Adaptive Background Learning for Vehicle Detection and Spatio-Temporal Tracking

Chengcui Zhang[1], Shu-Ching Chen[1*], Mei-Ling Shyu[2], Srinivas Peeta[3]
[1]Distributed Multimedia Information System Laboratory
School of Computer Science, Florida International University, Miami, FL 33199, USA
[2]Department of Electrical and Computer Engineering, University of Miami,
Coral Gables, FL 33124, USA
[3]School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA

Abstract

Traffic video analysis can provide a wide range of useful information such as vehicle identification, traffic flow, to traffic planners. In this paper, a framework is proposed to analyze the traffic video sequence using unsupervised vehicle detection and spatio-temporal tracking that includes an image/video segmentation method, a background learning/subtraction method and an object tracking algorithm. A real-life traffic video sequence from a road intersection is used in our study and the experimental results show that our proposed unsupervised framework is effective in vehicle tracking for complex traffic situations.

1. Introduction

The advent of Intelligent Transportation Systems (ITS) technologies has significantly enhanced the ability to provide timely and relevant information to road users through advanced information systems. One ITS technology, Advanced Traffic Management Systems (ATMS) [8], aims at using advanced sensor systems for on-line surveillance and detailed information gathering on traffic conditions. Techniques from computer vision, such as 2-D image processing, 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 [7, 4].

For traffic intersection monitoring, digital cameras are fixed and installed above the area of the intersection. A classic technique to extract the moving objects (vehicles) is background subtraction. Various approaches to background subtraction and modeling techniques have been discussed in the literature [5, 6]. Background subtraction is a technique to remove non-moving 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 non-zero difference. The main assumption for its application is that the camera remains stationary. However, in most cases, the non-semantic content (or background) in the images or video frames is very complex. Therefore, an effective way to obtain the background information can enable better segmentation results.

In the proposed framework, an unsupervised video segmentation method called the Simultaneous Partition and Class Parameter Estimation (SPCPE) algorithm [1] is applied to identify the vehicle objects in the video sequence. In addition, we propose a new method for background learning and subtraction to enhance the basic SPCPE algorithm in order to get more accurate segmentation results, so that the more accurate spatio-temporal relationships of objects can be obtained. Experiments are conducted using a real-life traffic video sequence 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 spatio-temporal information of objects can be obtained through object tracking.

This article is organized as follows. The next section introduces the proposed learning-based object tracking framework. The experiments, results, and the analysis of the proposed multimedia traffic video data indexing framework are discussed in Section 3. A real-life traffic video sequence is used for the experiments. Conclusions are presented in Section 4.

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

The proposed unsupervised spatio-temporal vehicle tracking framework includes background learning and subtraction, vehicle object identification and tracking.

In the proposed framework, an adaptive background learning method is proposed. Our method consists of the following steps:

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.

0-7803-8185-8/03/$17.00 © 2003 IEEE

3. Generate the current background image based on the learned information seen so far.
4. Do 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 until all the frames have been segmented.

The details are described in the following subsections. The proposed segmentation method can identify vehicle objects, but does not differentiate between them (into cars, buses, etc.). Therefore, *a priori* knowledge (size, length, etc.) 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 though they "disappear" for some duration due to the background subtraction. This aspect is especially critical under congested or queued traffic conditions.

2.1 Vehicle Object Segmentation

The SPCPE (Simultaneous Partition and Class Parameter Estimation) algorithm is an unsupervised image segmentation method to segment video frames [1]. A given class description determines a partition, and vice versa. Hence, the partition and the class parameters have to be estimated simultaneously. 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. Since the successive frames in a video do not differ much, the partitions of the adjacent frames do not differ significantly. Each frame is partitioned by using the partition of the previous frame as an initial condition to speed up the convergence rate of the algorithm. Also, we only use two-class segmentation for this specific application domain, where class 1 corresponds to 'background' and class 2 represents 'foreground'.

One difficulty with automatic segmentation is the lack of the intervention of the user: we do not know a priori which pixels belong to which class. In [9], we further advance the SPCPE method by using the initial partition derived by wavelet analysis, which is proven to produce more reliable segmentation results. More technical details can be found in [1, 9].

