Augmented Transition Network as a Semantic Model for Video Data

Shu-Ching Chen* — **Mei-Ling Shyu**** — **R. L. Kashyap*****

* *Florida International University, School of Computer Science,*
Miami, FL 33199, USA

chens@cs.fiu.edu
** *University of Miami, Department of Electrical and Computer Engineering*
Coral Gables, FL 33124-0640, USA

shyu@miami.edu
*** *Purdue University, School of Electrical and Computer Engineering*
West Lafayette, IN 47907, USA

kashyap@ecn.purdue.edu

ABSTRACT. *An abstract semantic model called the augmented transition network (ATN), which can model video data and user interactions, is proposed in this paper. An ATN and its subnetworks can model video data based on different granularities such as scenes, shots and key frames. Multimedia input strings are used as inputs for ATNs. Key frame selection is based on temporal and spatial relations of semantic objects in each shot. The relations of semantic objects are captured from our proposed unsupervised video segmentation method which considers the problem of partitioning each frame as a joint estimation of the partition and class parameter variables. Unlike existing semantic models which only model multimedia presentation, multimedia database searching, or browsing, ATNs together with multimedia input strings can model these three in one framework.*

RÉSUMÉ. *Dans cet article est présenté un modèle sémantique théorique appelé ATN (augmented transition network) qui est capable de générer des données vidéo et de créer des interactions avec l'utilisateur. Un ATN et ses sous-reseaux peuvent développer des bases de données video basées sur différents objets tel que des scènes, des images ou des "key frames." Les chaines de données multimédia sont utilisées comme sources d'entrées par les ATN. La sélection des "key frames" est elle basée sur une relation temporel et spatiale des objets sémantiques. Cette même relation est aquise par notre méthode de segmentation vidéo autonome proposée, laquelle se charge de partitioner chaque image en créeant une estimation jointe des variables de paramètres de classes et de la partition en elle même. Contrairement aux modèles sémantiques existant qui se contentent de générer soit une présentation multimédia, soit une base de données de recherche multimedia, ou tout simplement un affichage (appelé également "browsing"), les ATN combinés à des donnée d'entrée multimedia peuvent générer ces trois derniers à la fois en un unique cadre de travaille.*

KEYWORDS: *Augmented Transition Network (ATN), Multimedia Input String*

MOTS-CLÉS : *Resean de transition augmenté (ATN), Données d'entrée multimedia*

1. Introduction

Recently, multimedia database systems have emerged as a fruitful area for research due to the recent progress in high-speed communication networks, large capacity storage devices, digitized media, and data compression technologies over the last few years. Multimedia information has been used in a variety of applications including manufacturing, education, medicine, entertainment, etc. Unlike the traditional database systems which have text or numerical data, a multimedia database or information system may contain different media such as text, image, audio, and video. The important characteristic of such a system is that all of the different media are brought together into one single unit, all controlled by a computer.

An increasing number of digital library systems allow users to access not only textual or pictorial documents, but also video data. Video is popular in many applications such as education and training, video conferencing, video on demand, news service, and so on. Digital library applications based on huge amount of digital video data must be able to satisfy complex semantic information needs and require efficient browsing and searching mechanisms to extract relevant information [HOL 98]. Traditionally, when users want to search for certain content in videos, they need to fast forward or rewind to get a quick overview of interest on the video tape. This is a sequential process and users do not have a chance to choose or jump to a specific topic directly. In most cases, users have to browse through parts of the video collection to get the information they want, which address the contents and the meaning of the video documents. Also, users should have the opportunity to retrieve video materials by using database queries. Since video data contains rich semantic information, database queries should allow users to get high level content such as *scenes* or *shots* and low level content according to the temporal and spatial relations of semantic objects. A semantic object is an object appearing in a video frame such as a “car.” How to organize video data and provide the visual content in compact forms becomes important in multimedia applications [YEO 97]. Hence, a semantic model should have the ability to model visual contents at different granularities so that users can quickly browse large video collections.

With the emerging demand on content based video processing approaches, more and more attention is devoted to segmenting video frames into regions such that each region, or a group of regions, corresponds to an object that is meaningful to human viewers [FER 97, COU 97]. This kind of object based representation of the video data is being incorporated into standards like MPEG4 and MPEG7 [FER 97]. A video clip is a temporal sequence of two dimensional samples of the visual field. Each sample is an image which is referred to as a frame of the video. Segmentation of an image, in its most general sense, is to divide it into smaller parts. In image segmentation, the input image is partitioned into regions such that each region satisfies some homogeneity criterion. The regions, which are usually characterized by homogeneity criteria like intensity values, texture, etc., are also referred to as *classes*. Video segmentation is a very important step in processing video clips. One of the emerging applications in video processing is its storage and retrieval from multimedia databases and content based indexing. Video data can be temporally segmented into smaller groups depending on the scene activity where each group contains several frames. Clips are divided into scenes and scenes into shots. A shot is considered the smallest group of frames that represent a semantically consistent unit.

Videos include verbal and visual information that is spatially, graphically, and temporally spread out. This makes indexing video data more complex than textual data. Typically, indexing covers only the topical or content-dependent characteristics. The extra-topical or content-independent characteristics of visual information are not indexed. These characteristics include color, texture, or objects represented in a picture that topical indexing would not include, but users may rely on when making relevance judgments [KOM 98]. Hence, it is very important to provide the users such visual cues in browsing. For this purpose, *key frames* extracted from the videos are one of the methods to provide visual surrogates of video data.

Many video browsing models propose to allow users to visualize video content based on user interactions [ARM 94, DAY 95, FLI 95, MIL 92, OOM 93, SMO 94, YEO 97]. These models choose representative images using regular time intervals, one image in each shot, all frames with focus key frame at specific place, and so on. Choosing key frames based on regular time intervals may miss some important segments and segments may have multiple key frames with similar contents. One image in each shot also may not capture the temporal and spatial relations of semantic objects. Showing all key frames may confuse users when too many key frames are displayed at the same time. To achieve a balance, we propose a key frame selection mechanism based on the number, temporal, and spatial changes of the semantic objects in the video frames.

