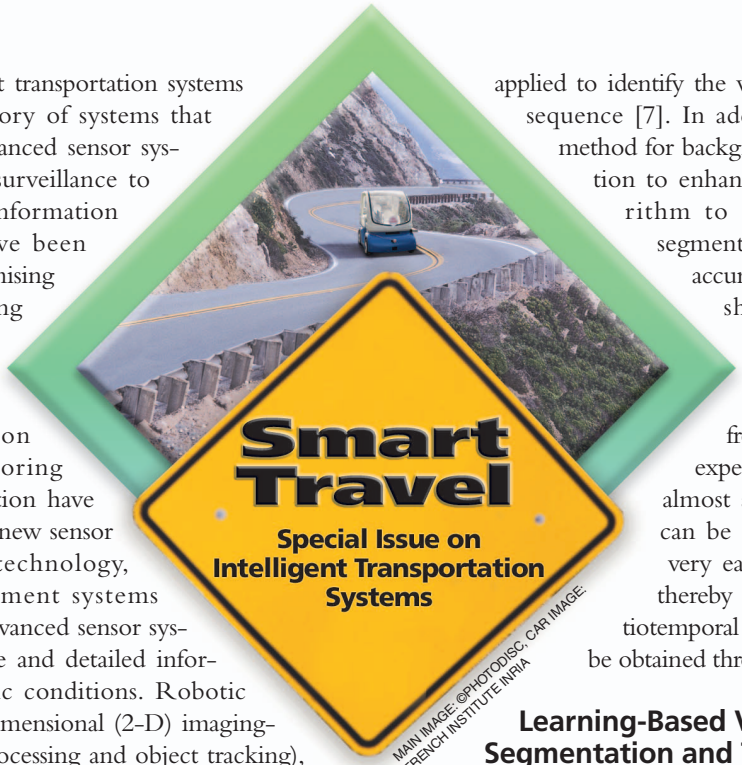
Spatiotemporal Vehicle Tracking

The Use of Unsupervised Learning-Based Segmentation and Object Tracking



Recently, intelligent transportation systems (ITS), one category of systems that make uses of advanced sensor systems for online surveillance to gather detailed information on traffic conditions, have been identified as the most promising way to address the growing mobility problems. Along with the exponential growth in computational capability and information technology, traffic monitoring and large-scale data collection have been enabled by the use of new sensor technologies. One ITS technology, advanced traffic management systems (ATMS) [1], aims to use advanced sensor systems for online surveillance and detailed information gathering on traffic conditions. Robotic vision [2], especially two-dimensional (2-D) imaging-based vision (2-D image processing and object tracking), can be applied to traffic video analysis to address queue detection, vehicle classification, and vehicle counting. In particular, vehicle classification and vehicle tracking have been extensively investigated [3], [4]. Issues associated with extracting traffic movement and accident information from real-time video sequences are discussed in [5].

For traffic intersection monitoring, digital cameras are fixed and installed above the area of the intersection. One classic technique used to identify the moving objects (vehicles) is background subtraction [6]. Various approaches to background subtraction, as well as modeling techniques, have been discussed in the literature [4], [5]. In the proposed framework, an unsupervised video segmentation method called the simultaneous partition and class parameter estimation (SPCPE) algorithm is applied to identify the vehicle objects in the video sequence [7]. In addition, we propose a new method for background learning and subtraction to enhance the basic SPCPE algorithm to generate more accurate segmentation results so that more accurate spatiotemporal relationships of objects can be obtained. Experiments are conducted using real-life traffic video sequences from road intersections. The experimental results indicate that almost all moving vehicle objects can be successfully identified at a very early stage of the processing, thereby ensuring that accurate spatiotemporal information of objects can be obtained through object tracking.

Learning-Based Vehicle-Object Segmentation and Tracking for Traffic Video Sequences

Traffic video analysis at intersections can provide a rich array of useful information, including vehicle identification, queue detection, vehicle classification, traffic volume, and incident detection. The proposed unsupervised spatiotemporal vehicle tracking framework includes background learning and subtraction and vehicle-object identification and tracking.

Background subtraction is a technique used to remove nonmoving components from a video sequence, where a reference frame of the stationary components in the image is created. Once created, the reference frame is subtracted from any subsequent images. The pixels resulting from new (moving) objects will generate a nonzero difference. The main assumption for the application of background subtraction is that the

BY SHU-CHING CHEN, MEI-LING SHYU, SRINIVAS PEETA, AND CHENGCUI ZHANG

1070-9932/05/$20.00©2005 IEEE

camera remains stationary. However, in most cases, the nonsemantic content (or background) in the images or video frames is very complex. Therefore, an effective way to obtain background information can generate better segmentation results.

In the framework under discussion, we propose an adaptive background-learning method. Our method consists of the following steps (as illustrated in Figure 1):

1) Subtract the successive frames to get the motion difference images.
2) Apply segmentation on the difference images to get the estimation of foreground regions and background regions.
3) Generate the current background image based on the learned information seen so far.
4) Perform background subtraction and object segmentation for those frames that contribute to the generation of the current background. Meanwhile, the extracted vehicle objects are tracked from frame to frame. Upon finishing the current processing, go back to Steps 1–3 to generate the next background image. Repeat this process until all the frames have been segmented.

While the proposed segmentation method can identify vehicle objects, it does not differentiate between them (cars, buses, etc.). Therefore, a priori knowledge (size or length) of different vehicle classes should be provided to enable such classification. In addition, since the vehicle objects of interest are the moving ones, stopped vehicles will be considered as static objects and will not be identified as mobile objects until they start moving again. However, the object tracking technique ensures that such vehicles are seamlessly tracked even though they "disappear" for some duration due to the background subtraction. This aspect is especially critical under congested or queued traffic conditions.

