Journal of Intelligent Information Systems, 19:1, 61–77, 2002
© 2002 Kluwer Academic Publishers. Manufactured in The Netherlands.

A Multimedia Data Mining Framework: Mining Information from Traffic Video Sequences

SHU-CHING CHEN
Distributed Multimedia Information System Laboratory, School of Computer Science, Florida International University, Miami, FL 33199, USA

MEI-LING SHYU
Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124, USA

CHENGCUI ZHANG
JEFF STRICKROTT
Distributed Multimedia Information System Laboratory, School of Computer Science, Florida International University, Miami, FL 33199, USA

Received May 27, 2001; Revised July 5, 2001; Accepted January 31, 2002

Abstract. The analysis and mining of traffic video sequences to discover important but previously unknown knowledge such as vehicle identification, traffic flow, queue detection, incident detection, and the spatio-temporal relations of the vehicles at intersections, provide an economic approach for daily traffic monitoring operations. To meet such demands, a multimedia data mining framework is proposed in this paper. The proposed multimedia data mining framework analyzes the traffic video sequences using background subtraction, image/video segmentation, vehicle tracking, and modeling with the multimedia augmented transition network (MATN) model and multimedia input strings, in the domain of traffic monitoring over traffic intersections. The spatio-temporal relationships of the vehicle objects in each frame are discovered and accurately captured and modeled. Such an additional level of sophistication enabled by the proposed multimedia data mining framework in terms of spatio-temporal tracking generates a capability for automation. This capability alone can significantly influence and enhance current data processing and implementation strategies for several problems vis-à-vis traffic operations. Three real-life traffic video sequences obtained from different sources and with different weather conditions are used to illustrate the effectiveness and robustness of the proposed multimedia data mining framework by demonstrating how the proposed framework can be applied to traffic applications to answer the spatio-temporal queries.

Keywords: multimedia data mining, spatio-temporal relationships, multimedia augmented transition network (MATN), vehicle tracking

1. Introduction

The rapid progress of data collection tools, advanced database system technologies, and the World Wide Web (WWW) technologies has precipitated the explosive growth of vast amounts of data in various forms. Hence, there is an increasing need for multimedia data mining to discover important and previously unknown knowledge from the complex types of data. For example, municipalities (Caltrans, Montgomery) have installed video camera systems to monitor and extract traffic control information from their highways.

The analysis and mining of traffic video sequences to discover information, such as vehicle identification, traffic flow, queue detection, incident detection, and the spatio-temporal relations of the vehicles at intersections, provide an economic approach for daily traffic monitoring operations. Issues associated with extracting traffic movement and recognizing accident information from real time video sequences are discussed in Cucchiara et al. (2000), Dailey et al. (2000), Huang et al. (1994) and Kamijo et al. (1999). Two common themes exist in these works. First, the video information must be segmented and turned into objects. Second, the behavior of those objects is monitored (they are tracked) for immediate decision making purposes. What is missing in these efforts is to model and index the data for on-line analysis, storage or later pattern mining.

In order to identify and track the temporal and relative spatial positions of vehicle objects in video sequences, it is necessary to have object-based representation of video data. For this purpose, attention has been 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 (Courtney, 1997; Fan and Sung, 2000; Ferman et al., 1997). While most of the previous work is based on low-level global features, such as color histogram and texture, our video segmentation method focuses on obtaining object level segmentation, i.e., obtaining objects in each frame and their traces across the frames. In Chen et al. (2000a, 2001) we have addressed the issues of unsupervised image segmentation, object modeling with multimedia input strings to capture the spatial-temporal behavior of the objects, and the application of these techniques to the domain of traffic monitoring.

When the image segmentation problem is considered for a fixed camera domain, a classic technique to resolve the foreground objects is background subtraction (Gonzalez and Woods, 1993). This involves the creation of a background model that is subtracted from the input image to create a difference image. The new difference image only contains objects not in the background or new features that have not yet been incorporated into the background. Various approaches to background subtraction and modeling techniques have been discussed in the literature (Dailey et al., 2000; Grimson et al., 1998; Haritaoglu et al., 2000; Stauffer and Grimson, 1999), ranging from modeling the intensity variations of a pixel via a mixture of Gaussian distributions to simple differencing of successive images. In Toyama et al. (1999) the authors provide some simple guidelines and evaluation of the various techniques for background modeling. We are in the beginning phases of evaluating the performance benefits of background subtraction methods for the various domains of our image segmentation applications. To that aim we have evaluated the effectiveness of the image averaging techniques over stationary (non-changing) portions of the image data set.

In this paper, a multimedia data mining framework for traffic video sequences is proposed. The proposed framework considers image/video segmentation with initial background subtraction, vehicle tracking, and modeling with the multimedia augmented transition network (MATN) model and multimedia input strings (Chen et al., 1999; Chen and Kashyap, 2001), in the domain of traffic monitoring over traffic intersections. The multimedia input strings are used to capture the spatio-temporal relationships of vehicle objects thereafter. Based on the historical information of vehicle objects, some previously unknown or non-intuitive knowledge (such as traffic flow and queue detection) can be figured out and be used to support decision making. The video segmentation method mentioned here is unsupervised. An

advantage of the proposed framework is that it uses the segmentation result of the previous video frame to speed up the segmentation process of the current video frame.

Experiments are conducted to illustrate the effectiveness and robustness of the proposed framework using three real-life traffic video sequences. A portion of each traffic video sequence was used to demonstrate how the proposed framework can be applied to traffic applications to answer the spatio-temporal queries such as "Estimate the traffic flow of the road intersection from 8:00 AM to 8:30 AM." The query requires the use of the proposed multimedia data mining framework to discover information such as the number of vehicles passing through the corresponding road intersection in a given time duration as well as the types of the vehicles (e.g., "car", "bus", etc.). The process can be done on-line or off-line.

The organization of this paper is as follows. In the next section, the knowledge discovery process that includes background subtraction, the unsupervised segmentation algorithm, vehicle tracking techniques, MATN model, and multimedia input strings are introduced. Experimental results and analysis of the proposed multimedia data mining framework are discussed in Section 3. Along with the discussion, three real-life example traffic video sequences are used. Conclusions are presented in Section 4.