The Augmented transition network (ATN), developed by Woods [WOO 70], has been used in natural language understanding systems and question answering systems for both text and speech. We use the ATN as a semantic model to model multimedia presentations [CHE 97a], multimedia database searching, the temporal, spatial, or spatio-temporal relations of various media streams and semantic objects [CHE 97b, SHY 98b]. As shown in [CHE 97c], ATNs need fewer nodes and arcs to represent a multimedia presentation compared with Petri-net models such as OCPN [LIT 90]. Multimedia input strings adopt the notations from regular expressions [KLE 56] and are used to represent the presentation sequences of temporal media streams, spatio-temporal relations of semantic objects, and keyword compositions. In addition to using ATNs to model multimedia presentations and multimedia database searching, how to use ATNs and multimedia input strings as video browsing models is discussed in this paper. Moreover, key frame selection based on the temporal and spatial relations of semantic objects in each shot will be discussed. In previous studies, formulations and algorithms for multiscale image segmentation and unsupervised video segmentation and object tracking were introduced [SIS 98, SIS 99b, SIS 99c]. Our video segmentation method focuses on obtaining object level segmentation, i.e., obtaining objects in each frame and their traces across the frames. Hence, the temporal and spatial relations of semantic objects required in the proposed key frame selection mechanism can be captured. We apply our video segmentation method on a small portion of a soccer game video and use the temporal and spatial relations of semantic objects to illustrate how the key frame selection mechanism works.

The organization of this paper is as follows. Section 2 discusses the use of ATNs and multimedia input strings to model video browsing. Key frame selection algorithm is introduced in section 3. Section 3 also gives an example soccer game video. Conclusions are presented in section 4.

2. Video Browsing Using ATNs

In an interactive multimedia information system, users should have the flexibility to browse and decide on various scenarios they want to see. This means that two-way communications should be captured by the conceptual model. Digital video has gained increasing popularity in many multimedia applications. Instead of sequential access to the video content, structuring and modeling video data so that users can quickly and easily browse and retrieve interesting materials becomes an important issue in designing multimedia information systems.

Browsing provides users the opportunity to view information rapidly since they can choose the content relevant to their needs. It is similar to the table of contents and the index of a book. The advantage is that users can quickly locate the interesting topic and avoid the sequential and time-consuming process. In a digital video library, in order to provide this capability, a semantic model should allow users to navigate a video stream based on shots, scenes, or clips. The ATN can be used to model the spatio-temporal relations of multimedia presentations and multimedia database systems. It allows users to view part of a presentation by issuing database queries. In this paper, we further design a mechanism by using the ATN to model video browsing so that users can navigate the video contents. In this manner, querying and browsing capabilities can be provided by using ATNs.

2.1. *Hierarchy for a Video Clip*

As mentioned in [YEO 97], a video clip can be divided into *scenes*. A *scene* is a common event or locale which contains a sequential collection of *shots*. A *shot* is a basic unit of video production which captures between a record and a stop camera operation. Figure 1 is a hierarchy for a video clip. At the topmost level is the video clip. A clip contains several *scenes* at the second level and each *scene* contains several *shots*. Each *shot* contains some contiguous *frames* which are at the lowest level in the video hierarchy. Since a video clip may contain many video frames, it is not good for database retrieving and browsing. How to model a video clip, based on different granularities, to accommodate browsing, searching and retrieval at different levels is an important issue in multimedia database and information systems. A video hierarchy can be defined by the following three properties:

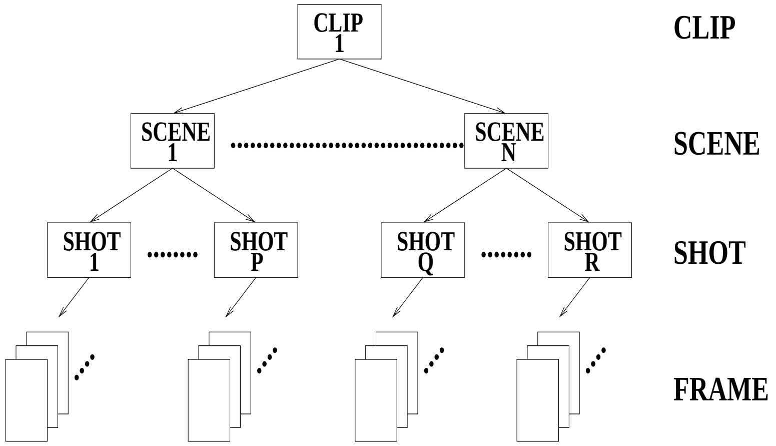


Figure 1. *A hierarchy of video media stream*

1. V = {S_1, S_2, . . ., S_N}, S_i denotes the ith scene and N is the number of scenes in this video clip. Let B(S_1) and E(S_1) be the starting and ending times of scene S_1, respectively. The temporal relation B(S_1) $<$ E(S_1) $<$ B(S_2) $<$ E(S_2) $<$ . . . is preserved.

2. S_i = {T_1^i, . . ., $T_{n_i}^i$}, T_j^i is the jth shot in scene S_i and n_i is the number of shots in S_i. Let B(T_1^i) and E(T_1^i) be the starting and ending times of shot T_1^i where B(T_1^i) < E(T_1^i) < . . . < B($T_{n_i}^i$) < E($T_{n_i}^i$).

3. T_j^i = {$R_1^{i,j}$, . . ., $R_{l_j}^{i,j}$}, $R_1^{i,j}$ and $R_{l_j}^{i,j}$ are the starting and ending frames in shot T_j^i and l_j is the number of frames for shot T_j^i.