In a traffic video monitoring sequence, when a vehicle object stops in the intersection area (including the vehicle approaching the intersection), our framework may deem it part of the background information. Since the vehicle objects move into the intersection area before stopping, they are identified as moving vehicles before they stop due to the characteristics of our framework. Hence, their centroids identified before they stop will be in the intersection area. For these vehicles, the tracking process is frozen until they begin moving again; they are identified as "waiting" objects rather than "disappearing" objects. That is, the tracking process will follow the same procedure as before unless one or more new objects abruptly appear in the intersection area. Then, the matching and tracking of the previous "waiting" objects will be triggered to continue tracking the trails of these vehicles.

The Unsupervised Video Segmentation Method (SPCPE)

The SPCPE algorithm is an unsupervised image segmentation method used to partition video frames [7]. A given class description determines a partition and vice versa. Therefore, the partition and the class parameter 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; we do not have a priori knowledge of which pixels belong to which class. In the SPCPE algorithm, the 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 jointly estimate the partition and the class parameters. Since the successive frames in a video do not differ to a great extent, the partitions of adjacent frames do not show significant differences. Each frame is partitioned using the partition of the previous frame as an initial condition to speed up the convergence rate of the algorithm. Usually, the number of iterations needed for convergence is around two to three, except for the first frame for which a randomly generated initial partition is used.

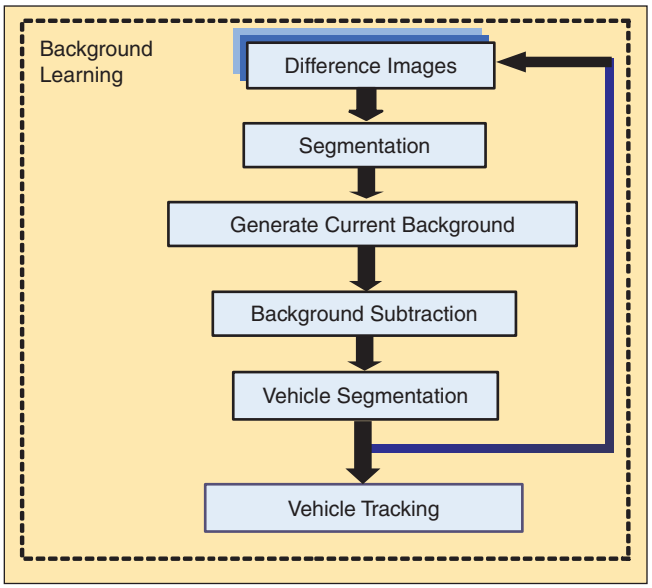


Figure 1. *The basic workflow of the proposed method.*

Background Learning and Extraction

The basic idea of our background-learning method is to generate the current background image based on the segmentation results extracted from a set of frame-to-frame difference images. The method described in [8] is probably closest to our paradigm. In [8], binary thresholding of difference images is used as the segmentation method, and the background image is updated periodically based on the weighted sum of the current binary object mask and the previous background. They also conduct vehicle tracking and vehicle classification based on these methods. Rather than binary thresholding, we use the SPCPE algorithm to segment a difference image into a two-class segmentation map which can serve as a foreground/background mask, where "class one" includes the background points and "class two" records the foreground points. By simply collecting a small portion of the continuous segmentation maps, we can reach a point where every pixel position has at least one segmentation

map with its corresponding pixel value equal to one (background pixel). In other words, every background point within this time interval has appeared and been identified at least once in the segmentation maps.

Figure 2 illustrates the process for background learning using an image sequence from Frame 1,121 to Frame 1,228. As shown in Figure 2(b), the difference images are computed by subtracting successive frames and applying linear