2.2 Background Learning and Extraction

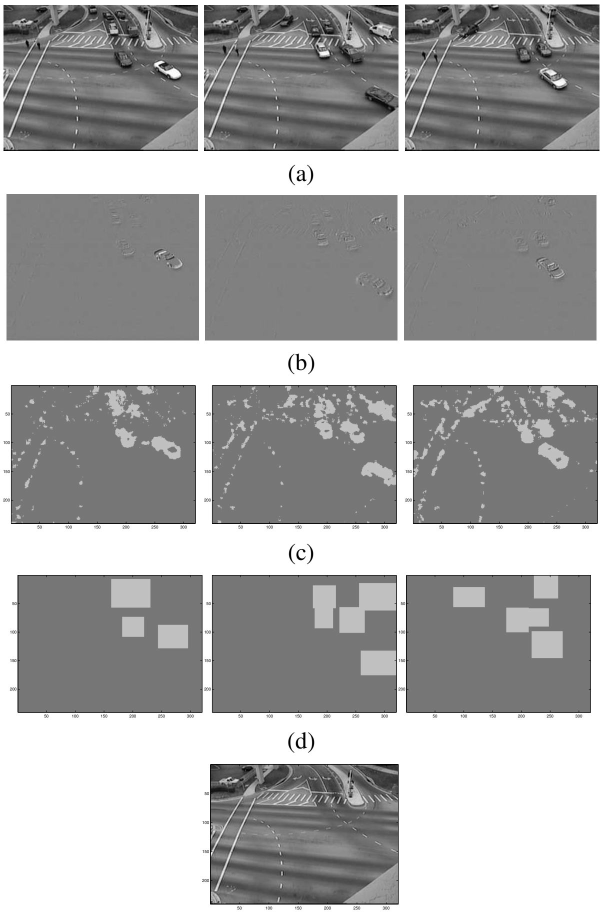
The basic idea of our background learning method is to generate the current background image based on the segmentation results extracted from a set of frame-to-frame difference images. Many existing methods use binary thresholding on difference image as the segmentation method. In some work, a further step is taken to update the background image periodically based on the weighted sum of the current binary object mask and the previous background. In our method, instead of using 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 1 includes the background points and class 2 records the foreground points. By just 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 1 (background pixel). In other words, every background point within this time interval has appeared and been identified at least once in the segmentation maps. Let S_i ~ S_j denote the segmentation maps for difference images i to j; the corresponding background image can be generated if:

$$\forall (r \in row, c \in col) \exists S_k [(S_k[r,c] = 1) \wedge (i \leq k \leq j)]$$

where *row* is the number of rows and *col* is the number of columns for a video frame.

Figure 1 illustrates the process for background learning using an image sequence from frame 1121 to frame 1228. As shown in Figure 1(b), the difference images are computed by subtracting successive frames and applying linear normalization. Then, these difference images are segmented into 2-class segmentation maps (Figure 1(c)) using advanced SPCPE, followed by a rectification procedure that can eliminate most of the noise and keep the background information robustly. Figure 1(d) shows such rectified segmentation maps for Figure 1(c). The rectified segmentation maps are then used to generate a background image (as shown in Figure 1(e))if the condition specified in Equation (6) is satisfied. The step to extract the background information is done 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 background image due to noise or miss-detected motions can be reduced significantly.

In a traffic video monitoring sequence, when a vehicle object stops in the intersection area (including the approaches to the intersection), our framework may deem it as part of the background information. In this case, 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 start moving again and they are identified as "waiting" 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.



(a)

(b)

(c)

(d)

(e)

Figure 1: Unsupervised background learning and subtraction in the traffic video sequence. (a) Image sequence from frame 1121 to frame 1228; (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); (e) The generated background for this sequence.

The key point here is that it is not necessary to obtain a perfectly 'clean' background image at each time interval. In fact, including the static objects as part of the background will not affect the extraction of moving objects. Further, 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.

2.3 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 [3], which enables the proposed framework to provide useful and accurate traffic information for ATMS.

Definitions: 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.

Definition 1: A bounding box B (of dimension 2) is defined by the two endpoints S and T of its major diagonal [Gonzalez93]:

B = (S, T), where S = (s_1, s_2) and T = (t_1, t_2) and $s_i \leq t_i$ for i = 1, 2. Due to this definition, the area of B: $Area_B = (t_1 - s_1) \times (t_2 - s_2)$.

Definition 2: The centroid ctd_O of a bounding box B corresponding to an object O is defined as follows:

$ctd_O = [ctd_{O1}, ctd_{O2}]$, where

$$ctd_{O1} = (\sum_{i=1}^{No} O_{xi}) / No \; ; \; ctd_{O2} = (\sum_{i=1}^{No} O_{yi}) / No \; ;$$

where No is the number of pixels belonging to object O within bounding box B, O_{xi} represents the x-coordinate of the ith pixel in object O, and O_{yi} represents the y-coordinate of the ith pixel in object O.