In property 1, V represents a video clip and contains one or more *scenes* denoted by S_1, S_2, and so on. *Scenes* follow a temporal order. For example, the ending time of S_1 is earlier than the starting time of S_2. As shown in property 2, each *scene* contains some *shots* such as T_1^i to $T_{n_i}^i$. *Shots* also follow a temporal order and there is no time overlap among shots so B(T_1^i) $<$ E(T_1^i) $<$ . . . $<$ B($T_{n_i}^i$) $<$ E($T_{n_i}^i$). A *shot* contains some key frames to represent the visual contents and changes in each shot. In property 3, $R_k^{i,j}$ represents key frame k for shot T_j^i. The details of how to choose key frames based on temporal and spatial relations of semantic objects in each shot will be discussed in section 3.

2.2. *Using ATNs to Model Video Browsing*

An ATN can build up the hierarchy property by using its subnetworks. Figure 2 is an example of how to use an ATN and its subnetworks to represent a video hierarchy. An ATN and its subnetwork are capable of segmenting a video clip into different granularities and still preserve the temporal relations of different units.

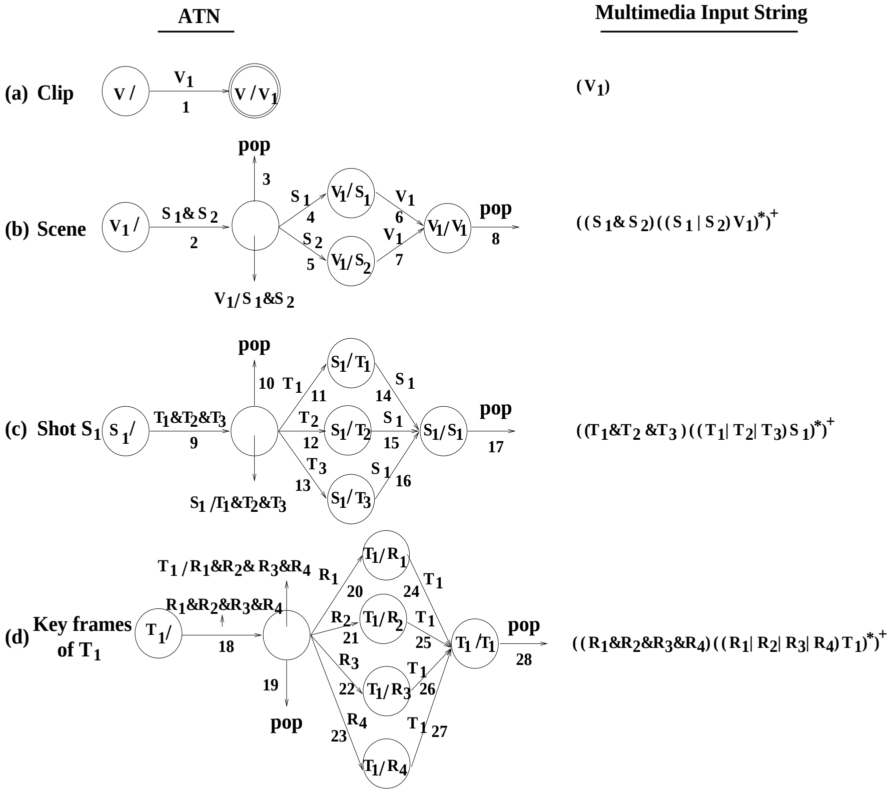


Figure 2. *Augmented Transition Network for video browsing: (a) is the ATN network for a video clip which starts at the state V/. (b)-(d) are part of the subnetworks of (a). (b) is to model scenes in video clip V_1. (c) is to model shots in scene S_1. Key frames for shot T_1 is in (d).*

Table 1 shows the traces of ATN for presentation V in Figure 2. This table is used to explain how ATN works for video browsing. The part of the steps are shown as follows:

Step 1: The current state is V and the arc to be followed is arc number 1 with arc label V_1. The input symbol V_1 is a subnetwork name (as shown in Figure 2(b)). Since input symbol V_1 (video clip) is a subnetwork name, the state name (V/V_1) at the head of arc 1 is put into a stack which is shown at backup states in Table 1. The control passes to the subnetwork V_1 (Figure 2(b)) after the state name is put into the stack.

Step 2: The current state is $V_1/$ which is the starting state of a subnetwork as shown in Figure 2(b). Arc number 2 is followed and the arc label is $S_1\&S_2$. Arc label $S_1\&S_2$ means a video clip V_1 consists of two scenes to let users choose and they are S_1 and S_2. Assuming the user chooses S_1, arc number 4 is followed and the arc label (input symbol) is S_1. Since S_1 is also a subnetwork name, the state name V_1/S_1 at the head of this arc is pushed into the stack so that this state name is on top of the state name V/V_1. Therefore, there are two state names in the stack at this stage. The control passes to the subnetwork in Figure 2(c).

Step 3: The current state is $S_1/$. Arc number 9 with arc label $T_1\&T_2\&T_3$ is followed. This arc label denotes that scene S_1 consists of three shots: T_1, T_2, and T_3.

In Figure 2(a), the arc label V_1 is the starting state name of its subnetwork in Figure 2(b). When the input symbol V_1 is read, the name of the state at the head of the arc (V/V_1) is pushed into the top of a push-down store. The control is then passed to the state named on the arc which is the subnetwork in Figure 2(b).

In Figure 2(b), when the input symbol X_1 ($S_1\&S_2$) is read, two frames which represent two video scenes S_1 and S_2 are both displayed for the selections. In the original video sequence, S_1 appears earlier than S_2 since it has a smaller number. The "&" symbol in multimedia input strings is used to denote the concurrent display of S_1 and S_2. ATNs are capable of modeling user interactions where different selections will go to different states so that users have the opportunity to directly jump to the specific video unit that they want to see. In our design, vertical bars "|" in multimedia input strings and more than one outgoing arc in each state at ATNs are used to model the "or" condition so that user interactions are allowed. Assume S_1 is selected, the input symbol S_1 is read. Control is passed to the subnetwork in Figure 2(c) with starting state name S_1/. The "*" symbol indicates the selection is optional for the users since it may not be activated if users want to stop the browsing. The subnetwork for S_2 is omitted for the simplicity.