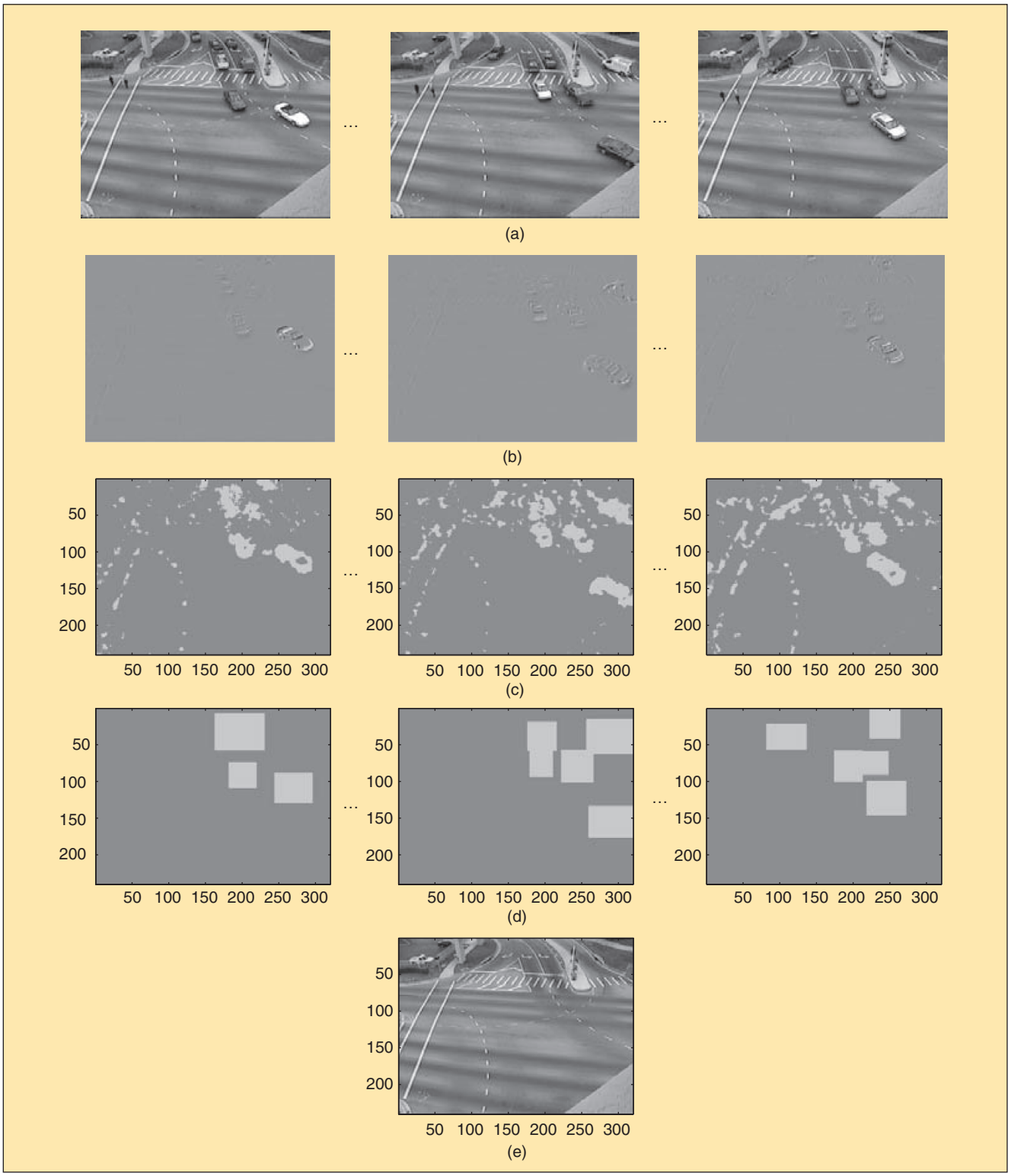


Figure 2. *Unsupervised background learning and subtraction in the traffic video sequence: (a) image sequence from Frame 1,121 to Frame 1,228; (b) successive difference (normalized) images for frames in (a); (c) segmentation maps for difference images in (b); (d) rectified segmentation maps for difference images in (c); and (e) the generated background for this sequence.*

normalization. Then, these difference images are segmented into two-class segmentation maps as in Figure 2(c) using SPCPE. This is followed by a rectification procedure that can eliminate most of the noise and store the background information robustly. Figure 2(d) shows such rectified segmentation maps for Figure 2(c). The rectified segmentation maps are then used to generate a background image if the specified condition is satisfied. The extraction of the background information is completed by taking the corresponding background pixels from individual frames within this time interval. Instead of simply averaging these background pixel values, we further analyzed its histogram distribution and picked the values in dominant bin(s) as the trusted values. With this extra sophistication, the false positives in the background image caused by noise or misdetected motions can be reduced significantly. Two such examples are shown in Figure 3; Figure 3(a) contains two constructed background images with lots of false positives, and Figure 3(b) shows the improved results based on the principles described above. As seen in this figure, the two generated background images contain some static vehicle objects, such as the gray car waiting in the middle of the intersection and the cars that stopped behind the zebra crossing. Once they begin to move, a new background image is created to reflect the motion changes in these objects.

The key point here is that it is not necessary to obtain a perfectly "clean" background image for each time interval. In fact, including the static objects as part of the background will not affect the extraction of the moving objects. Furthermore, once the static objects begin to move, the underlying background can be discovered automatically. Instead of finding one "general background image," the proposed background-learning method aims to provide periodical background images based on motion difference and robust image segmentation. In this manner, it is insensitive to illumination changes and does not require any human efforts in the loop.

Vehicle-Object Tracking

In order to index the vehicle objects, the proposed framework must have the ability to track the moving vehicle objects (segments) within successive video frames, allowing it to provide useful and accurate traffic information for ATMS.

Traffic video analysis at intersections can provide a rich array of useful information, including vehicle identification and queue detection.

After video segmentation, the segments (objects), with their bounding boxes and centroids, are extracted from each frame. Intuitively, two segments that are spatially the closest in the adjacent frames are connected. Euclidean distance is used to measure the distance between their centroids for vehicle tracking. However, it is still necessary to handle the occlusion situations in vehicle tracking.

A more sophisticated object-tracking algorithm integrated in the proposed framework is given in [9]. It can handle the situation of two objects overlapping under certain assumptions (e.g., the two overlapped objects should have similar sizes). In this case, if two overlapped objects of similar size have ever separated from each other in the video sequence, they can be split and identified as two objects with their bounding boxes fully recovered using the object-tracking algorithm. The results are demonstrated in Figure 4(a)–(d), where two vehicles experience some overlapping in Frames 138 and 142 but are identified as two separate objects in Frame 132. Figure 4(d) demonstrates the final