2. Mining information from traffic video sequences

Traffic video analysis can discover and provide useful information, such as queue detection, vehicle classification, traffic flow, and incident detection at intersections. To the best of our knowledge, the current multimedia based transportation applications and research work either do not connect to databases or have limited capabilities to index and store the collected data (such as traffic videos) in their databases. Therefore, those applications cannot provide organized, unsupervised, conveniently accessible, and easy-to-use multimedia information to the traffic planners. In order to discover and provide some important but previously unknown knowledge from the traffic video sequences to the traffic planners, multimedia data mining techniques need to be employed. The proposed multimedia data-mining framework includes background subtraction, vehicle object identification and tracking, MATN model, and multimedia input strings. The additional level of sophistication enabled by the proposed framework, in terms of spatio-temporal tracking, generates a capability for automation. This capability alone can significantly influence and enhance current data processing and implementation strategies for several problems vis-á-vis traffic operations.

MATNs and multimedia input strings are used to model the temporal and relative spatial relations of the vehicle objects. An unsupervised video segmentation method, i.e., the SPCPE algorithm (see Section 2.2), can identify vehicle objects. In our framework, the background subtraction technique is introduced to enhance the basic SPCPE algorithm to provide better segmentation results, so that more accurate spatio-temporal relationships of the vehicle objects can be obtained. In the following subsections: first, the background subtraction technique is described, and second, an overview of the SPCPE algorithm and the vehicle tracking techniques are presented. After that presentation, we will briefly describe the use of MATNs and multimedia input strings to model key video frames. A portion of the traffic video clips is used to demonstrate how video indexing is modeled by the MATNs and multimedia input strings.

2.1. *Background subtraction*

Background subtraction is a technique to remove non-moving components from a video sequence. The main assumption for its applications is that the camera remain stationary. The basic principle is to create a reference frame of the stationary components in the image. Once created, the reference frame is subtracted from any subsequent images. Those pixels resulting from new (moving) objects will generate a difference not equal to zero (i.e., difference $\neq 0$).

Our approach is similar to that of Haritaoglu et al. (1998) and Kamijo et al. (1999), where a reference frame is constructed by accumulating and averaging images of the target area (the intersection in our case) for some time interval. As mentioned in Haritaoglu et al. (2000), this is not a robust technique as it is sensitive to intensity variations. That is, it can generate false positives (false detection of moving objects) solely due to lighting variations (such as cloud motion blocking the sun). In addition it can also generate false negatives (detection of non-moving objects as a moving object) due to the addition of stationary objects to the scene that are not part of the reference frame. Toyama et al. (1999) provides a good summary of the problems associated with background modeling. In order to lessen the intensity variation sensitivity, we propose to use an image averaging technique together with our unsupervised segmentation algorithm for background subtraction in this paper. This allows us to quickly evaluate an upper limit on the performance improvement to our segmentation algorithm. In the reported experiments, three traffic video sequences under different weather conditions are used to test the robustness of our framework.

In our current approach, video sequences containing non-moving objects were manually selected from the video data and then averaged together. Figure 1 gives an example result of background subtraction for frame 32 in one of the traffic video sequences. This traffic video sequence consists of about 16 minutes of video from a traffic intersection in Miami, FL under approximately constant lighting conditions. The raw difference image is obtained by subtracting the reference frame (as shown in figure 1(b)) from the current image (as shown in figure 1(a)), and then the results are further normalized to create the difference image (as shown in figure 1(c)). The results of the differencing step are fed to our unsupervised segmentation algorithm as the input images. Binary thresholding of the difference image (as shown in figure 1(d)) can be used as an initial partition to improve the speed of convergence (see Section 2.2) in our segmentation algorithm. For simplicity, a fixed binary threshold is used in our framework due to the fact that the initial partition does not need to be very precise.

2.2. *Unsupervised video segmentation (SPCPE) method*

The Simultaneous Partition and Class Parameter Estimation (SPCPE) algorithm is an unsupervised video segmentation method to partition video frames. A given class description determines a partition. Similarly, a given partition gives rise to a class description, so the partition and the class parameters have to be estimated simultaneously. In practice, the class descriptions and their parameters are not readily available. An additional difficulty arises when images have to be partitioned automatically without the intervention of the user. Thus, we do not know a priori which pixels belong to which class. In the SPCPE algorithm, the

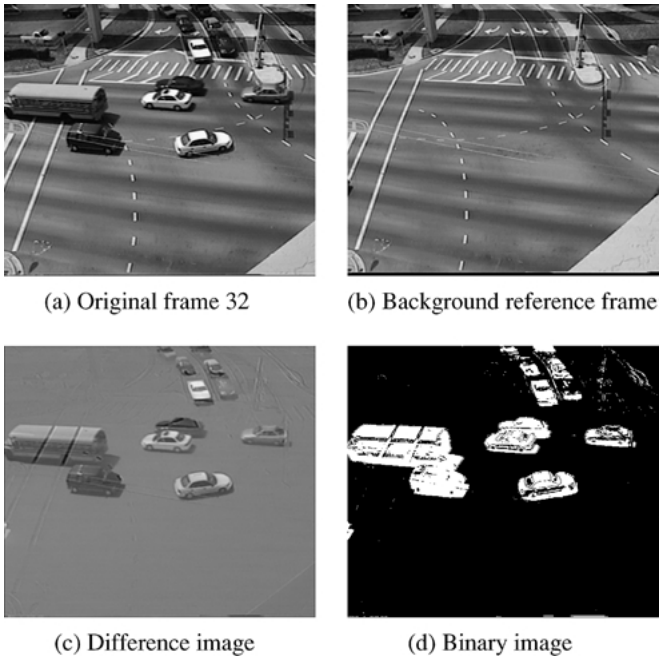


Figure 1. Example result of background subtraction.

partition and the class parameters are treated as random variables. The method for partitioning a video frame starts with an arbitrary partition and employs an iterative algorithm to estimate the partition and the class parameters jointly (Chen et al., 2000b; Sista and Kashyap, 2000). Since (for a smoothly varying video sequence) successive frames in a video do not differ much, the partitions of adjacent frames do not differ significantly. Each frame is partitioned using the partition of the previous frame as an initial condition to speed up the convergence rate of the algorithm. The initial partition for the first frame can be either a randomly generated partition, a learned partition for the domain, or a binary image derived from the background difference.

