Proceedings of the
2007 IEEE Intelligent Vehicles Symposium
Istanbul, Turkey, June 13-15, 2007

A Street Scene Surveillance System for Moving Object Detection, Tracking and Classification

Huei-Yung Lin and Juang-Yu Wei

Abstract— **This paper presents a vision system for street scene surveillance. In addition to the capabilities of detection and tracking of moving objects, it is also able to recognize and classify the targets based on the walking rhythm. The classification results are further used for event analysis and video retrieval of interested scenes. The proposed technique is computational efficient and can be used for embedded real-time applications. The experimental results are presented for several image sequences of the real scenes.**

I. INTRODUCTION

Video surveillance for environment monitoring has gained significant research interests due to the availability of low-cost imaging hardware and high demand of commercial applications. It is traditionally used in the areas of industrial inspection, automated manufacturing, office or building security, and scientific experiments [8]. Recently, outdoor surveillance of public spaces for safety and security purposes has become one of the most important subjects in the computer vision research community. These issues are also the essential problems for the development of intelligent transportation systems (ITS). The ultimate goal is usually to build an intelligent surveillance system to reduce the usage of manpower and make the tasks more reliable. Some examples of using video information for intelligent transportation systems include highway traffic control, vehicle parking access control, driver assistance systems for intelligent vehicles, and pedestrian detection and tracking.

For highway traffic control, surveillance systems are used to estimate the vehicle flow and classify the types of observed vehicles based on the detected moving objects in the video sequences [5], [11]. Parking access control acquires the images of the vehicle's license plate for pattern recognition and identification using the existing database [10], [3]. Video surveillance for driver assistance is aimed to detect the changes of the environment, or monitor the driver's distraction level and used for intelligent safety deployment [13], [12]. Similar work dealing with the vehicles or drivers have been extensively investigated for the past few decades. As for the street scene surveillance related intelligent transportation, most of the previous research focused on the recognition and tracking of walking pedestrians [9], [2], [4]. There are, however, relatively few work concerning the street surveillance involving both the pedestrians and vehicles for safety and security purposes.

In this paper we proposed a video surveillance system for the street environment. In addition to the real-time detection and tracking of moving objects in the scene, another important objective is to classify the types of interested targets and then used for event analysis and video retrieval. For example, some surveillance applications might be interested in the pedestrian or vehicle counts for a given video stream, yet others might want to separate the image sequences with different classes of moving objects for efficient data storage or fast information access. The results can be used for traffic modeling or management, and searching of special events (such as car accidents) with least human visual investigation, respectively.

Under the normal street environment, there are generally three different types of moving objects, namely the vehicles (including cars and trucks), motorcycles (including bicycles), and pedestrians. The major work for our purpose is to first detect and track the moving objects, and then recognize the tracked targets based on the above three possibilities. Thus, the proposed street surveillance system mainly consists of the following three modules:

Detection and Tracking : Detect the foreground changes in the video stream and identify the moving target for visual tracking.

Recognition and Classification : Recognize and classify the identified target based on the predefined object models.

Event Summarization : Index the video sequence for event summarization using the moving target classification results.

Due to the mixture of non-rigid human body motion and rigid vehicle motion, conventional object tracking algorithm such as Kalman filter is not suitable for this system [1]. In the object detection and tracking module, a prediction and matching approach is adopted, provided that the uniform object motion is assumed. Recognition of the identified objects is achieved by checking two criteria: the height/width ratio and walking rhythm of the interested target (vehicle, motorcycle, or pedestrian). The surveillance video sequences are finally classified and indexed based on the target recognition results.

This paper is organized as follows. Section II describes the fundamentals of image processing techniques used in this work as well as the object detection and tracking module. In Section III, we present the proposed recognition module for moving target classification. Experimental results

The support of this work in part by the National Science Council of Taiwan under Grant NSC-95-2221-E-194-075 is gratefully acknowledged.

Huei-Yung Lin is with the Department of Electrical Engineering, National Chung Cheng University, Chia-Yi 621, Taiwan, R.O.C. lin@ee.ccu.edu.tw

Juang-Yu Wei is with the Department of Electrical Engineering, National Chung Cheng University, Chia-Yi 621, Taiwan, R.O.C.