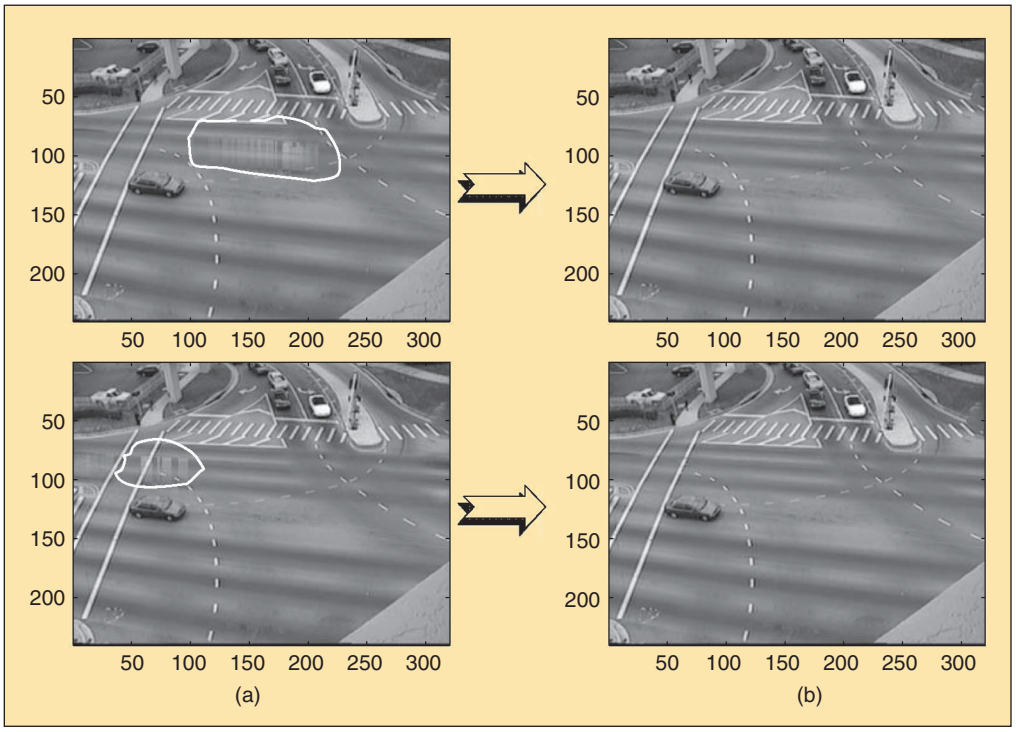


Figure 3. *Improved background extraction: (a) two generated background images with false positives marked by white circles and (b) the improved results.*

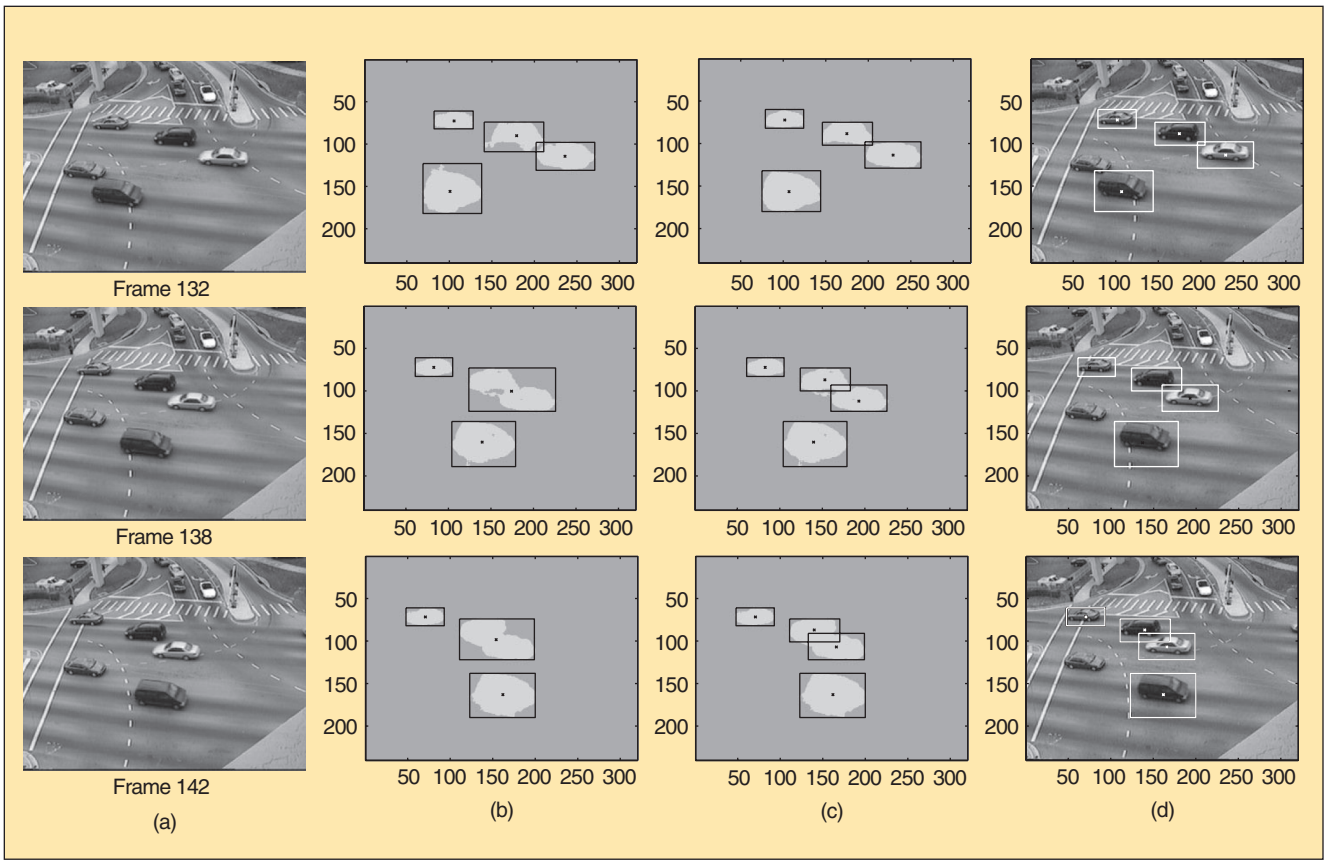


Figure 4. *Handling two-object occlusion in object tracking: (a) video Frames 132, 138, and 142; (b) segmentation maps for frames in (a) without occlusion handling; (c) results by applying occlusion handling; (d) the final results produced by overlaying the bounding boxes in (c) to frames in (a).*