In Figure 2(c), when the input symbol $T_1\&T_2\&T_3$ is read, three frames T_1, T_2, and T_3 which represent three shots of scene S_1 are displayed for the selection. If the shot T_1 is selected, the control will be passed to the subnetwork in Figure 2(d) based on the arc symbol T_1/. The same as in Figure 2(b), temporal flow is maintained.

3. The Proposed Key Frame Selection Approach

The next level under *shots* are key frames. Key frame selections play an important role to let users examine the key changes in each video shot. Since each shot may still have too many video frames, it is reasonable to use key frames to represent the *shots*. The easiest way of key frame selection is to choose the first frame of the shot. However, this method may miss some important temporal and spatial changes in each shot. The second way is to include all video frames as key frames and this may have computational and storage problems, and may increase users' perception burdens. The third way is to choose key frames based on fixed durations. This method is still not a good mechanism since it may give us many key frames with similar contents. Therefore, how to select key frames to represent a video *shot* is an important issue for digital library browsing, searching, and retrieval [YEU 95]. To achieve a balance, we propose

Table 1. *The trace of ATN for the browsing sequence in Figure 2.*

Step	Current State	Input Symbol	Arc Followed	Backup States
1	$V/$	V_1	1	V/V_1
2	$V_1/$	$S_1\&S_2$	2	V/V_1
3	$V_1/S_1\&S_2$	S_1	4	V_1/S_1 V/V_1
4	$S_1/$	$T_1\&T_2\&T3$	9	V_1/S_1 V/V_1
5	$S_1/T_1\&T_2\&T3$	T_1	12	S_1/T_1 V_1/S_1 V/V_1
6	$T_1/$	$R_1\&R_2\&R3\&R_4$	18	S_1/T_1 V_1/S_1 V/V_1
7	$T_1/R_1\&R_2\&R3\&R_4$	R_1	20	S_1/T_1 V_1/S_1 V/V_1
8	T_1/R_1	T_1	24	T_1/T_1 S_1/T_1 V_1/S_1 V/V_1
9	$T_1/$	$R_1\&R_2\&R_3\&R_4$	18	S_1/T_2 V_1/S_1 V/V_1
10	$T_1/R_1\&R_2\&R_3\&R_4$	None	19	V_1/S_1 V/V_1
11	S_1/T_2	S_1	15	S_1/S_1 V_1/S_1 V/V_1
12	$S_1/$	$T_1\&T_2\&T_3$	9	S_1/S_1 V_1/S_1 V/V_1
13	$S_1/T_1\&T_2\&T_3$	None	10	V_1/S_1 V/V_1
14	S_1/S_1	None	17	V/V_1
15	V_1/S_1	V_1	6	V_1/V_1 V/V_1
16	$V_1/$	$S_1\&S_2$	2	V_1/V_1 V/V_1
17	$V_1/S_1\&S_2$	None	3	V/V_1
18	V_1/V_1	None	8	NIL
19	Finish			

a key frame selection mechanism based on the number, temporal, and spatial changes of the semantic objects in the video frames. Other features may also be possible for the key frame selections, but we focus on the number, temporal, and spatial relations of semantic objects in this study. Therefore, spatio-temporal changes in each shot can be represented by these key frames. For example, in each shot of a soccer game, players may change positions in subsequent frames and the number of players appearing may change at the time duration of the shot.

3.1. *Simultaneous Partition and Class Parameter Estimation (SPCPE) Algorithm*

Let the set of semantic objects in the kth frame ($R_k^{i,j}$) of the jth shot T_j^i in the ith scene S_i be denoted by $O_k^{i,j}$. We define the key frame selections as follows:

Definition 1: Given two contiguous video frames $R_a^{i,j}$ and $R_b^{i,j}$ in T_j^i, let the sets of the semantic objects in these two video frames be $O_a^{i,j}$ and $O_b^{i,j}$. $R_b^{i,j}$ is a key frame if and only if any of following two conditions is satisfied:

(1) $O_a^{i,j} \cap O_b^{i,j} \neq O_a^{i,j} \cup O_b^{i,j}$;

(2) Any semantic object spatial location changes between $O_a^{i,j}$ and $O_b^{i,j}$.

As mentioned previously, the video segmentation method can provide the required information for the key frame selection mechanism. Therefore, the video segmentation method is applied to each frame before the above two conditions are checked. 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. So the minimum we obtain through our descent method depends strongly on the starting point or the initial partition. In a video, the successive frames do not differ much due to the high temporal sampling rate. Hence the partitions of adjacent frames do not differ significantly. Starting with the estimated partition of the previous frame, if we apply our descent algorithm on the current frame we may obtain a new partition that is not significantly different from the partition of the previous frame. For the first frame, since there is no previous frame, we use a randomly generated initial partition.

We treat the partition as well as the class parameters as random variables and pose the problem as one in joint estimation [SIS 98, SIS 99b]. Suppose there are two classes. Let the two classes be described by the probability densities denoted by $p_1(y_{ij})$ and $p_2(y_{ij})$. Also, let the partition variable be $\boldsymbol{c} = \{\boldsymbol{c}_1, \boldsymbol{c}_2\}$ and the classes be parameterized by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$. Consider an image of N_r rows and N_c columns (i.e., image of size $N_r \times N_c$) with intensities given by $Y = \{y_{ij} : 1 \leq i \leq N_r, 1 \leq j \leq N_c\}$. We estimate the best partition as that which maximizes the a posteriori probability (MAP) of the partition variable given the image data Y. Now, the MAP estimates of $\boldsymbol{c} = \{\boldsymbol{c}_1, \boldsymbol{c}_2\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ are given by

$$(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\theta}}) = \text{Arg}\max_{(\boldsymbol{c},\boldsymbol{\theta})} P(\boldsymbol{c}, \boldsymbol{\theta} \mid Y) = \text{Arg}\max_{(\boldsymbol{c},\boldsymbol{\theta})} P(Y \mid \boldsymbol{c}, \boldsymbol{\theta}) P(\boldsymbol{c}, \boldsymbol{\theta}). \qquad [1]$$