1-4244-1068-1/07/$25.00 ©2007 IEEE.

Fig. 1. Processing pipeline of the detection and tracking module.

performance analysis of the system are presented in Section IV, followed by event summarization in Section V. Section VI concludes the paper and discusses possible directions of future work.

II. Foreground Detection and Object Tracking

The processing pipeline for the proposed object detection and tracking module is shown in Fig. 1. Given a sequence of images captured by a video camera, the first step of visual surveillance is to segment the moving objects or foreground regions from the static background scene. The simplest method is the direct image subtraction from the video sequence, which attempts to locate the changes in two consecutive image frames. Although this technique is easy to implement, it cannot be used to detect an object without continuous motion in the scene (e.g., the object remains static for a short period of time and then moves away). More sophisticated tracking techniques, such as optical flow, use the image brightness constancy to detect the object motion [7]. In addition to the high computation cost of the algorithms, they are also sensitive to the illumination changes and thus not suitable for the outdoor environment.

In this work, the background image subtraction method is used to segment the dynamic object in the scene. A background model of the scene is first generated based on the statistics of several input image frames. An intensity threshold is then used to segment the moving object from the difference between the background model and the current image frame. Since the background region of the scene will also change slightly due to the non-uniform outdoor illumination condition, it should be continuously updated after a period of time. To model the background scene caused by the illumination changes exclusively, the image is first converted to the HSV color space and only the brightness component is updated [6]. This process ensures that the pixel intensity values do not change significantly and the background scene is more robust for object segmentation. By thresholding the derived difference image, segmented and detected objects are represented by a binary image. After the morphological erosion and dilation operations for noise reduction, the bounding boxes of the targets are given by the nonzero horizontal and vertical projections of the blobs. Figure 2 shows the result of difference image and the detected object location.

To make the tracking algorithm more robust under different illumination conditions, a prediction and matching approach is applied on the detected target region to simultaneously track multiple objects in the image sequence [1]. The position of a target at time $t+1$ is predicted based on its positions at times t and $t-1$, and this prediction is then used to match with the detected objects. Since the moving speed of the object is assumed to be uniform, the displacement of the detected targets between two consecutive frames is used to approximate the same object's position in the next image frame. The new target position is updated in the image sequence if an object match is found at time $t+1$ by the following criteria:

$$(d_x, d_y) = (x_t - x_{t-1}, y_t - y_{t-1})$$
$$(x_{t+1}, y_{t+1}) = (x_t + d_x, y_t + d_y)$$

where (d_x, d_y) is the displacement vector of the moving object at time t, and (x_i, y_i) is the object location at image frame i.

The capability of multiple object tracking is essential for most surveillance systems. One major challenge of this