Let ctd_M and ctd_N, respectively, be the centroids of segments M and N that exist in consecutive frames, and δ be a threshold. The Euclidean distance between them should not exceed the threshold δ if M and N represent the same object in consecutive frames:

$$DIST(ctd_M - ctd_N) = \sqrt{(ctd_{M1} - ctd_{N1})^2 + (ctd_{M2} - ctd_{N2})^2} \leq \delta$$

In addition to the use of the Euclidean distance, some size restriction is applied to the process of object tracking. If two segments in successive frames represent the same object, the difference between their sizes should not be large. The details of object tracking can be found in [2].

Handling occlusion situations in object tracking: Unsupervised background learning and subtraction enables fast and satisfactory segmentation results, greatly benefiting the handling of object occlusion situations. A more sophisticated object tracking algorithm integrated in the proposed framework is given in [2], which 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 with similar sizes have ever separated from each other in the video sequence, then they can be split and identified as two objects with their bounding boxes being fully recovered using the object tracking algorithm. The results are demonstrated in Figure 2(a)-(c), where two vehicles have some overlapping in frame 142, but are identified as two separate objects in frame 132. Figure 2(c) demonstrates the final results by applying the occlusion handling method proposed in [2]. The segmentation results accurately identify all the vehicle objects' bounding boxes and centroids.

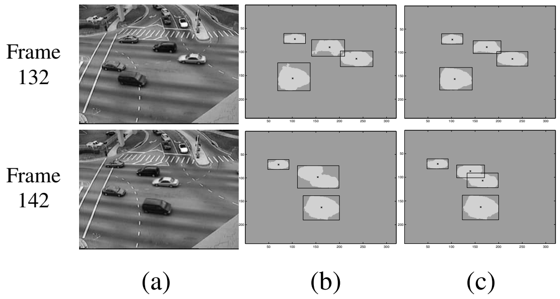
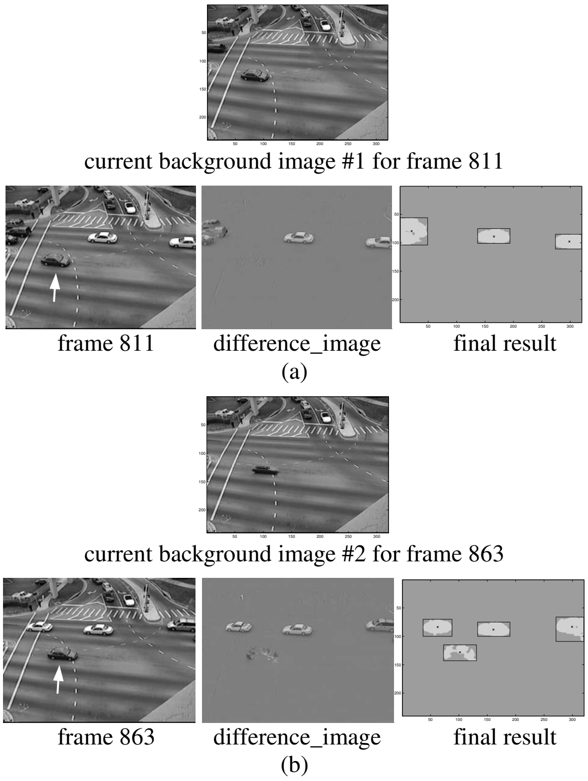


Figure 2: Handling two object occlusion in object tracking. (a) Video frames 132 and 142; (b) Segmentation maps for frames in (a) without occlusion handling; (c) The final detection results by applying occlusion handling.

However, there are cases where a large object overlaps with a small one. For example, a large bus merges with a small car to form a new big segment ("*overlapping*" segment) though they are two separate segments. In this scenario, the car object and the bus object that were separate in previous frames cannot find their corresponding segments in the currently processed frame by centroid-matching and size restriction. For these "*overlapping*" segments, we proposed a difference binary map reasoning method in [3] to identify which objects the "*overlapping*" segment may include. The idea is to obtain the difference binary map by subtracting the segmentation map of the previous frame from that of the current frame and check the amount of difference between the segmentation results of the consecutive frames. If the amount of difference between the segmentation results of the two consecutive frames is less than a threshold, the car and bus objects in the previous frame can be roughly mapped into the area of the "*overlapping*" segment in the current frame. Therefore, the corresponding links between the "*overlapping*" segment and the car and bus objects in the previous frame can be created, which 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 the previous frame.