The mathematical description of a class specifies the pixel values as functions of the spatial coordinates of the pixel. The parameters of each class can be computed directly by using a least squares technique. Suppose we have two classes. Let the partition variable be $\boldsymbol{c} = \{\boldsymbol{c}_1, \boldsymbol{c}_2\}$ and the classes be parameterized by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$. Also, suppose all the pixel values y_{ij} (in the image data Y) belonging to class k ($k = 1, 2$) are put into a vector Y_k. Each row of the matrix Φ is given by $(1, i, j, ij)$ and a_k is the vector of parameters $(a_{k0}, \ldots, a_{k3})^T$.

$$
\begin{aligned}
y_{ij} &= a_{k0} + a_{k1} i + a_{k2} j + a_{k3} ij, \quad \forall (i, j) \;\; y_{ij} \in \boldsymbol{c}_k \\
Y_k &= \Phi a_k \\
\hat{a}_k &= (\Phi^T \Phi)^{-1} \Phi^T Y_k
\end{aligned}
$$

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

$$(\hat{\boldsymbol{c}}, \hat{\boldsymbol{\theta}}) = \operatorname{Arg}\max_{(\boldsymbol{c},\boldsymbol{\theta})} P(\boldsymbol{c}, \boldsymbol{\theta} \mid Y)$$
$$= \operatorname{Arg}\max_{(\boldsymbol{c},\boldsymbol{\theta})} P(Y \mid \boldsymbol{c}, \boldsymbol{\theta}) P(\boldsymbol{c}, \boldsymbol{\theta})$$

We assume that the pixel values and parameters are independent and the parameters are uniformly distributed. We also assume that the error function[1] of y_{ij} is represented by a Gaussian with mean 0 and variance 1. Let $J(\boldsymbol{c}, \boldsymbol{\theta})$ be the functional to be minimized. With these assumptions, the joint estimation can be simplified to the following form:

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

The minimization of J can be carried out alternately on $\boldsymbol{c}$ and $\boldsymbol{\theta}$ in an iterative manner. Let $\hat{\boldsymbol{\theta}}(\boldsymbol{c})$ represent the least squares estimate of the class parameters for a given partition $\boldsymbol{c}$. The final expression for $J(\boldsymbol{c}, \hat{\boldsymbol{\theta}}(\boldsymbol{c}))$ can be derived easily and is given by

$$J(\boldsymbol{c}, \hat{\boldsymbol{\theta}}(\boldsymbol{c})) = \operatorname{Arg}\min_{(\boldsymbol{c}_1,\boldsymbol{c}_2)} \left\{ \frac{N_1}{2} \ln \hat{\rho}_1 + \frac{N_2}{2} \ln \hat{\rho}_2 \right\}$$

where $\hat{\rho}_1$ and $\hat{\rho}_2$ are the estimated model error variances of the two classes, and N_1 and N_2 are the numbers of pixels in the two classes. The algorithm starts with an arbitrary partition of the data and computes the corresponding class parameters. With 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.

2.3. Vehicle tracking

The first step for vehicle tracking is to extract the segments in each class from each frame (the SPCPE algorithm provides this functionality). Then the minimum bounding rectangle (MBR) and the centroid point for each segment (vehicle object) are obtained. The next step for vehicle tracking is to connect the related segments in successive frames. The idea is to connect two segments that are spatially the closest in the adjacent frames (Sista and Kashyap, 2000). In other words, the Euclidean distances between the centroids of the segments in the adjacent frames are used as the criteria to track the related segments. In addition, size restrictions are employed to determine the related segments in the successive frames.

Two cases exist when identifying the inter frame relationship between individual objects whose MBRs become overlapped in succeeding frames: overlapped objects of similar sizes (two cars), and overlapped objects of dissimilar sizes (cars, trucks). A solution to the first case was proposed in Chen et al. (2000a), leaving us to address the second case here. An example of this condition is illustrated in figure 2 where a school bus and a small car that were detected as two objects in one frame (frame 15), are merged into a new bigger MBR in frame (frame 16).

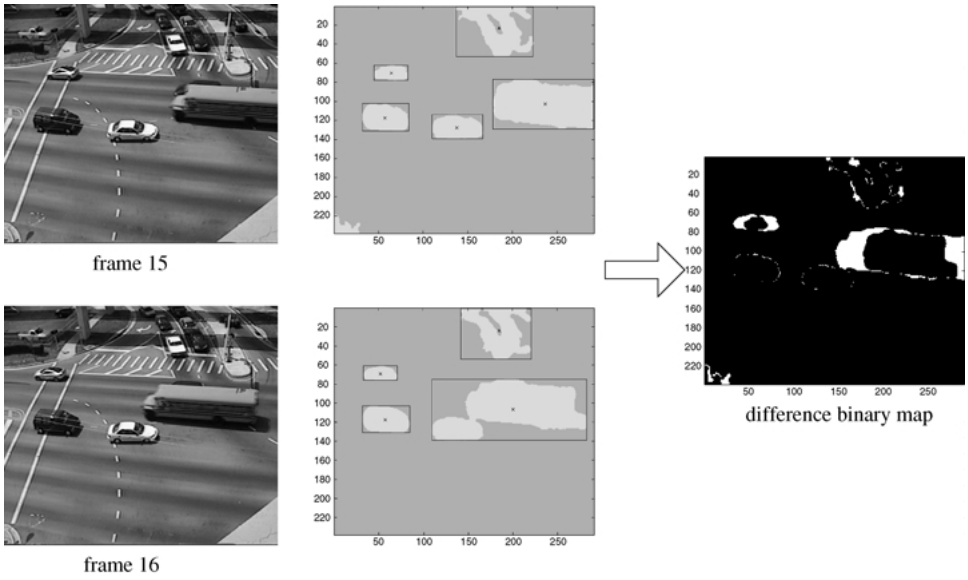


Figure 2. Vehicle tracking.

In our application, given that objects are constrained in their motions (objects move only in a limited number of directions), and that we have information regarding the previous state of the objects before being overlapped (and post state for that matter), we can determine the identity of the objects in the '*overlapping*' segment (MBR). A difference binary map knowledge discovery method is proposed to discover the identity of the objects contained in the '*overlapping*' segments.