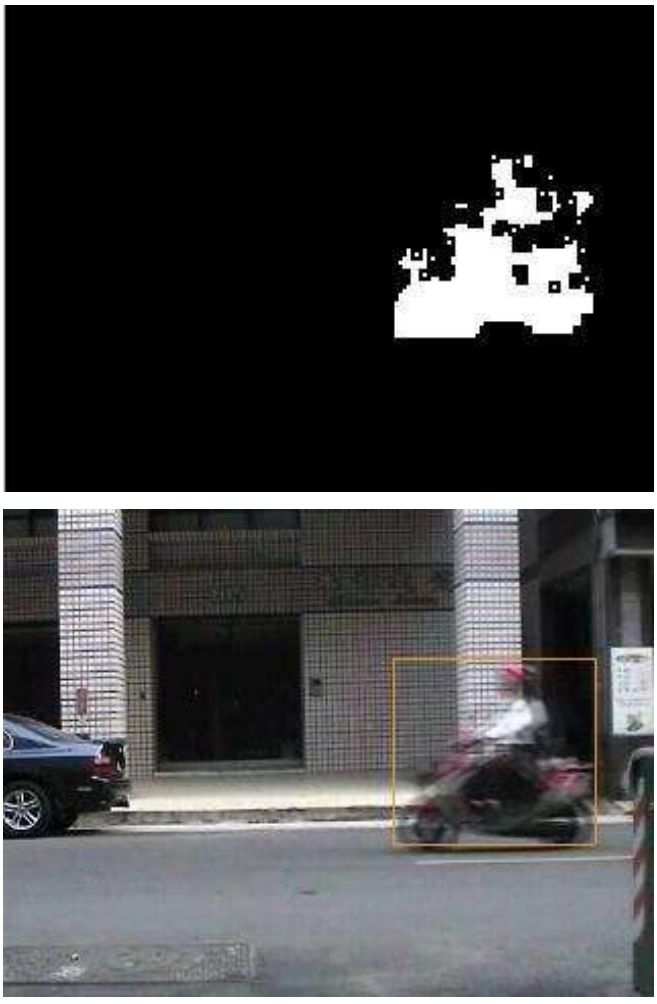


Fig. 2. The difference image and detected object location.

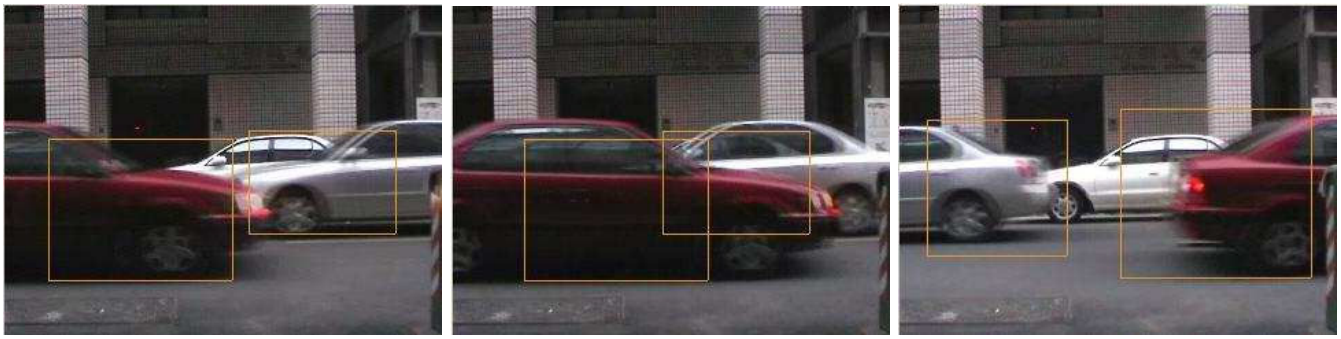


Fig. 3. Multiple object tracking with occlusion (two vehicles).

Fig. 4. Multiple object tracking with occlusion (two motorcycles).

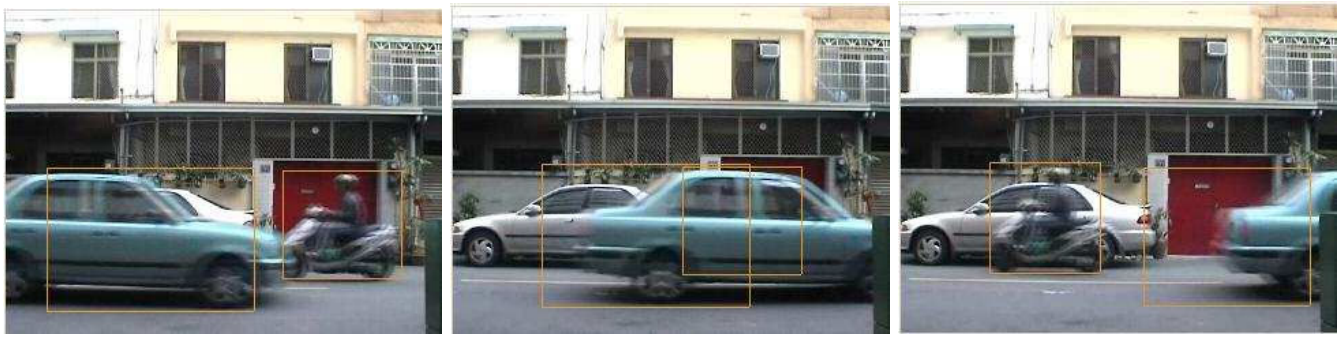


Fig. 5. Multiple object tracking with occlusion (one vehicle and one motorcycle).