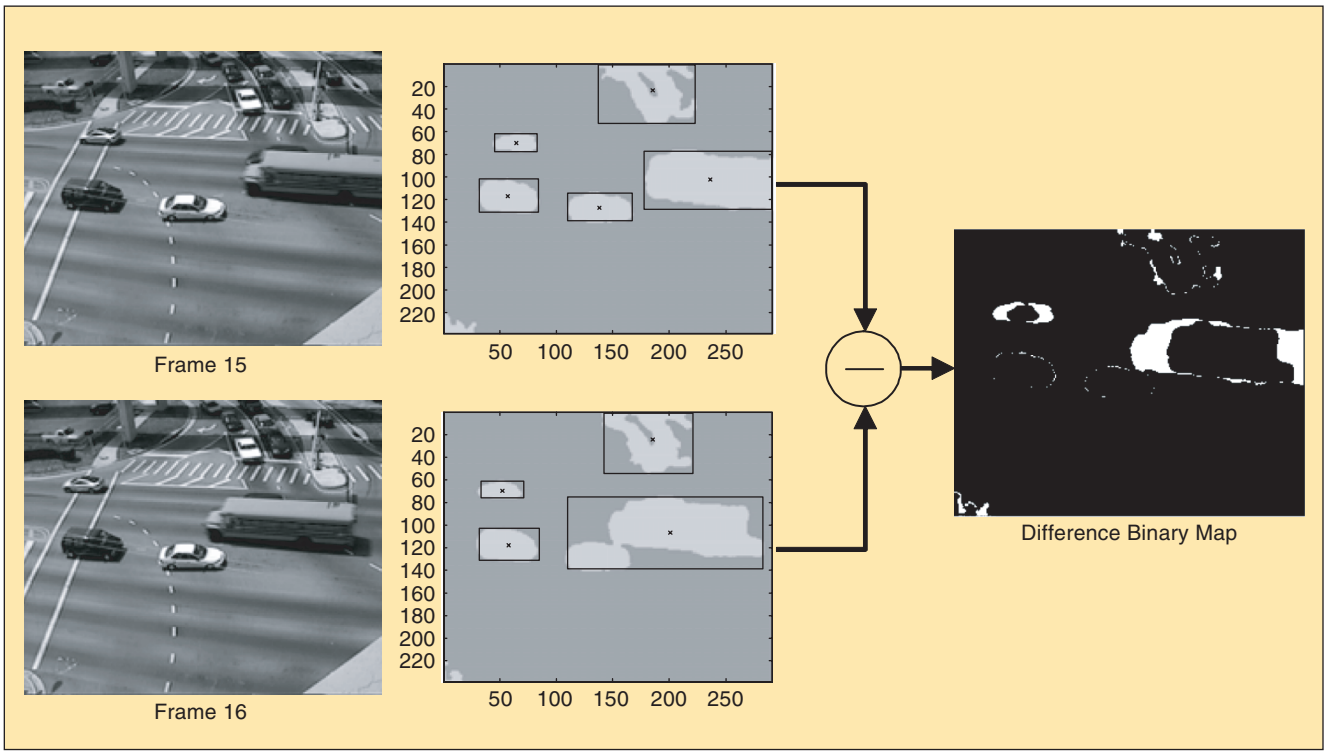


Figure 5. *Handling object occlusion in object tracking.*

results by applying the occlusion handling method proposed in [9]. The segmentation results accurately identify all the vehicle objects' bounding boxes and centroids.

However, there are cases where a large object overlaps a small one. For example, in Figure 5, the large bus merges with the small white car to form a new, big segment in Frame 16, although they are two separate segments in Frame 15. In this scenario, the car object and the bus object that were separate in Frame 15 cannot find their corresponding segments in Frame 16 by centroid-matching and size restriction. However, from the new, big segment in Frame 16, we can reason that this is an overlapping segment that actually includes more than one vehicle object. For this purpose, a difference-binary-map reasoning method is proposed in this article to identify which objects the overlapping segment may include. The idea is to obtain the difference binary map by subtracting the segment result of Frame 15 from that of Frame 16 and then to check the amount of difference between the segmentation results of the consecutive frames. As shown in the difference binary map in Figure 5, the white areas indicate the amount of difference between the segmentation results of the two consecutive frames. Consequently, the car and bus objects in Frame 15 can be roughly mapped into the area of the big segment in Frame 16 with relatively small differences. And, the vehicle objects in the big segment in Frame 16 can be obtained by reasoning that this segment is most probably related to the car