The idea is to obtain the difference binary map between the pre and post overlapping frames by subtracting the segmentation result of the overlapped frame from that of the non-overlapped frame and comparing the amount of difference. As shown in the difference binary map in figure 2, the white areas in the difference binary map indicate the amount of difference between the segmentation results of the two consecutive frames. The car and school bus objects in frame 15 can be roughly mapped into the area of the big segment in frame 16 with relatively minor differences. Hence, we can discover the vehicle objects in the big segment in frame 16 by reasoning that it is most probably related to the car and school bus objects from frame 15. In such a case, for the big segment (the '*overlapping*' segment) in frame 16, the corresponding links to the car and bus objects in frame 15 will be created. The approach is symmetric in that it can determine the identity of objects that start out overlapped and in later frames separate.

2.4. Modeling video key frames via MATNs and multimedia input strings

A multimedia augmented transition network (MATN) model can be represented diagrammatically by a labeled directed graph, called a *transition graph*. A multimedia input string is

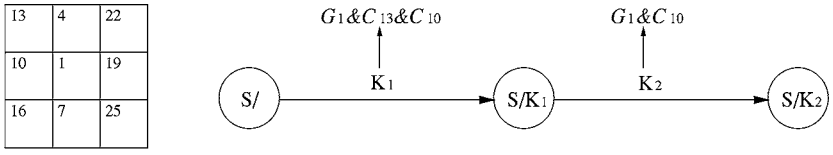
accepted by the grammar if there is a path of transitions which corresponds to the sequence of symbols in the string and which leads from a specified initial state to one of a set of specified final states.

A MATN can build up a video hierarchy (Chen et al., 1999). A video clip can be divided into *scenes*, a *scene* contains a sequential collection of *shots*, and each *shot* contains some contiguous frames that are at the lowest level in the video hierarchy (Yeo and Yeung, 1997). It is advantageous to use several key frames to represent a shot instead of showing all these frames. Key frames act as the indices for a shot. The key frame selection approach utilized was proposed in Chen et al. (1999) and is 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 selection, but we focus on the number, temporal, and spatial relations of the semantic objects. Therefore, these key frames represent the spatio-temporal changes in each shot. For example, in each shot of a traffic video sequence, a key frame is defined when the vehicles change their positions in the subsequent frames or the number of vehicles appearing changes.

As introduced in Chen and Kashyap (2001), one semantic object is chosen as the target semantic object in each video frame and the minimal bounding rectangle (MBR) concept is used. In order to distinguish the 3-D relative positions, twenty-seven numbers are used (Chen and Kashyap, 2001). These numbers represent the twenty-seven possible spatial relations between an object and its neighbors. In this paper, each frame is divided into nine sub-regions with the corresponding subscript numbers shown in figure 3(a). Each key frame is represented by an input symbol in a multimedia input string and the "&" symbol between two vehicle objects is used to denote that the vehicle objects appear in the same frame. The subscripted numbers are used to distinguish the relative spatial positions of the vehicle objects relative to the target object "ground" (figure 3(a)). For simplicity, two consecutive key frames are used to explain how to construct the multimedia input string and the MATN. The multimedia input string that represents these two key frames is as follows:

$$\underbrace{(G_1 \& C_{13} \& C_{10})}_{K_1}\underbrace{(G_1 \& C_{10})}_{K_2}$$

There are two input symbols, K_1 and K_2. The order of the vehicle objects in an input symbol is based on the relative spatial locations of the vehicle objects in the traffic video frame (from left to right and top to bottom). For example, the first key frame is represented



(a) the nine subregions and their corresponding subscript numbers

(b) an example MATN model

Figure 3. MATN and multimedia input strings for modeling the key frames of traffic video shot S.

by input symbol K_1, where G_1 indicates that G is the target object, C_{13} means that the first car object is on the left of and above G, and C_{10} means that the second car object is on the left of G. For the next key frame, its multimedia input string is almost the same as that of the first key frame except that the car C_{13} that appeared in the first key frame has already left the road intersection in the next key frame. Hence, the number of vehicle objects decreases from two to one. This is an example to show how a multimedia input string can represent the change of the number of semantic (vehicle) objects.

Figure 3(b) is the MATN for the above two key frames of the example traffic video sequence. The starting state name for this MATN is S/. As shown in figure 3(b), there are two arcs with arc labels the same as the two input symbols (K_1 and K_2). The different state nodes in the MATN model the temporal relations of the selected key frames. The multimedia input strings model the relative spatial relations of the vehicle objects.

3. Experiments and discussions

In this paper, the enhanced video segmentation method is applied to three real-life traffic video sequences from different sources under various weather conditions by considering two classes. The first frame is partitioned into two classes using an initial random partition. After obtaining the final partition of the first frame (via the SPCPE algorithm), we compute the partitions of the subsequent frames using the previous partitions as the initial partition parameter for the subsequent segmentation steps (since there is little significant difference between consecutive video frames). The convergence speed of the SPCPE algorithm is increased by using the previous partition results and thus provides support for real-time processing.

In the experiments, the original frames, difference frames, segmentation results, and bounding boxes for a few frames in each video sequence are presented. Those frames shown in the experiment results are the key frames after applying the key frame selection method introduced in Chen et al. (1999). In the experiment results (figures 4–6), the leftmost column gives the original video frames, the second column shows difference images obtained by subtracting the background reference frame from the original frames, the third column shows the vehicle segments extracted from the video frames, and the rightmost column shows the bounding boxes of the vehicle objects.

As only the vehicles are important for our application, we use the rightmost column (the simplified segmentation results) to show the relative spatial relationships of the vehicle segments for each key frame in each video sequence. Also, symbolic representations are used to represent the spatial relationships of the vehicle objects in each frame. In the proposed symbolic representation, each vehicle segment is indexed in a multimedia input string based on the spatial relation of its centroid. The subscript numbers are used to denote the relative spatial relations of the vehicle objects with respect to the target object from the viewer's perspective. We observe that the two-class partitioning schema can capture most of the relevant scene information (in regard to traffic applications). In other words, one class captures relevant vehicle information and the second class captures most of the ground information (the background non-vehicle information). Hence, in the multimedia input strings, the ground (G) is selected as the target object and the segments are denoted

by C for the *cars*, B for the *buses*, W for the *waiting* segments, and O for the *overlapping* segments.