Let $J(\boldsymbol{c}, \boldsymbol{\theta})$ denote the functional that needs to be minimized, i.e., the sum of terms. With appropriate assumptions, this joint estimation can be simplified to the following form:

$$\begin{aligned} (\hat{\boldsymbol{c}}, \hat{\boldsymbol{\theta}}) &= \text{Arg}\min_{(\boldsymbol{c},\boldsymbol{\theta})} J(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ J(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \sum_{y_{ij} \in \boldsymbol{c}_1} -\ln p_1(y_{ij}; \boldsymbol{\theta}_1) + \sum_{y_{ij} \in \boldsymbol{c}_2} -\ln p_2(y_{ij}; \boldsymbol{\theta}_2). \qquad [2] \end{aligned}$$

The joint estimation method is called the *simultaneous partition and class parameter estimation (SPCPE)* algorithm. The algorithm starts with an arbitrary partition of the data and computes the corresponding class parameters. Using these class parameters and the data a new partition is estimated. Both the partition and the class parameters are iteratively refined until there is no further change in them. The details of the video segmentation method are shown in [SIS 99c].

Given a video shot T^i_j, let K^i_j be the set of key frames selected for T^i_j and m a frame index. Initially the first frame is always selected so $K^i_j = \{R^{i,j}_1\}$.

```
1. Initialization:
   • $K^i_j = \{R^{i,j}_1\}$;
   • Execute SPCPE algorithm for the first frame;
2. for $m$ = 2 to $l_j$
   • Execute SPCPE algorithm to get the temporal and spatial relations of the
     semantic objects;
   • if ( ($O^{i,j}_m \cap O^{i,j}_{m-1} \neq O^{i,j}_m \cup O^{i,j}_{m-1}$) OR
     Spatial_location_change($O^{i,j}_m$, $O^{i,j}_{m-1}$) ) then
     $K^i_j = K^i_j \cup R^{i,j}_m$;
   endfor;
```

The first condition of definition 1 models the number of semantic object changes in two contiguous video frames at the same shot. The first part of the if-statement in the above solution algorithm is used to check this situation. The latter part of the if-statement checks the second condition of definition 1, which is to model the temporal and spatial changes of semantic objects in two contiguous video frames of the shot. Using the same definition of three dimensional relative positions for semantic objects as shown in [CHE 97b], we choose one semantic object to be the target semantic object in each video frame. We adopt the minimal bounding rectangle (MBR) concept in R-tree [GUT 84] so that each semantic object is covered by a rectangle. In order to distinguish the relative positions, twenty-seven numbers are used to distinguish the relative positions of each semantic object relative to the target semantic object and are represented by subscripted numbers. 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.

3.2. *Implementation and Results*

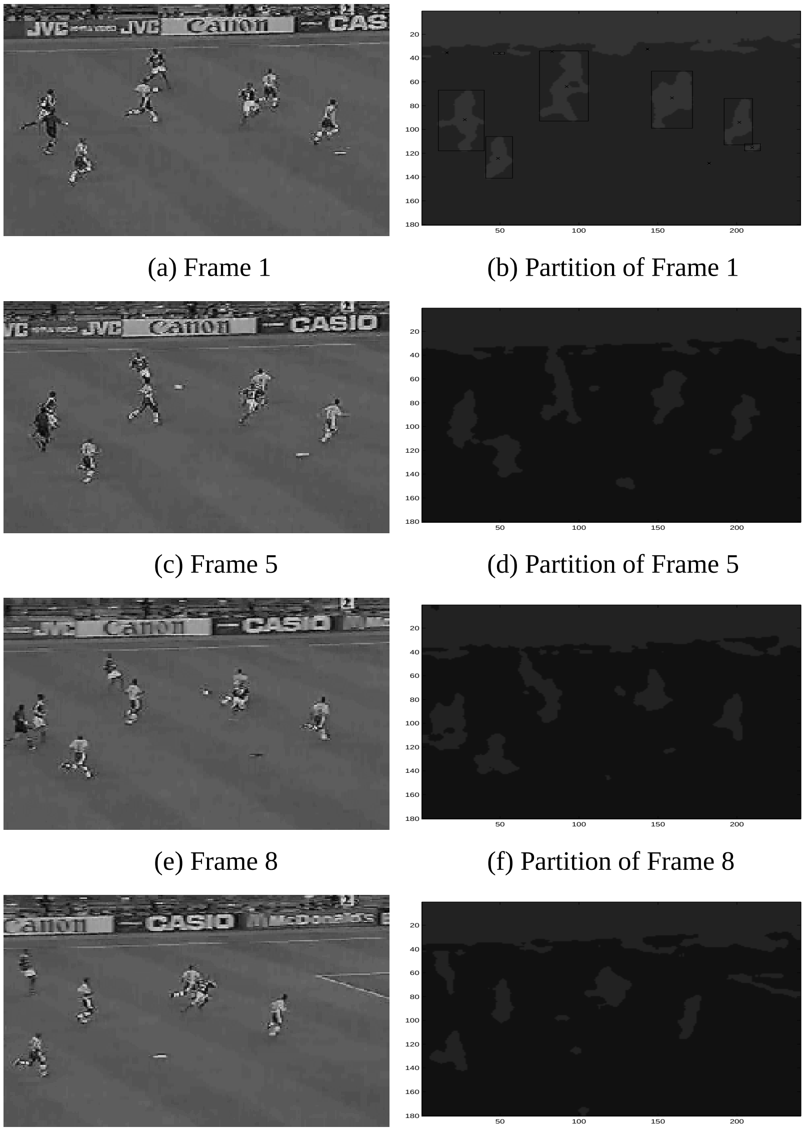
The example soccer video consists of 60 frames. It is a gray scale video that shows the part of the game where a goal is scored. Each frame is of size 180 rows and 240 columns. A small portion of the soccer video game is used to illustrate the way the proposed key frame selection mechanism works. Although we have several distinct regions in each frame of the video, only two of them are important from the content based retrieval perspective, namely the ball and the players. There are some important aspects in this video that make automatic object tracking difficult. They are as follows:

– The soccer ball vanishes between players for a few frames and reappears later.