requirement is the occlusion detection of multiple moving targets. However, if the object trajectory can be modeled by a uniform linear motion, then it is possible to use the area size of the tracked target to detect the occlusion – the target region usually becomes fairly large instantaneously if the occlusion happens between two moving objects. Thus, the positions of two individual moving objects can be recorded prior to the occlusion, and then used to track the subsequent targets after the occlusion ends. To determine the termination of the occlusion for a street surveillance scene, the following three rules are adopted:

- The blob size of the connected occluded targets is much larger than the size of the individual targets.
- The positions of newly detected objects are close to the positions predicted prior to the occlusion.
- The vertical positions of the detected objects remain constant for horizontal object motion.

In our implementation the objects move nearly horizontally along the image scanlines, which generally yields robust occlusion detection. Figs. 3 – 5 illustrate the multiple object tracking based on the above rules. Although the object tracking positions might not be accurate immediately after the termination of the occlusion, they will be modified automatically in the subsequent image frames.

III. Recognition and Classification

To recognize the identified and tracked objects in the image sequence, we are primarily interested in three classes of objects: vehicle, motorcycle (or bicycle) and pedestrian. For general cases, the vehicles and motorcycles can be modeled as rigid bodies, but the pedestrians should be considered as deformable objects. Thus, the basic recognition and classification criteria are based on the height/width ratio and "walking" rhythm of the target. The flowchart of the

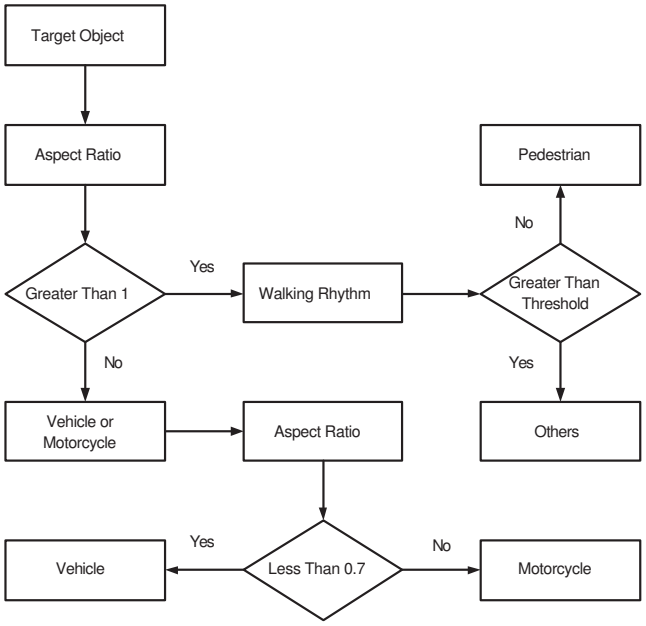


Fig. 6. Recognition and classification flowchart.

proposed recognition and classification module is illustrated in Fig. 6.

Given an identified target object, the first step is to check the ratio of its height to width. For general cases the height to width ratios of the pedestrians are greater than 1, and the ratios of vehicles and motorcycles are less than 1. Moreover, the ratios of vehicles are usually smaller than those of motorcycles. Thus, these two types of objects can be further distinguished by a threshold of about 0.7 from experience. Checking the target's width/height ratio for recognition is a simple technique, but it might cause some potential problems. First, it is possible that tall motorcycle or bicycle riders make the object's height larger than the object's width. In this case, the target will be classified as pedestrian even if a motorcycle is actually detected. Second, if there is occlusion between two target objects, then the detection and tracking process will stop until the occlusion is terminated. Consequently, the targets can only be recognized and classified based on their appearances near the left and right borders of the image. In these cases the identified target usually has greater height/width ratio than the ground truth since it is partially truncated due to the camera's limited field of view.

