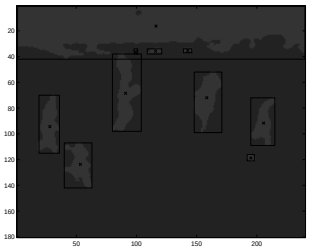
8. All the paths with positive cumulative edge weights are ranked in the following manner. The top ranked path is the one with maximal $\sum_i D_T(i,j_T)$, maximal $\sum_i D_{T-1}(i,j_{T-1})$, $\cdots$, and maximal $\sum_i D_1(i,j_1)$ and the state sequence is $j_1 \rightarrow \cdots \rightarrow j_{T-1} \rightarrow j_T$. The second ranked path could be the one with maximal $\sum_i D_T(i,j_T)$, maximal $\sum_i D_{T-1}(i,j_{T-1})$, $\cdots$, and second ranked $\sum_i D_1(i,j_1)$, if it exists. The same ranking process is executed to rank all the paths with positive cumulative edge weights.

Under the ranking process, the top-ranked path indicates a state sequence that provides the information for the query with maximal cumulative edge weight. If the top-ranked path cannot provide the information required for the query, then the second ranked path is considered. This is repeated until the information needed by the query can be obtained. From our experience, most of the edges have zero edge weights at each time slot. Hence, only the node S_j with positive $\sum_i D_t(i,j)$ at time t is involved in the computation of the edge weights and the cumulative edge weights for the next time slot.

3.2. Information Retrieval

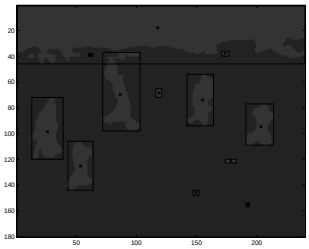
After the required subset of media objects is obtained, information retrieval is performed by traversing the ATNs associated with those identified media objects. If a media object contains images, video frames, or texts, then its subnetworks are traversed. The input for an ATN or a subnetwork is a multimedia input string which is constructed to model the spatio-temporal relations of semantic objects and keyword compositions. To make information retrieval of the *MMM* mechanism systematic and automatic, the *MMM* mechanism extends its data modeling functionality by incorporating a video segmentation method that identifies the semantic objects' spatial relations in the video. The temporal relations are captured by the sequence of the video frames.

(a) Frame 1



(b) Partition of Frame 1

(c) Frame 6



(d) Partition of Frame 6

Figure 1: On the left are the original frames; while on the right are shown their corresponding segments. The centroid of each segment is marked with an 'x' and the segment is shown with a bounding box around it.

Table 1: Part of the three dimensional relative positions for semantic objects.

No.	Relative Coordinates	No.	Relative Coordinates
1	$x_s \approx x_t, y_s \approx y_t, z_s \approx z_t$	13	$x_s < x_t, y_s < y_t, z_s \approx z_t$
10	$x_s < x_t, y_s \approx y_t, z_s \approx z_t$	19	$x_s > x_t, y_s \approx y_t, z_s \approx z_t$

We have applied the video segmentation method to an example soccer video in [2] and parts of the segmentation results (as shown in Figure 1) are used in this paper. In Figure 1, the original frames are on the left and the corresponding segments are on the right. Since only the ball and the players are important from the content based retrieval perspective, we use **P** and **B** to represent "players" and "soccer ball" with **G** ("ground") being selected as the target semantic object. Under the method, the players and the ball are

combined into a single segment if they are close to each other. For example, the ball is clubbed into a single segment with two other players in Frame 1, and the ball is far away so that it becomes a segment in itself in Frame 6.

- The constructed multimedia input strings:
  - Frame 1: $G_1\&P_{10}\&P_{13}\&P_1\&P_1\&P_{19}$.
  - Frame 6: $G_1\&P_{10}\&P_{13}\&P_1\&B_1\&P_1\&P_{19}$.