3.1. Experiment 1

3.1.1. Experiment setup. The first traffic video sequence was captured with a *Sony Handycam CCD TR64* and digitized with a simple *Brooktree Bt848* based capture card on a *Windows NT 2000 Celeron* based platform. The video sequence consists of about 16 minutes of video from a traffic intersection with approximately constant lighting conditions. The original video frames were of size 480 rows $\times$ 640 columns, 24 bit color and frame rate sampled at 5 frames per second. For simplicity and real-time processing purpose, we transform the color video frames to grayscale images and resize them to half of the original size (240 rows $\times$ 320 columns). The traffic video sequence shows the traffic flow of an intersection on *US*1, one of the busiest state roads in Miami, FL, USA.

3.1.2. Experiment results. The experiment results for frames 4, 9, 15, 16 and 35 are shown in figure 4. As mentioned earlier, in the leftmost column (figure 4(a)) are the original frames. The second column (figure 4(b)) shows the difference images after background subtraction. The final segmentation results are shown in the third column (figure 4(c)). As can be seen from figure 4(c), almost all of the vehicle objects are captured as separate segments (objects) except for those vehicles in the two lanes located in the upper part of the video frame (which has been captured as one segment because they are too close together due to the shooting angle of the camera). Some of the vehicles have been combined with other objects into a single segment when they are closely located. For example, in frame 16, the school bus is overlapped with the car that was waiting in the middle of the intersection while the school bus was moving westbound. Other cars in the main area of the intersection are successfully identified in all of these frames.

In the multimedia input strings (as shown in figure 4(d)), the ground (G) is selected as the target object. For those cars combined together into a single segment (in the upper part of video frame), we use domain knowledge that there are two lanes located in the upper part of the scene where the vehicles are waiting before they enter the intersection. We use the symbol W for this special segment indicating that this is a "*waiting*" segment that may include more than one vehicle waiting to enter the intersection. Our data also contains vehicle objects in the main area of intersection that are combined into one segment. For example, the car object and the school bus object were combined into one segment in frame 16, while they were separate segments in the preceding frame (frame 15). As discussed in Section 2.3, this occlusion situation can be detected by the proposed difference binary map knowledge discovery method. We use symbol O to denote an "*overlapping*" segment which has corresponding links to the related segments in the preceding frame.

As can be seen from figure 4, the "*waiting*" segment always remains at the same location in the scene. In order to answer the query for traffic flow estimation, these "*waiting*" segments will not be counted. As mentioned earlier, G_1 indicates that the ground (G) is the target object and the subscript numbers have the same relative spatial meanings. In frames 4 and 9, two cars in the middle of the intersection (C_{10} and C_1) were waiting to pass while another car

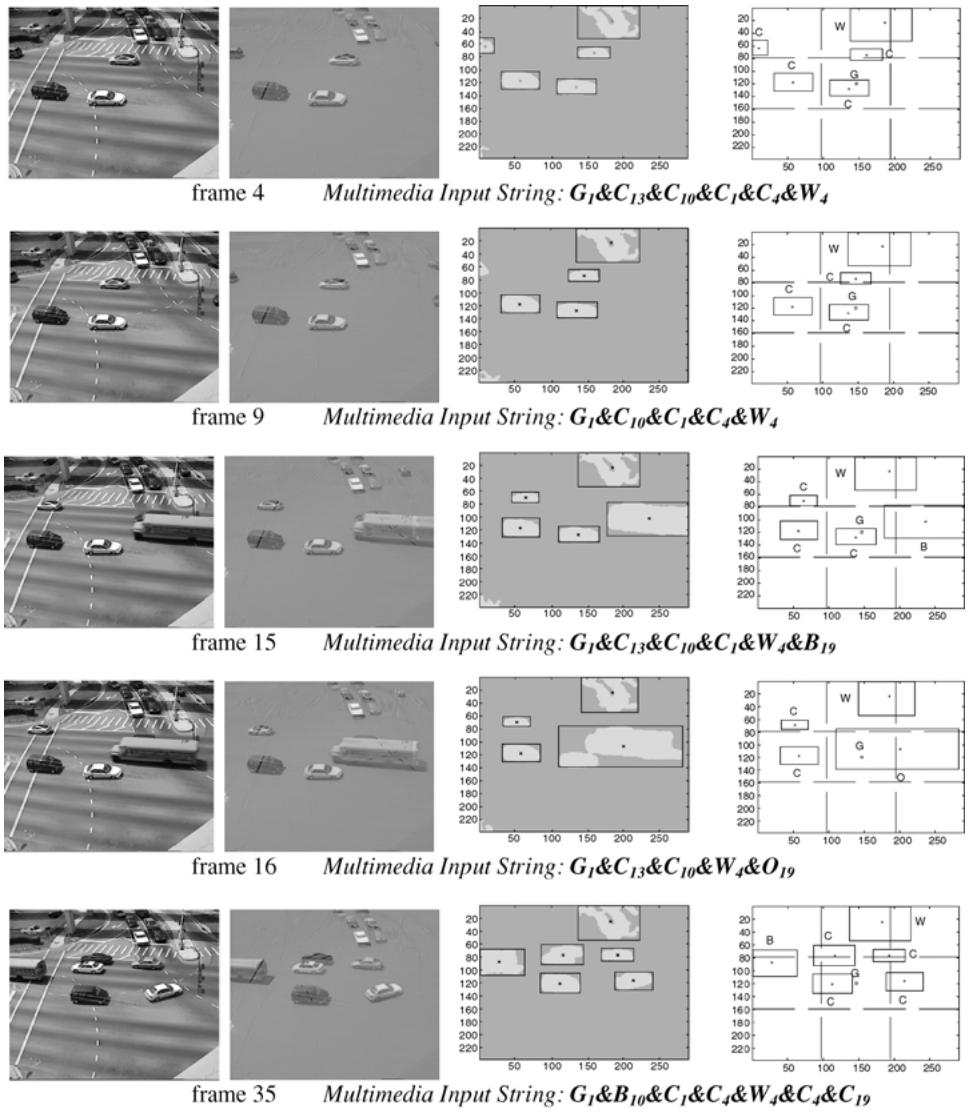


Figure 4. Segmentation results as well as the multimedia input strings for frames 4, 9, 15, 16 and 35 in the first traffic video. The leftmost column gives the original video frames; the second column shows difference images obtained by subtracting the background reference frame from the original frames; the third column shows the vehicle segments extracted from the video frames, and the rightmost column shows the bounding boxes of the vehicle objects.