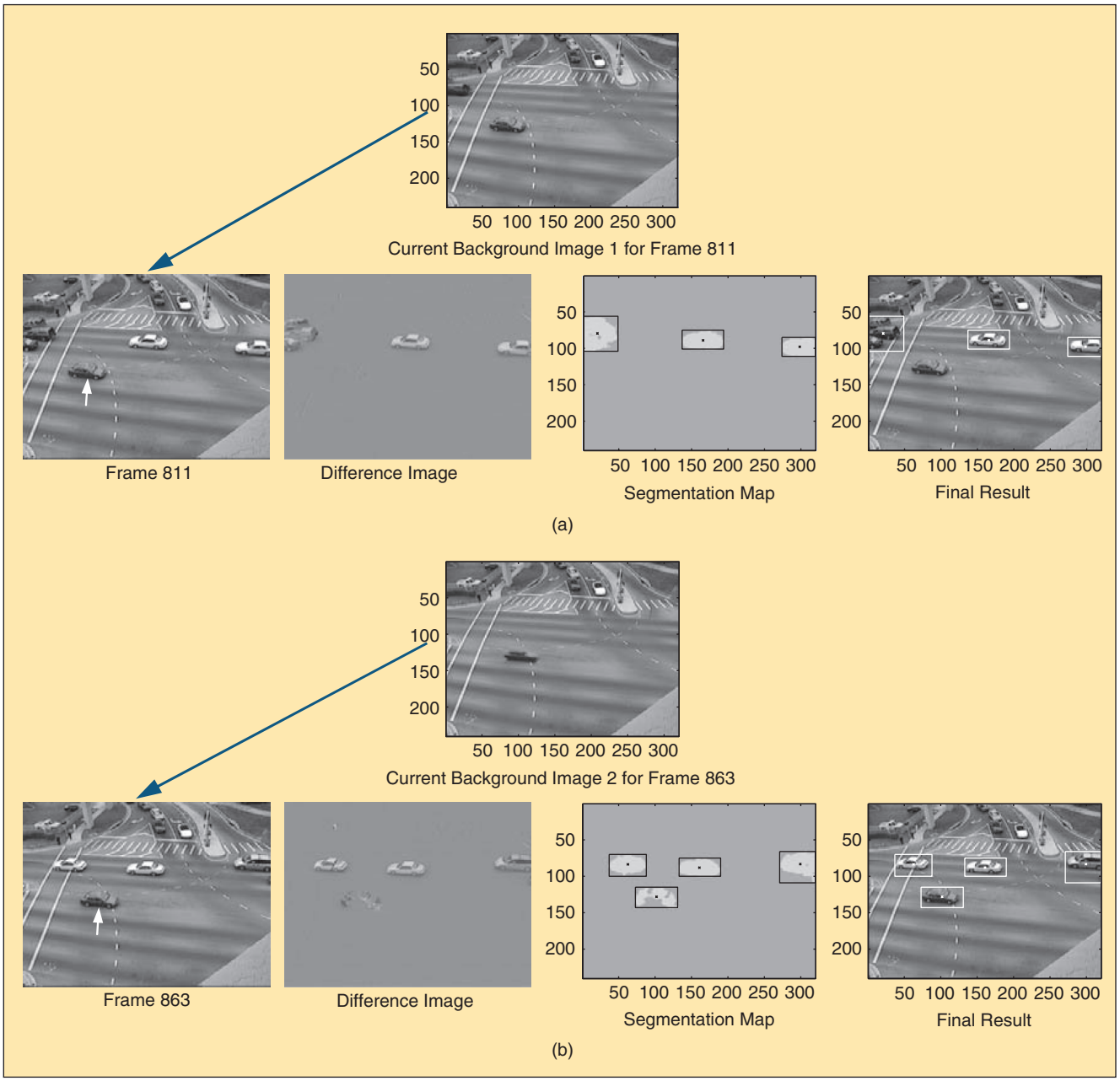


Figure 6. *Segmentation results for Video Sequence 1: (a) segmentation result for Frame 811 using background Image 1; (b) segmentation result for Frame 863 using background Image 2.*

The proposed unsupervised spatiotemporal vehicle tracking framework includes background learning and subtraction and vehicle-object identification and tracking.

and bus objects in Frame 15. Therefore, for the big segment (the overlapping segment) in Frame 16, the corresponding links to the car and bus objects in Frame 15 can be created. This means that the relative motion vectors of that big segment in the following frames will be automatically appended to the trace tubes of the bus and car objects in Frame 15.

Experimental Analysis

Two real-life traffic video sequences (one taken by us and the other downloaded from the Web site of KOGS/IAKS Universitat Karlsruhe [10]) are used to analyze spatiotemporal vehicle tracking utilizing the proposed learning-based vehicle tracking framework. We label these two video sequences as Video Sequence 1 and Video Sequence 2, respectively. Both are grayscale videos that show the traffic flows on two different road intersections for some time duration. The background information can be very complex due to road pavement, trees, zebra crossings, pavement markings/signage, and ground. The proposed new framework is fully unsupervised, meaning that it can enable the automatic background-learning process to facilitate the unsupervised vehicle segmentation process without any human intervention. During the segmentation, the first frame is partitioned into two classes using random initial