– The regions corresponding to the players merge together and separate out after a few frames.

– Some spurious patches, typically on the ground, suddenly appear as blobs and disappear giving the impression of an object.

We will apply our video segmentation method to this data, assuming that there are two classes. The first frame is partitioned using the multiscale frame segmentation with two classes. The algorithm is initialized with a random starting partition. After obtaining the partition of the first frame, we compute the partitions of the subsequent frames. From the results on frames 1 through 60, a few frames – 1, 5, 8 and 13 – are shown in Figure 3 along with the original frames adjacent to them. As can be seen from Figure 3, the players, the soccer ball and the sign boards in the background (JVC, Canon, etc.) are all captured by a single class. The ground in the soccer field is captured by another class. Some of the players who are close together have been combined into a single segment. Similarly, the ball is merged into a single segment with two other players. For example, in frame 1, the ball and two players are part of one segment; whereas by the fifth frame, the soccer ball is far away so that it becomes a segment in itself. This continues until it goes in between two other players. Notice the patch on the ground which was near the right most player in the first frame, moves to the left uniformly owing to the camera panning to the right. In frame 5 we can see a spurious patch appearing out of nowhere. On the whole, the initial conditions from the previous frames seem to be guiding the segmentation of the current frame in an effective manner. There are some artifacts on the ground, specifically the one closest to the rightmost player, which show up as patches in the final partition. Inspection of the other frames shows that it is also present in them and not something spurious.



(a) Frame 1

(b) Partition of Frame 1

(c) Frame 5

(d) Partition of Frame 5

(e) Frame 8

(f) Partition of Frame 8

(g) Frame 13

(h) Partition of Frame 13

Figure 3. *Figures (a),(c),(e),(g) are the original Frames 1,5,8,13 (on the left) and (b),(d),(f),(h) show their corresponding partitions (on the right). (b) shows the segments extracted from the first frame of the Soccer video. The centroid of each segment is marked with an 'x' and the segment is shown with a bounding box around it. The segments corresponding to the moving players and the ball are captured in every frame automatically.*

The segments of Frame 1, extracted by applying the seeding and region growing method are shown in Figure 3(b). There are 15 segments in this frame out of which only 5 correspond to the players and the ball. The ball and 2 players are merged into one segment, and there are 2 other segments where two players are put into a single segment. The rest of the two segments consist of one player in each segment. We have implemented the programs to find the bounding boxes and the centroids for the segments. Therefore, the segments are displayed with the bounding boxes around them and the centroids are marked with an ‘x’ in Figure 3(b). The small segments with only a centroid and without any apparent bounding box are the ones with very few pixels. Most of them are on the top of the frame and at the bottom of the sign boards. They arise out of smoothing the broken soccer boundary line.

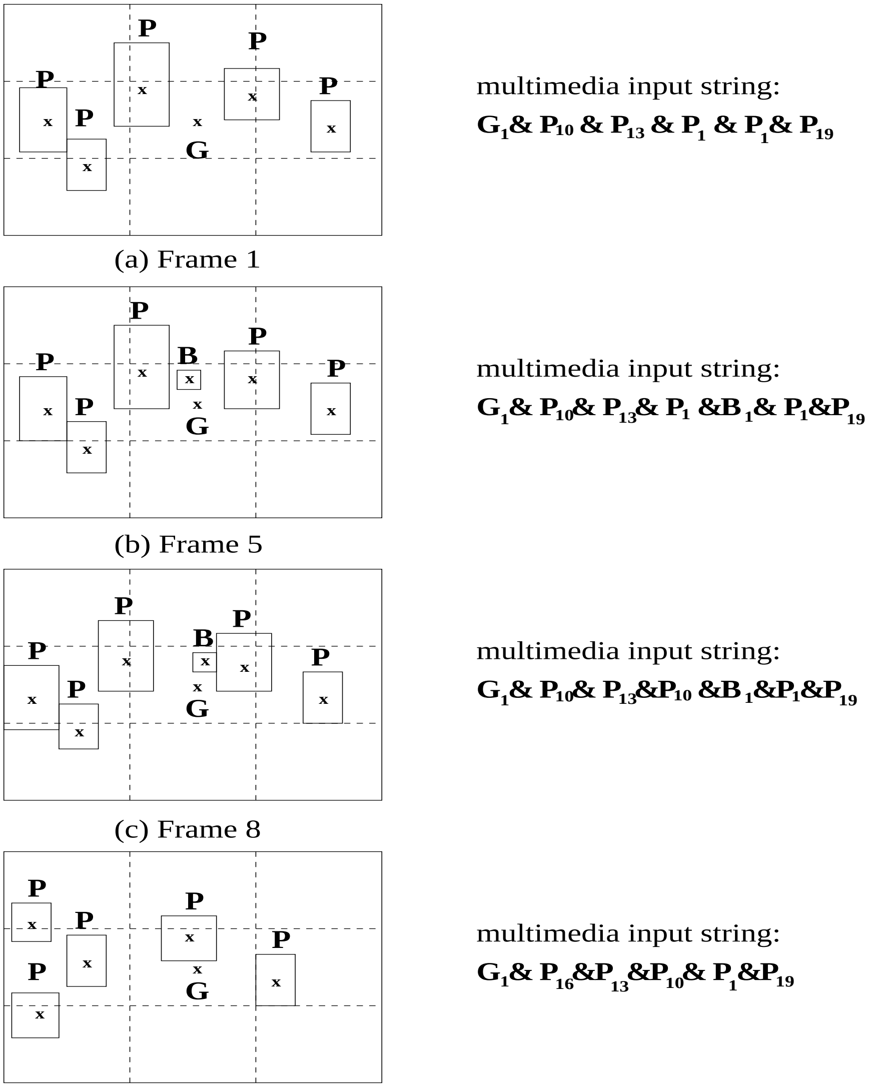
Since only the ball and the players are important from the content based retrieval perspective, we use Figure 4 to simplify the segments for each frame. As shown in Figure 4, the ground (**G**) is selected as the target semantic object and the segments are denoted by **P** for the players or **B** for the soccer ball. As mentioned earlier, if two semantic objects are too close to each other, they are merged into a single segment. Hence, the soccer ball is put into a single segment only when it is far away from the players (in Frames 5 and 8) and each segment **P** may consist of multiple players and/or the soccer ball. In this example, each frame is divided into nine subregions. More or fewer subregions in a video frame may be used to allow more fuzzy or more precise queries as necessary. The corresponding multimedia input strings are on the right of Figure 4. In our design, each key frame is represented by an input symbol in a multimedia input string and the “&” symbol between two semantic objects is used to denote that the semantic objects appear in the same frame. The subscripted numbers are used to distinguish the relative positions of the semantic objects relative to the target semantic object “ground”. Table 2 shows part of the three dimensional spatial relations introduced in [CHE 97b]. (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. Since two dimensions are considered in this example, $z_s \approx z_t$. The multimedia input strings can be used for multimedia database searching via substring matching. The details of multimedia database searching are shown in [CHE 97b].