(C_4) was driving slowly through the upper part of the intersection westbound. In addition, the car (C_{13} in frame 4) was leaving the intersection westbound. In frame 15, a school bus appeared as B_{19} from the east side. However, in frame 16, the school bus and the white car (C_1 in frame 15) were combined into one overlapping segment (O_{19}). In frame 35, the school bus (B_{10}) was separated from the other cars and left the intersection on the west side, while the two cars (C_{10} and C_1) in frames 4, 9 and 15 made the left turn and moved towards the northeast bound so that their relative spatial locations changed to C_1 and C_{19} in frame 35.

3.2. Experiment 2

3.2.1. Experiment setup. The second traffic video sequence was downloaded from the research website of University Karlsruhe (IRA), which includes 1,733 grayvalue frames in GIF format. This traffic sequence was recorded by a stationary camera under normal weather condition. The original size of each frame is 740×560. In our experiment, since our focus is on the traffic intersection, each original image was cut into a 353×473 piece which excludes the trees and buildings since they are not related to the traffic flow monitoring purpose.

3.2.2. Experiment results. One observation on this video sequence is that its visual quality is significantly worse than that of the video sequence in Experiment 1. That is, the objects (e.g., car, pedestrian, etc.) are not as sharp as in the first traffic video sequence. However, as can be seen from figure 5, the testing results are still promising. In frames 3 and 80, all the three vehicles are successfully captured as separate segments. In frame 3, the relative spatial locations of the three vehicles are $C_{13}\&C_{16}\&C_{13}$, while in frame 80, their relative spatial locations change to $C_{13}\&C_{16}\&C_4$. In frame 242, another vehicle (C_{13}) drives into the intersection area from the northwest corner, and the two vehicles stopped side by side behind the crosswalk have been captured as one "*overlapping*" object (O_4). In frame 280, a dark-gray car (C_4) joins the intersection area driving towards northwest. Note that the pedestrian objects have also been successfully identified during segmentation, which is useful for human object tracking. However, we filter them out in this experiment in order to emphasize the vehicle monitoring purpose.

3.3. Experiment 3

3.3.1. Experiment setup. In order to test the robustness of our framework, a third traffic video sequence consisting of 300 frames was used. This video sequence was also downloaded from the research website of University Karlsruhe (IRA) and shows the same traffic intersection as in the second video sequence but with a different weather condition. As shown in figure 6, this video was taken in winter where there was snow on the lanes and light snow was falling with a strong wind. The original size of each frame is 768×576; while in our experiment, the segmentation procedure is only applied to the area with size 268×251, which is the traffic intersection area of our experimental interest. As mentioned

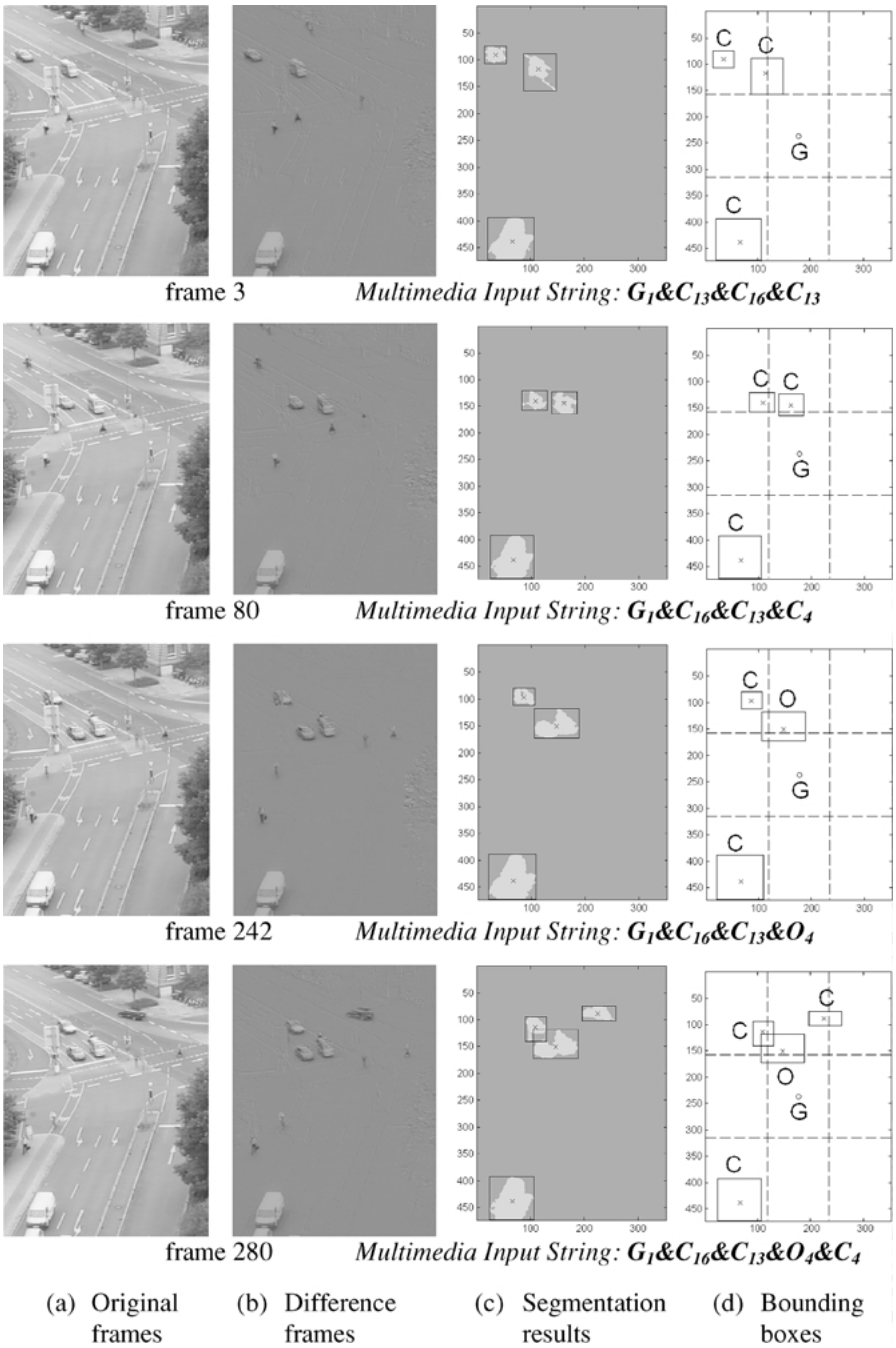


Figure 5. Segmentation results as well as the multimedia input strings for frames 3, 80, 242 and 280 in the second traffic video.