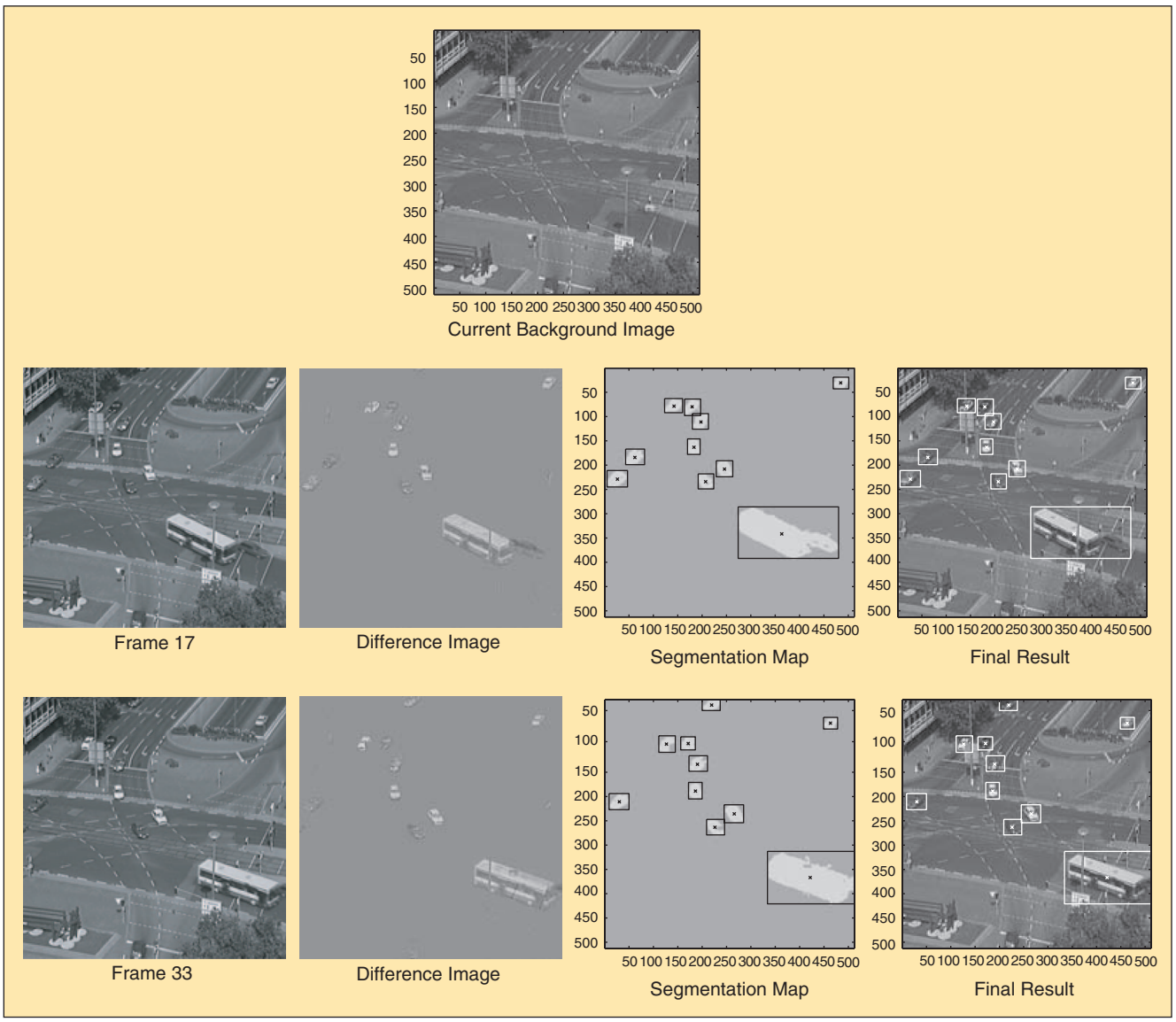


Figure 7. *Segmentation results for Video Sequence 2.*

partitions. After obtaining the partition of the first frame, the partitions of the subsequent frames are computed using the previous partitions as the initial partitions because there is no significant difference between consecutive frames. By doing so, the segmentation process will converge quickly, thereby providing support for real-time processing.

To demonstrate the effectiveness of the proposed background-learning process, the algorithm is coded in C++ and run on a 3.06-GHz Pentium 4 PC with 1-GB RAM. The total CPU time for processing 2,940 JPEG frames (each 240 by 320 pixels) is about 1,639 s (i.e., less than 0.56 s/frame). This performance ensures that a long run is not necessary to fully determine accurate background information. In our experiments, the current background information can usually be obtained within 20~100 consecutive frames, and it is good enough for the future segmentation process. In fact, by combining the background-learning process with the unsupervised segmentation method, our framework can enable the adaptive learning of background information.

Figure 6 shows the segmentation results for a few frames (811 and 863) along with the original frames, the background images, and the difference images. As shown in Figure 6(a), the background image for Frame 811 contains some static vehicle objects such as the gray car waiting in the middle of the intersection and the cars that stopped behind the zebra crossing. Since they are identified as part of the background in this time interval because they lack motion, they will not be extracted by the segmentation process, as shown in Figure 6(a). However, the gray car (marked by a white arrow in the original frames) that was previously waiting in the middle begins to move at around Frame 860, triggering the generation of a new current background, which is shown in Figure 6(b) and labeled as Image 2. From Figure 6(b), it is obvious that the gray car is fading, although not completely, from the background image. However, this fading is sufficient enough to result in the identification of the gray car in Frame 863, as can be seen from the segmentation map and final result in Figure 6(b). Moreover, as shown in Frame 863 in Figure 6(b), our method can successfully illustrate the slow-motion effect, unlike many methods that have difficulties dealing with it since it can be easily confused with noise-level information.

Figure 7 shows our experimental results for Video Sequence 2. As illustrated by the figure, the background of this traffic video sequence is very complex. Some vehicle objects (such as the small gray vehicles in the upper left part of Frame 33) can be easily ignored or confused with the road surface and surrounding environment. While there is an existing body of literature that addresses relatively simple backgrounds, our framework can address far more complex situations. The experimental results shown in Figure 7 are very promising because a) the background is perfectly reconstructed by using 38 out of the total 50 consecutive frames, and b) a single class can capture almost all vehicle objects, even those vehicles that look small and obscure in the upper left area of the video frames. Also, the rightmost column in Figure 7 shows that nearly all vehicle objects are identified as separate objects.

Our background-learning method generates the current background image based on the segmentation results extracted from a set of frame-to-frame difference images.

Based on our experimental experience with these two traffic video sequences, we have the following observations:

1) The segmentation method adopted is very robust and insensitive to illumination changes. Also, the underlying class model in the SPCPE method is very suitable for vehicle-object modeling. Unlike some existing frameworks, in which one vehicle object is segmented into several small regions and later regrouped and linked with the corresponding vehicle object, no extra effort is required for such a merge in our method.
2) As described earlier, the long-run look ahead is not necessary to generate a background image in our framework. This implies that the moving objects can be extracted as soon as their motions are detected. Moreover, the background image can be quickly updated and adapted to new changes in the environment.
3) No manual initialization or prior knowledge of the background is needed.

Furthermore, since the position of the centroid of a vehicle is recorded during the segmentation and tracking process, this information can be used in the future for indexing the vehicle's relative spatial relations. The proposed framework has the potential to address a vast range of spatiotemporal-related database queries for ITS. For example, it can be used to reconstruct accidents at intersections in an automated manner to identify causal factors, thereby enhancing safety.