To overcome the above problems and ensure the correct identification of pedestrians, walking rhythm is further verified for the moving object with the height/width ratio greater 1. The idea is based on the fact that the target width is approximately a constant for vehicles and motorcycles but periodically changed for pedestrians, especially on their lower halves, due to the swing motion of the legs. Thus, pedestrians can be distinguished from vehicles or motorcycles by checking if there exist significant width changes. Since the time domain object width detection might not be accurate enough due to noise and shadow, and our goal is only to distinguish the periodic function from a constant, frequency domain approach will generally provide more robust pattern classification. Thus, Fourier transform is applied on the function of object width (as a function of time) to find the corresponding power spectrum, and then used to distinguish the vehicles and motorcycles from the pedestrians.

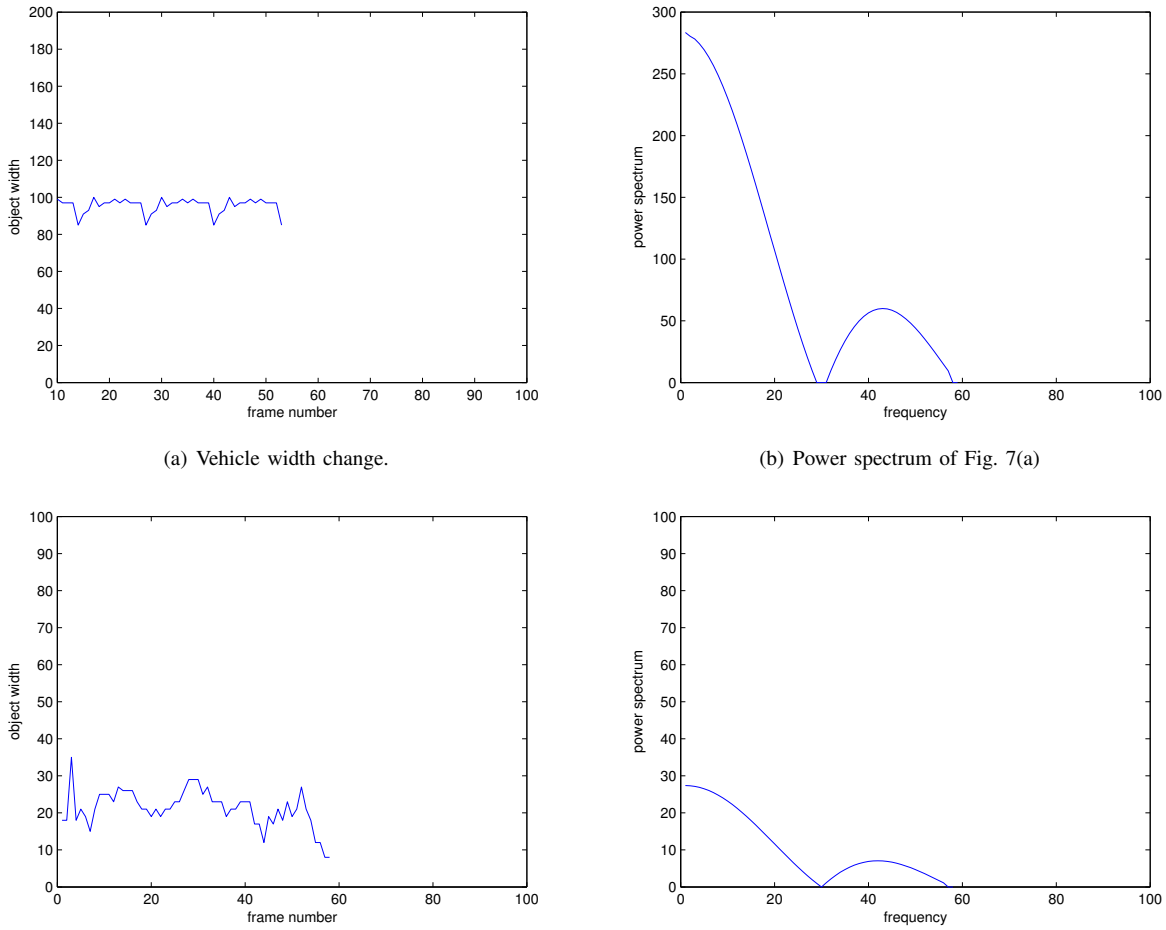
Fig. 7 shows the object width functions of a vehicle and a pedestrian (Figs. 7(a) and 7(c)), and their power spectra (Figs. 7(b) and 7(d)). The response in frequency domain is used for target classification. Since the low frequency components of both cases are large, the second lobe of the power spectrum is used to distinguish between the vehicles and pedestrians. In the implementation, a threshold of 40 is selected to separate these two classes based on the training sample video sequence.