In a multimedia input string, the "&" symbol between two semantic objects denotes that the semantic objects appear in the same frame and the subscripted numbers distinguish the relative positions of the semantic objects relative to **G**. Table 1 lists part of the three dimensional spatial relations where (x_t, y_t, z_t) and (x_s, y_s, z_s) represent the X-, Y-, and Z-coordinates of the target and any semantic object, respectively. The "$\approx$" symbol means the difference between two coordinates is within a threshold value. The appearance sequence of the semantic objects in a multimedia input string is based on the spatial locations of the semantic objects in the video frame from left to right and top to bottom. For example, in Frame 1, G_1 indicates that **G** is the target semantic object. P_{10} means the first **P** is on the left of **G**, etc. In Frame 6, the soccer ball **B** appears at the same subregion as **G** and the rest of the semantic objects remain at the same locations. Thus, the spatio-temporal relations of the semantic objects are captured and modeled by the *MMM* mechanism automatically and users can retrieve video materials via database queries.

To systematically retrieve information, each high level query is first translated into a multimedia input string. Since those most likely required media objects are identified by the stochastic process, only the ATNs associated with those identified media objects are traversed. If any ATN consists of images, video frames, or texts, then the corresponding subnetworks are also traversed. Therefore, information retrieval becomes the problem of substring matching between the multimedia input string of the query and the multimedia input strings for the ATNs and/or their subnetworks.

4. CONCLUSIONS

In this paper, a probabilistic-based mechanism called *Markov Model Mediator (MMM)* that integrates both the data modeling and image processing capabilities to facilitate the functionality of a video database management system is presented. The *MMM* mechanism provides a systematic and automatic means for information retrieval. It is systematic since the structure of the media objects of the video data is modeled by the state sequence of the *MMM* states and a stochastic process is developed to identify the most likely required media objects for a specific query. It is automatic since the required spatio-temporal relations of the semantic objects in the video data are captured by the proposed unsupervised video segmentation method and modeled by the multimedia input strings. Users can retrieve video materials by issuing database queries. Information retrieval is performed by conducting substring matching between the multimedia inputs of the ATNs and/or their subnetworks and the multimedia input string of a query.

5. REFERENCES

[1] F. Arman, R. Depommer, A. Hsu, and M.Y. Chiu, "Content-based browsing of video sequences," *ACM Multimedia 94*, pp. 97-103, Aug. 1994.

[2] S-C. Chen, S. Sista, M-L. Shyu, and R.L. Kashyap, "An Indexing and Searching Structure for Multimedia Database Systems," the IS&T/SPIE conference on Storage and Retrieval for Media Databases 2000, pp. 262-270, January 23-28, 2000, San Jose, CA, U.S.A.

[3] S-C. Chen and R.L. Kashyap, "A Spatio-Temporal Semantic Model for Multimedia Presentations and Multimedia Database Systems," accepted for publication *IEEE Transactions on Knowledge and Data Engineering*, 2000.

[4] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Computer*, Vol. 28, No. 9, pp. 23-31, September 1995.

[5] O. Frank and D. Strauss, "Markov graphs," *Journal of the American Statistical Association*, 81, pp. 832-842, 1986.

[6] A. Guttman, "R-tree: A Dynamic Index Structure for Spatial Search," in Proc. ACM SIGMOD, pp. 47-57, June 1984.

[7] Q. Li and L.S. Huang, "A dynamic data model for a video database management system," *ACM Computing Surveys*, vol. 27, no. 4, pp. 602-606, December 1995.

[8] L.R. Rabiner and B.H. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, 3(1), pp. 4-16, January 1986.

[9] M-L. Shyu, S-C. Chen, and R.L. Kashyap, "Information Retrieval Using Markov Model Mediators in Multimedia Database Systems," 1998 International Symposium on Multimedia Information Processing, pp. 237-242, Dec. 14-16, 1998.

[10] S. Sista and R.L. Kashyap, "Unsupervised video segmentation and object tracking," *IEEE Int'l Conf. on Image Processing*, Japan, 1999.

[11] G. Wiederhold, "Mediators in the architecture of future information systems," *IEEE Computers*, pp. 38-49, March 1992.