Conclusions

In this article, we present a framework for spatiotemporal vehicle tracking using unsupervised learning-based segmentation and object tracking. An adaptive background-learning and subtraction method is proposed and applied to two real-life traffic video sequences to obtain more accurate spatiotemporal information on the vehicle objects. The proposed background-learning method, paired with the image segmentation, is robust under many conditions. As demonstrated in our experiments, almost all vehicle objects are successfully identified through this framework. A key advantage of the proposed background-learning algorithm is that it is fully automated and unsupervised; it performs the generation of background images

using a self-triggered mechanism. This is very useful in video sequences in which it is difficult to acquire a clean image of the background. Hence, the proposed framework can deal with very complex situations compared to intersection monitoring.

In our future work, we intend to: 1) perform a more comprehensive study under a wider range of conditions; 2) index and store the vehicle tracking information; and 3) fuse different types of media data from video data.

Acknowledgments

For Shu-Ching Chen, this research was supported in part by NSF CDA-9711582, NSF EIA-0220562, and the Office of the Provost/FIU Foundation.

Keywords

Robotic vision, vehicle tracking, video analysis, segmentation, intelligent transportation systems.

References

[1] S. Peeta and H.S. Mahmassani, "Multiple user classes real-time traffic assignment for online operations: A rolling horizon solution framework," *Transport. Res.*, vol. 3, no. 2, pp. 83–98, 1995.

[2] A.C. Kak and G.N. DeSouza, "Robotic vision: What happened to the visions of yesterday?" in *Proc. 2002 Int. Conf. Pattern Recognition*, Quebec, Canada, Aug. 2002, pp. 839–847.

[3] S. Kamijo, Y. Matsushita, and K. Ikeuchi, "Traffic monitoring and accident detection at intersections," *IEEE Trans. Intell. Transport. Syst.*, vol. 1, no. 2, 2000, pp. 108–118.

[4] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 1999, pp. 246–252.

[5] D.J. Dailey, F. Cathey, and S. Pumrin, "An algorithm to estimate mean traffic speed using uncalibrated cameras," *IEEE Trans. Intell. Transport. Syst.*, vol. 1, no. 2, pp. 98–107, June 2000.

[6] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*. Reading, MA: Addison-Wesley, 1993.

[7] S.-C. Chen, S. Sista, M.-L. Shyu, and R.L. Kashyap, "An indexing and searching structure for multimedia database systems," in *Proc. S&T/SPIE Conf. Storage and Retrieval for Media Databases 2000*, San Jose, CA, Jan. 23–28, 2000, pp. 262–270.

[8] S. Gupte, O. Masoud, R.F.K. Martin, and N.P. Papanikolopoulos, "Detection and classification of vehicles," *IEEE Trans. Intell. Transport. Syst.*, vol. 3, no. 1, pp. 37–47, Mar. 2002.

[9] S.-C. Chen, M.-L. Shyu, C. Zhang, and R.L. Kashyap, "Object tracking and augmented transition network for video indexing and modeling," in *Proc. 12th IEEE Int. Conf. Tools with Artificial Intelligence (ICTAI 2000)*, Vancouver, BC, Canada, Nov. 13–15, 2000, pp. 428–435.

[10] Institute of Algorithms and Cognitive Systems [Online]. Available: http:// i21www.ira.uka.de/image_sequences/

Shu-Ching Chen has been an associate professor in the School of Computer Science (SCS), Florida International University (FIU), since August 2004. Previously, he was an assistant professor in SCS at FIU since August 1999. He received his Ph.D. from the School of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana, in December 1998. He also received master's degrees in computer science, electrical engineering, and civil engineering from Purdue University. His main research interests include distributed multimedia database systems, data mining, and intelligent transportation systems. He was awarded the University Outstanding Faculty Research Award from FIU in 2004.

Mei-Ling Shyu received her Ph.D. from the School of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana, in 1999. She also received an M.S. degree in computer science, an M.S. degree in electrical engineering, and an M.S. degree in restaurant, hotel, institutional, and tourism management from Purdue University in 1992, 1995, and 1997, respectively. She has been an assistant professor at the University of Miami's Department of Electrical and Computer Engineering since January 2000.

Srinivas Peeta has been an associate professor of civil engineering at Purdue University since August 2000. Previously, he was an assistant professor in the Transportation and Infrastructure Systems group (1994–2000). His research interests include the use of operations research, control theory, computational intelligence techniques, and sensor technologies to model and evaluate the dynamics of large-scale transportation networks, especially in the context of advanced information systems. His Ph.D. dissertation won the 1994 best dissertation award in the Transportation Science Section of the Institute for Operations Research and Management Science. He received a CAREER award from the National Science Foundation in 1997. He has authored more than 45 papers in archival journals and has refereed conference proceedings.

Chengcui Zhang has been an assistant professor of computer and information science at the University of Alabama at Birmingham (UAB) since August 2004. She received her Ph.D. from the School of Computer Science at Florida International University (FIU), Miami. She also received her baccalaureate and master's degrees in computer science from Zhejiang University in China. Her research interests include multimedia databases, multimedia data mining, and image and video database retrieval. She has been the recipient of several awards, including the UAB ADVANCE Junior Faculty Research Award from the National Science Foundation and the Presidential Fellowship and the Best Graduate Student Research Award at FIU.

Address for Correspondence: Shu-Ching Chen, School of Computer Science, Florida International University, 11200 SW 8th Street, Miami, FL 33199 USA. Phone: +1 305 348 3480. Fax: +1 305 348 3549. E-mail: chens@cs.fiu.edu.