Assume Figures 4(a), (b), (c), and (d) are four key frames for shot T_1. The multimedia input string to represent these four key frames is as follows:

Multimedia input string:

$$\underbrace{(G_1\&P_{10}\&P_{13}\&P_1\&P_1\&P_{19})}_{M_1}\underbrace{(G_1\&P_{10}\&P_{13}\&P_1\&B_1\&P_1\&P_{19})}_{M_2}$$
$$\underbrace{(G_1\&P_{10}\&P_{13}\&P_{10}\&B_1\&P_1\&P_{19})}_{M_3}\underbrace{(G_1\&P_{16}\&P_{13}\&P_{10}\&P_1\&P_{19})}_{M_4}$$

As shown in the above multimedia input string, there are four input symbols which are M_1, M_2, M_3, and M_4. The appearance sequence of the semantic objects in an



multimedia input string:
$\mathbf{G_1 \& P_{10} \& P_{13} \& P_1 \& P_1 \& P_{19}}$

(a) Frame 1

multimedia input string:
$\mathbf{G_1 \& P_{10} \& P_{13} \& P_1 \& B_1 \& P_1 \& P_{19}}$

(b) Frame 5

multimedia input string:
$\mathbf{G_1 \& P_{10} \& P_{13} \& P_{10} \& B_1 \& P_1 \& P_{19}}$

(c) Frame 8

multimedia input string:
$\mathbf{G_1 \& P_{16} \& P_{13} \& P_{10} \& P_1 \& P_{19}}$

(d) Frame 13

Figure 4. *Segments with bounding boxes and centroids for Frames 1,5,8,13 in Figure 3 on the left and their corresponding multimedia input strings on the right. Each segment is displayed with the bounding box around it and the centroid is marked with an ‘x’. Here,* **G, P**, *and* **B** *represent “ground”, “players”, and “soccer ball”, respectively. The “ground” (***G***) is selected as the target semantic object and the subscripted numbers in a multimedia input string are used to distinguish the relative positions of the semantic objects relative to* **G**. *Each frame is divided into nine subregions and the centroid of each segment is used as a reference point for spatial reasoning.*

Table 2. *Part of the three dimensional relative positions for semantic objects: The first and the third columns indicate the relative position numbers while the second and the fourth columns are the relative coordinates.* (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.*

Number	Relative Coordinates
1	$x_s \approx x_t, y_s \approx y_t, z_s \approx z_t$
10	$x_s < x_t, y_s \approx y_t, z_s \approx z_t$
13	$x_s < x_t, y_s < y_t, z_s \approx z_t$
16	$x_s < x_t, y_s > y_t, z_s \approx z_t$
19	$x_s > x_t, y_s \approx y_t, z_s \approx z_t$

input symbol is based on the spatial locations of the semantic objects in the video frame from left to right and top to bottom. For example, Figure 4(a) is represented by input symbol M_1. G_1 indicates that **G** is the target semantic object. P_{10} means the first **P** is on the left of **G**, P_{13} means the second **P** is below and to the left of **G**, P_1 means the third **P** and the fourth **P** are at the same subregion as **G**, and P_{19} means the fifth **P** is on the right of **G**. Figure 4(b) is modeled by input symbol M_2 in which the soccer ball **B** appears at the same subregion as **G** and the rest of the semantic objects remain at the same locations. In this case, the number of semantic objects is increased from six to seven. This is an example to show how to use a multimedia input string to represent a number of semantic object changes. Figure 4(c) is represented by input symbol M_3. The third **P** moves from the same subregion of **G** to above and left of **G** so the associated number changes from 1 to 10 from which the relative spatial relations can also be modeled by the multimedia input string. Input symbol M_4 models Figure 4(d). In this situation, **B** disappears and the first **P** changes its spatial location from the left to above and left of **G** in Figure 4(c). So, the number associated with the first **P** changes from 10 to 16 and **B** does not exist in M_4. The order of these four key frames is modeled by four input symbols concatenated together to indicate that M_1 appears earlier than M_2 and so on.