IV. Event Summarization

The objective of event summarization for the street scene surveillance is to index the video sequence for efficient storage and fast information retrieval. It is commonly a major issue for a large data set of video recording. For example, searching for a specific event might be tedious work if one has to go through the whole video sequence manually. Thus, in this work we use a simple scheme to separate and index the video stream based on the target classification. The image frames corresponding to static scenes are first removed from the video sequence. The starting and ending frame numbers of pedestrians, motorcycles and vehicles are indexed for the remaining video stream. As illustrated in Figure 8, searching for the scenes consisting of any combination of pedestrians (green), motorcycles (red), and vehicles (blue) based on the time stamps (or indices) of the target appearance interval can be achieved very efficiently. Furthermore, new video sequences can be generated by the event summarization for other specific applications.

V. Experimental Results

The proposed street surveillance system has been tested on several real scenes. Figure 9 shows the background images of four data sets used for performance evaluation. Each data set contains 27, 18, 20, and 20 minutes of video sequence at 15 fps, respectively. Figure 10 illustrates the graphical user interface of the surveillance system. The moving object detection and recognition results are shown in Table I. For the first three video sequences the illumination conditions do not change significantly over time, and the recognition rate is about 96.8%. If the environment is under strong sunshine, as appeared in the last data set, the reflection of the target surface usually makes the preprocessing more difficult. Consequently, the recognition rate is slightly lower mainly due to the misclassification of the cases with a single vehicle and two adjacent pedestrians. The overall recognition rates for pedestrian, motorcycle and vehicle are 87.5%, 98.4% and 94.7%, respectively, for the video sequences shown in Table I.



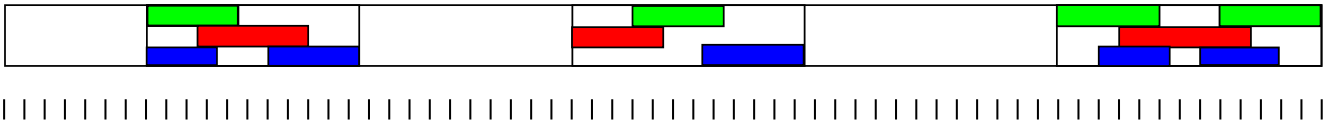
(a) Vehicle width change.

(b) Power spectrum of Fig. 7(a)

(c) Pedestrian walking rhythm.

(d) Power spectrum of Fig. 7(c)

Fig. 7. Walking rhythms and the corresponding power spectra.



Frame No.

Fig. 8. Event summarization schematic.

Fig. 9. Experimental environment background scenes.

VI. Conclusion

The application of image analysis techniques to intelligent transportation systems has become an important research topic in the past few years. In this work, we have presented a vision system for street surveillance purposes. The moving objects in the street scenes are classified as either pedestrians,

TABLE I

MOVING OBJECT DETECTION AND RECOGNITION RESULTS (GT: GROUND TRUTH, TP: TRUE POSITIVE, FP: FALSE POSITIVE, FN: FALSE NEGATIVE).