3. Experimental Analysis

A real life traffic video sequence is used to analyze spatio-temporal vehicle tracking using the proposed learning-based vehicle tracking framework. It is a grayscale video sequence that shows the traffic flows on a road intersections for some time duration. The background information can be very complex by having road pavement, trees, zebra crossing, pavement markings/signage, and ground. The proposed new framework is fully unsupervised in that it can enable the automatic background learning process that greatly facilitates the unsupervised vehicle segmentation process without any human intervention. During the segmentation, the first frame is partitioned with two classes using random initial 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 since there is no significant difference between consecutive frames. By doing so, the segmentation process will converge fast, thereby providing support for real-time processing. The effectiveness of the proposed background learning process ensures that a long run is not necessary to fully determine the accurate background information. In our experiments, the current background information can be usually obtained within 20~100 consecutive frames and 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.


current background image #1 for frame 811

frame 811 difference_image final result
(a)

current background image #2 for frame 863

frame 863 difference_image final result
(b)

Figure 3: Segmentation results for video sequence. (a) Segmentation result for frame 811 using the background image #1; (b) Segmentation result for frame 863 using the background image #2.

Figure 3 shows the segmentation results for a few frames (811 and 863) along with the original frames, background images, and the difference images. As shown in Figure 3(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 those cars that stopped behind the zebra crossing. Since, in this time interval, they are identified as part of the background as they lack motion, they will not be extracted by the segmentation process as shown in Figure 3(a). However, the gray car (marked by white arrow in original frames) that was previously waiting in the middle begins to move around frame 860, triggering the generation of a new current background, shown in Figure 3(b) labeled as #2. From Figure 3(b), it is obvious that the gray car is fading, although not completely, from the background image. However, this fading is sufficient 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 3(b). Moreover, as shown in frame 863 in Figure 3(b), our method can successfully illustrate the slow motion effect; unlike many methods that have difficulties in dealing with it since it can be easily confused with noise-level information.

Based on our experimental experience, 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 other existing framework, in which one vehicle object has been segmented into several small regions and later these regions are grouped and linked with the corresponding vehicle object, no extra effort is needed 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 begin to show up. 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.

Further, 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 its relative spatial relations. The proposed framework has the potential to address a large range of spatio-temporal related database queries for ITS. For example, it can be used to reconstruct accidents at intersections in an automated manner to identify causal factors to enhance safety.

4. Conclusions

In this paper, we present a framework for spatio-temporal vehicle tracking using unsupervised learning-based segmentation and object tracking. An adaptive background learning and subtraction method is proposed and applied to a real life traffic video sequence to obtain more accurate spatio-temporal information of the vehicle objects. The proposed background learning method paired with the image segmentation is robust under many situations. 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 automatic and unsupervised, and 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 vis-à-vis intersection monitoring.

Acknowledgement

For Shu-Ching Chen, this research was support in part by NSF EIA-0220562 and NSF HRD-0317692. For Mei-Ling Shyu, this research was support in part by NSF ITR (Medium) IIS-0325260.

References

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

[2] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, “Object Tracking and Augmented Transition Network for Video Indexing and Modeling,” 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2000), pp. 428-435, Vancouver, British Columbia, Canada, November 13-15, 2000.

[3] S.-C. Chen, M.-L. Shyu, C. Zhang, and J. Strickrott, “Multimedia Data Mining Framework: Mining Information from Traffic Video Sequences,” Journal of Intelligent Information System, Special Issue on Multimedia Data Mining, vol. 19, no. 1, pp. 61-77, July 2002.

[4] R. Cucchiara, M. Piccardi, and P. Mello, “Image Analysis and Rule-based Reasoning for a Traffic Monitoring System,” IEEE International conference on Intelligent Transportation Systems, vol. 1, no. 2, pp. 119-130, Tokyo, Japan, June 2000.

[5] D. J. Dailey, F. Cathey, and S. Pumrin, “An Algorithm to Estimate Mean Traffic Speed Using Uncalibrated Cameras,” IEEE Transactions on Intelligent Transportations Systems, vol. 1, no. 2, pp. 98-107, June 2000.

[6] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee, “Using Adaptive Tracking to Classify and Monitor Activities in a Site,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition Proceeding, pp. 22-31, 1998.

[7] S. Kamijo, Y. Matsushita, and K. Ikeuchi, “Traffic Monitoring and Accident Detection at Intersections,” IEEE Trans. Intelligent Transportation Systems, vol. 1, no. 2, pp. 108-118, 2000.

[8] S. Peeta and H. S. Mahmassani, “System Optimal and User Equilibrium Time-dependent Traffic Assignment in Congested Networks,” Annals of Operations Research, pp. 81-113, 1995.

[9] X. Li, S.-C. Chen, M.-L. Shyu, and S.-T. Li, “A Novel Hierarchical Approach to Image Retrieval Using Color and Spatial Information,” the Third IEEE Pacific-Rim Conference on Multimedia 2002 (PCM2002), Lecture Notes in Computer Science: Springer-Verlag, pp. 175-182, Taiwan.