4. Conclusions

Video data are widely used in today's multimedia applications such as education, video on demand, video conferencing and so on. Managing video data so that users can quickly browse video data is an important issue for the multimedia applications using video data. A good semantic model is needed if we want to meet the needs. In this paper, ATNs are used to model video hierarchy for browsing. Based on this design, users can view information quickly to decide whether the content is what they want to see. Key frames selection based on temporal and spatial relations of semantic objects is used in our design. The temporal and spatial relations of semantic objects are captured by the proposed unsupervised video segmentation method. From the soccer game video example, we can see that the players and the soccer ball are captured well. Since the first frame uses a random initialization and the subsequent frames use the results of the previous frames, the method is completely unsupervised. In addition, by incorporating the partition information of the previous frame into the segmentation process of the current frame, the temporal information is implicitly used. Under this design, these key frames preserve many of the visual contents and minimize the data size to mitigate the computation and storage problems in multimedia browsing environments. Moreover, based on the results of the segmentation, multimedia input strings are constructed. The multimedia input strings can be used for multimedia database searching via substring matching. Unlike the existing semantic models which only model presentation, query, or browsing, our ATN model provides these three capabilities in one framework.

Acknowledgements

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

5. References

[ARM 94] ARMAN F., DEPOMMER R., HSU A., CHIU M.Y., “Content-based browsing of video sequences”, *ACM Multimedia 94*, p. 97-103, Aug. 1994.

[CHE 97a] CHEN S.-C., KASHYAP R.L., “Temporal and Spatial Semantic Models for Multimedia Presentations”, in *1997 International Symposium on Multimedia Information Processing*, p. 441-446, Dec. 11-13, 1997,

[CHE 97b] CHEN S.-C., KASHYAP R.L., “A Spatio-Temporal Semantic Model for Multimedia Presentations and Multimedia Database Systems”, accepted for publication *IEEE Transactions on Knowledge and Data Engineering*, 2000.

[CHE 97c] CHEN S.-C., KASHYAP R.L., “Empirical Studies of Multimedia Semantic Models for Multimedia Presentations”, in *13th International Conference on Computer and Their Applications*, p. 226-229, March 25-27, 1998.

[COU 97] COURTNEY J.D., “Automatic Video Indexing via Object Motion Analysis”, *Pattern Recognition*, vol. 30, no. 4, p. 607-625, 1997.

[DAY 95] DAY Y.F., DAGTAS S., IINO M., KHOKHAR A., GHAFOOR A., “Object-Oriented Concept Modeling of Video Data”, *IEEE Int’l Conference on Data Engineering*, p. 401-408, March 1995.

[FER 97] FERMAN A.M., GUNSEL B., TEKALP A.M., “Object Based Indexing of MPEG-4 Compressed Video”, in *Proc. SPIE: VCIP*, p. 953-963, vol. 3024, San Jose, USA, February 1997.

[FLI 95] FLICKNER M., SAWHNEY H., NIBLACK W., ASHLEY J., HUANG Q., DOM B., GORKANI M., HAFNER J., LEE D., PETKOVIC D., STEELE D., YANKER P., “Query by Image and Video Content: The QBIC System", *IEEE Computer*, Vol. 28, No. 9, p. 23-31, September 1995.

[GUT 84] GUTTMAN A., "R-tree: A Dynamic Index Structure for Spatial Search", in *Proc. ACM SIGMOD*, p. 47-57, June 1984.

[HOL 98] HOLLFELDER S., EVERTS A., THIEL U., "Concept-Based Browsing in Video Libraries", *Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries*, 1998.

[SIS 98] KASHYAP R.L. and SISTA S., "Unsupervised Classification and Choice of Classes: Bayesian Approach", Technical Report TR-ECE 98-12, School of Electrical and Computer Engineering, Purdue University, July 1998.

[KLE 56] KLEENE S.C., "Representation of Events in Nerve Nets and Finite Automata, Automata Studies", *Princeton University Press*, Princeton, N.J., p. 3-41, 1956.

[KOM 98] KOMLODI A. and SLAUGHTER L., "Visual Video Browsing Interfaces Using Key Frames", *Proceedings of the CHI 98 Summary Conference on CHI 98 Summary: Human Factors in Computing Systems*, p. 337-338, 1998.

[LIT 90] LITTLE T.D.C., GHAFOOR A., "Synchronization and Storage Models for Multimedia Objects", *IEEE J. Selected Areas in Commun.*, Vol. 9, p. 413-427, Apr. 1990.

[MIL 92] MILLS M., COHEN J., WONG Y.Y., "A magnifier tool for video data", in *Proc. ACM Computer Human Interface (CHI)*, May, 1992, p. 93-98.

[OOM 93] OOMOTO E., TANAKA K., "OVID: Design and Implementation of a Video Object Database System", *IEEE Trans. on Knowledge and Data Engineering*, Vol. 5, No. 4, p. 629-643, August 1993.

[SHY 98b] SHYU M.-L., CHEN S.-C., KASHYAP R.L., "Information Retrieval Using Markov Model Mediators in Multimedia Database Systems", *1998 International Symposium on Multimedia Information Processing*, p. 237-242, Dec. 14-16, 1998.

[SIS 99b] SISTA S., KASHYAP R.L., "Bayesian Estimation for Multiscale Image Segmentation", *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, March 1999.

[SIS 99c] SISTA S., KASHYAP R.L., "Unsupervised video segmentation and object tracking", in *IEEE Int'l Conf. on Image Processing*, 1999.

[SMO 94] SMOLIAR S.W., ZHANG H.J. "Content-based video indexing and retrieval", *IEEE Multimedia*, p. 62-72, Summer, 1994.

[WOO 70] WOODS W., "Transition Network Grammars for Natural Language Analysis", *Comm. of the ACM*, **13**, October 1970, p. 591-602.

[YEO 97] YEO B.-L., YEUNG M.M., "Retrieving and Visualization Video", *Comm. of the ACM*, Vol. 40, No. 12, December 1997, p. 43-52.

[YEU 95] YEUNG M.M., LIU B., "Efficient Matching and Clustering of Video Shots", in *IEEE International Conference on Image Processing*, Vol I, October, 1995, p. 338-341.