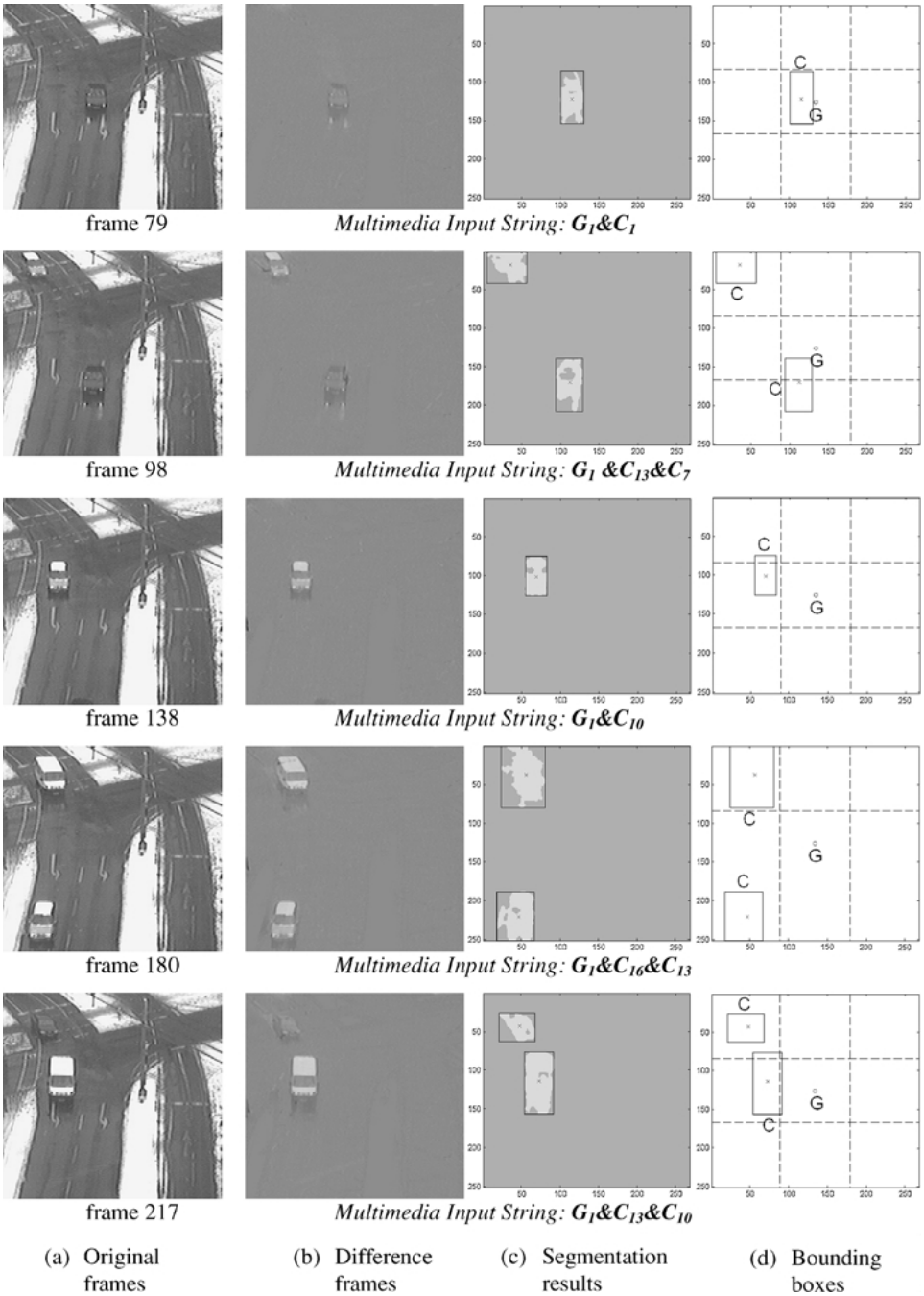


Figure 6. Segmentation results as well as the multimedia input strings for frames 79, 98, 138, 180 and 217 in the third traffic video.

earlier, since our purpose is for vehicle monitoring, only the traffic intersection area is considered.

3.3.2. Experiment results. Figure 6 shows the experiment results for frames 79, 98, 138, 180 and 217. As can be seen from figure 6, the background noise caused by the falling snow is significant and the quality of the video sequence becomes worse because of the bad weather. However, the experiment results are still satisfactory since all the vehicle objects have been captured as separate segments. Notice that the dark-gray car in frame 217 has also been successfully identified though this vehicle object is very obscure in the corresponding difference image (figure 6(b)) because of the similar color it has with the dark-gray background.

3.4. Discussions

The experiment results demonstrate the knowledge discovery process (i.e., the spatio-temporal vehicle tracking) from the traffic video sequences using the proposed multimedia data mining framework. The discovered information can be applied to traffic applications so that spatio-temporal queries can be answered. In addition, as can be seen from the experiment results, the backgrounds of the traffic video sequences are complex. Though related work has been done on the basis of highway traffic video sequences (Friedman and Russell, 1997; Huang et al., 1994), most of this work have had relatively simple backgrounds. Our framework, however, can deal with more complex situations such as the traffic video for intersection monitoring.

From the experiment results, it can be seen that the multimedia input strings can model not only the number of objects, but also the relative spatial relations. In this case, in order to estimate the intersection traffic flow, we can choose the east or west side of the intersection as a "judge line" in the frame to determine the traffic flow of the specified direction (east $\leftrightarrow$ west), and any vehicles passing through that line will be recorded. Using the information of centroid's position of each object, the traffic flow of a specified direction in the intersection area can be determined. Moreover, since the types of vehicles are also important for estimating the traffic flow, the sizes of the bounding boxes can be utilized to determine the vehicle types (such as 'car' or 'bus'). For those "*overlapping*" segments, since they have links to some specific vehicle segments, the corresponding number and types of vehicles in an overlapping segment can be obtained in order to count the traffic flow. Besides answering the traffic flow query, the proposed framework also has the potential to answer other spatio-temporal related database queries.