	Video #1				Video #2				Video #3				Video #4			
	GT	TP	FP	FN	GT	TP	FP	FN	GT	TP	FP	FN	GT	TP	FP	FN
Pedestrian	5	4	2	1	1	1	0	0	7	6	0	1	3	3	1	0
Motorcycle	75	72	1	3	65	65	0	0	99	98	2	1	73	72	2	1
Vehicle	39	38	0	1	25	22	0	3	23	22	0	1	27	26	0	1

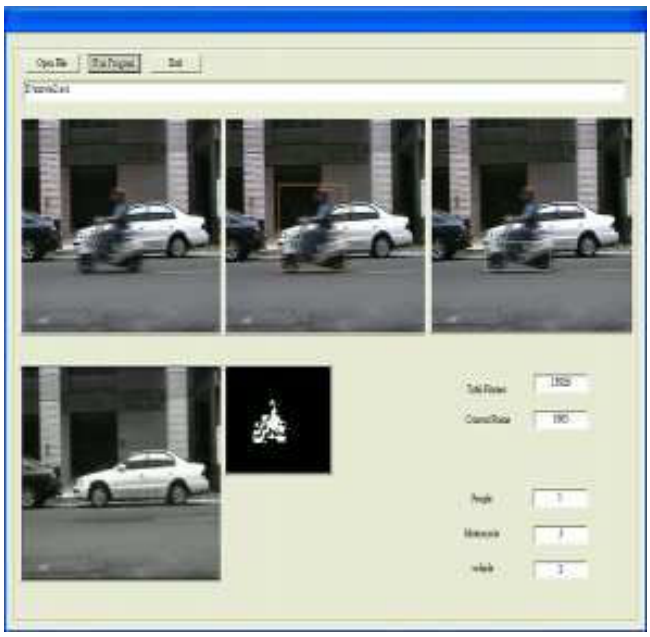


Fig. 10. Graphical user interface of the surveillance system.

motorcycles or vehicles based on the detection and tracking results. Furthermore, an event summarization scheme is used for quick identification of interested scenes. The proposed technique is computational efficient and can be implemented on embedded devices for real-time applications.

VII. ACKNOWLEDGEMENTS

The support of this work in part by the National Science Council of Taiwan under Grant NSC-95-2221-E-194-075 is gratefully acknowledged. The authors would like to thank Mr. Li-Wei Kao for preparation of some test videos.

REFERENCES

[1] F. Brémond and M. Thonnat, "Tracking multiple non-rigid objects in video sequences," *IEEE Transaction on Circuits and Systems for Video Technology Journal*, vol. 8, no. 5, 1998.

[2] A. Broggi, M. Bertozzi, F. A., and M. Sechi, "Shape-based pedestrian detection," in *IEEE Intelligent Vehicles Symposium*, 2000, pp. 215–218.

[3] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen, "Automatic license plate recognition," *IEEE Trans. Intelligent Transportation Systems*, vol. 5, no. 1, pp. 42–53, March 2004.

[4] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking pedestrian recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 3, September 2000.

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

[6] R. Gonzalez and R. Woods, *Digital Image Processing, 2nd Edition*. Prentice Hall, 2001.

[7] B. Horn, *Robot Vision*. MIT Press, 1986.

[8] R. Jain, R. Kasturi, and B. Schunck, *Machine Vision*. McGraw-Hill, 1995.

[9] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 878–885.

[10] T. Naito, T. Tsukada, K. Yamada, K. Kozuka, and S. Yamamoto, "Robust license-plate recognition method for passing vehicles underoutside environment," *IEEE Trans. Vehicular Technology*, vol. 49, no. 6, pp. 2309–2319, Nov. 2000.

[11] G. S., M. O., M. R. F. K., and P. N. P., "Detection and classification of vehicles," *IEEE Trans. Intelligent Transportation Systems*, vol. 3, no. 1, pp. 37–47, March 2002.

[12] G. P. Stein, O. Mano, and S. A., "Vision-based acc with a single camera: Bounds on range and range rate accuracy," in *IEEE Intelligent Vehicles Symposium*, 2003, pp. 120–125.

[13] M. Trivedi, S. Y. Cheng, E. Childers, and S. Krotosky, "Occupant posture analysis with stereo and thermal infrared video: algorithms and experimental evaluation," *IEEE Trans. Vehicular Technology*, vol. 53, no. 6, pp. 1698–1712, Nov. 2004.