4. Conclusions

The paper presents work on extracting and modeling the spatio-temporal relationships of vehicles from traffic video sequences. We presented three examples, under various lighting and weather conditions, of extracting representative key frames from videos of traffic intersections and modeling the spatial and temporal relationships of the vehicles in those key frames. The discovered spatio-temporal relationships of the vehicle objects were modeled

by the multimedia augmented transition network (MATN) model and the multimedia input strings. Vehicle object extraction was accomplished using an unsupervised segmentation algorithm. In order to eliminate aspects of the complex background in the traffic video frames, background subtraction techniques were employed. Using the background subtraction technique, both the efficiency of the segmentation process and the accuracy of the segmentation results were improved achieving more accurate video indexing and annotation. The discovered information can be applied to traffic applications so that spatio-temporal related database queries can be answered.

Much work needs to be done to automatically collect, index, store and analyze spatio-temporal information from real-time multimedia streams. The paper discusses an initial framework for the development of such systems. We intend to extend the present work in the following directions: (1) Integrate the methods of information indexing, analyzing and retrieving into our framework; (2) Extend our framework from daytime operations to nighttime domain.

Acknowledgments

For Shu-Ching Chen, this research was supported in part by NSF CDA-9711582.

Note

1. The model error is $e_{ij} = y_{ij} - (a_{k0} + a_{k1}i + a_{k2}j + a_{k3}ij)$.

References

Caltrans. Caltrans Live Traffic Cameras, http://video.dot.ca.gov/.

Chen, S.-C. and Kashyap, R.L. (2001). A Spatio-Temporal Semantic Model for Multimedia Database Systems and Multimedia Information Systems. *IEEE Trans. on Knowledge and Data Engineering*, 13(4), 607–622.

Chen, S.-C., Shyu, M.-L., and Zhang, C.C. (2001). An Intelligent Framework for Spatio-Temporal Vehicle Tracking. In *The 4th International IEEE Conference on Intelligent Transportation Systems*, Oakland, CA, USA (pp. 213–218).

Chen, S.-C., Shyu, M.-L., Zhang, C.C., and Kashyap, R.L. (2000a). Object Tracking and Augmented Transition Network for Video Indexing and Modeling. In *The 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2000)*, Vancouver, British Columbia, Canada (pp. 428–435).

Chen, S.-C., Sista, S., Shyu, M.-L., and Kashyap, R.L. (1999). Augmented Transition Networks as Video Browsing Models for Multimedia Databases and Multimedia Information Systems. In *The 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '99)*, Chicago, IL, U.S.A. (pp. 175–182).

Chen, S.-C., Sista, S., Shyu, M.-L., and Kashyap, R.L. (2000b). An Indexing and Searching Structure for Multimedia Database Systems. In *IS&T/SPIE conference on Storage and Retrieval for Media Databases 2000*, San Jose, CA, U.S.A. (pp. 262–270).

Courtney, J.D. (1997). Automatic Video Indexing via Object Motion Analysis. *Pattern Recognition*, 30(4), 607–625.

Cucchiara, R., Piccardi, M., and Mello, P. (2000). Image Analysis and Rule-based Reasoning for a Traffic Monitoring System. *IEEE Trans. on Intelligent Transportation Systems*, 1(2), 119–130.

Dailey, D.J., Cathey, F., and Pumrin, S. (2000). An Algorithm to Estimate Mean Traffic Speed Using Uncalibrated Cameras. *IEEE Transactions on Intelligent Transportations Systems*, 1(2), 98–107.

Fan, L. and Sung, K.K. (2000). Model-Based Varying Pose Face Detection and Facial Feature Registration in Video Images. In *The 8th ACM International Conference on Multimedia*, Los Angeles, CA (pp. 295–302).

Ferman, A.M., Guensel, B., and Tekalp, A.M. (1997). Object-Based Indexing of MPEG-4 Compressed Video. In *Proceedings of SPIE: Visual Communications and Image Processing*, San Jose, CA (pp. 953–963).

Friedman, N. and Russell, S. (1997). Image Segmentation in Video Sequences: A Probabilistic Approach. In *Proc. Thirteenth Conf. on Uncertainty in Artificial Intelligence (UAI '97)*, Providence; RI.

Gonzalez, R.C. and Woods, R.E. (1993). *Digital Image Processing*. Reading, MA: Addison-Wesley.

Grimson, W.E.L., Stauffer, C., Romano, R., and Lee, L. (1998). Using Adaptive Tracking to Classify and Monitor Activities in a Site. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Preceding* (pp. 22–31).

Haritaoglu, I., Harwood, D., and Davis, L. (1998). W 4—Who, Where, When, What: A Real-Time System for Detecting and Tracking People. In *IEEE Third International Conference on Face and Gesture Recognition*, Nara, Japan (pp. 222–227).

Haritaoglu, I., Harwood, D., and Davis, L. (2000). A Fast Background Scene Modeling and Maintenance for Outdoor Surveillance. In *The 15th IEEE International Conference on Pattern Recognition: Applications, Robotics Systems and Architectures*, Barcelona, Spain (pp. 179–183).

Huang, T., Koller, D., Malik, J., and Ogasawara, G. (1994). Automatic Symbolic Traffic Scene Analysis Using Belief Networks. In *Proceedings of the AAAI, 12th National Conference on Artificial Intelligence (AAAI '94)*, Seattle, WA (pp. 966–972).

IRA. http://i21www.ira.uka.de/image_sequences/.

Kamijo, S., Matsushita, Y., Ikeuchi, K., and Sakauchi, M. (1999). Automatic Symbolic Traffic Scene Analysis Using Belief Networks. In *IEEE International Conference on Intelligent Transportation Systems*, Tokyo Japan (pp. 703–708).

Montgomery Co. Department of Public Works Transportation. ATMS Video Monitoring System Live Traffic Camera Pictures, http://www.dpwt.com/jpgcap/camintro.html.

Sista, S. and Kashyap, R.L. (2000). Unsupervised Video Segmentation and Object Tracking. *Computers in Industry*, 42(2/3), 127–146.

Stauffer, C. and Grimson, W.E.L. (1999). Adaptive Background Mixture Models for Real-Time Tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 246–252).

Toyama, K., Krumm, J., Brumitt, B., and Meyers, B. (1999). Wallflower: Principles and Practice of Background Maintenance. In *7th International Conference on Computer Vision (ICCV '99)*, Island of Crete (pp. 255–261).

Yeo, B.-L. and Yeung, M.M. (1997). Retrieving and Visualizing Video. *Communications of the ACM*, 40(12), 43–52.