Real Time Moving Vehicle Detection and Reconstruction for Improving Classification

Tao Wang
Department of Computer Science
The Graduate Center of CUNY, New York, NY 10016
The City College of New York, New York, NY 10031
twang@cs.ccny.cuny.edu

Zhigang Zhu
Department of Computer Science
The City College of New York, New York, NY 10031
The Graduate Center of CUNY, New York, NY 10016
zhu@cs.ccny.cuny.edu

Abstract

Vehicle images captured by traffic and surveillance video cameras in various conditions usually exhibit several unexpected variations that worsen vehicle classification. These factors include occlusions, motion blur, and changes in perspective views. Complete and normalized views of vehicle images, if being able to reconstructed from the unsatisfactory data, will facilitate more accurate data labeling, feature extraction and multi-class vehicle classification. We propose a multimodal temporal panorama (MTP) approach to accurately extracting and reconstructing moving vehicles in real-time using a remote multimodal (audio/video) monitoring system. The MTP representation consists of: 1) a panoramic view image (PVI) for detecting vehicles using the concept of 1D vertical detection line; 2) an epipolar plane image (EPI), generated from 1D epipolar lines along the vehicles' moving paths, to characterize their speeds and directions; and 3) corresponding audio signals collected at the vehicle detection point to reduce false target detection in the PVI. Using the MTP approach, reconstructed vehicles all have the same side views, with less or no occlusions and motion blur. Using SVM classifiers for multiclass problems indicates that the classification accuracy using reconstruction results improves about 10% over that using corresponding vehicle images from original video for a dataset of about 140 vehicles. Our ultimate goal is to use the audio-visual vehicle data for multimodal vehicle classification and anomaly detection.

1. Introduction

In applications such as traffic monitoring [1, 2] and check-point vehicle inspection [3] where the cameras are set in well-chosen standpoints and stationary, the detection of vehicles and anomaly could be made easy by the use of motion detection and background subtraction. However, the vehicle classification and identification could be quite challenging. For vehicle detection, most methods [4, 5] assume that the desired vehicles can be detected by image differencing. Then various kinds of vehicle features like shapes, textures, etc. are extracted easily to make the vehicle classification straightforward. However, several environmental variations will significantly affect the accuracy of vehicle classification. This will be even more the case for long-range vehicle detection and inspection, where the sensors (cameras) can only be set in a remote location. In such a scenario, the standpoints of and views from the sensors to a road could be constrained due to large distances, and occlusions such as tress and other facilities. This will result in failure of vehicle detection and degrade the accuracy of later vehicle classification and recognition. Another environmental variation is that the perspective views (ranges, directions) of captured vehicles which also vary greatly. When a vehicle is observed along a lane, it will have different appearances/resolutions in different video frames over the period of time the vehicle can be seen. Also, the video data of the vehicle could be in low resolution and subject to motion blur. In addition to the vision-based approaches, there are some systems [6-9] using other sensors, such as sonar, infrared cameras, or laser Doppler vibrometer (LDV) to detect vehicles on road. It has been shown that the use of multimodal sensors provide better performances in object detection and classification.

In this paper we use a long-range multimodal sensor platform developed in [9] to monitor the road traffic that includes both visual and audio information. In our approach, we represent both visual and audio data in a multimodal temporal panorama (MTP) that we proposed in [8], which shows detection, motion, and acoustic information simultaneously. MTP provides a very effective user interface to visualize and analyze the alignment of the video and acoustic information of passing-by vehicles, thus facilitating the joint detection and classification of vehicles using both visual and audio information. It provides: 1) multi-modal information including visual presentation from a panoramic view image, motion presentation from an epipolar plane image, and acoustic information from an audio wave scroll; 2) real time detection, reconstruction of the vehicles' visual appearances, synchronized with their acoustic signatures; and (3) a very effective user interface for training data labeling in both video and audio domains. While the work described in [8] focuses on data representation and user interface for data labeling, in this paper we extend the work and have made three new contributions. First a

robust vehicle reconstruction algorithm is developed using both panoramic view images and epipolar plane images, and error analysis is performed. Since the generation of the MTP is done in real time, the reconstruction takes place immediately after a vehicle is detected. Second, audio information is used to remove some false detecting targets before reconstruction, even though our ultimate goal is to perform multimodal vehicle classification. Third, we generate a dataset of about 140 different types of vehicles, and perform vehicle classification comparison between reconstructed vehicle images and their corresponding images from the original videos. We note that the classification of the reconstructed vehicle images has significant performance improvement over that of the corresponding unreconstructed vehicles.

The rest of paper is organized as follows. Section 2 discusses some related work. Section 3 shows the data representation and detection using the MTP. Section 4 describes the vehicle reconstruction algorithm. Section 5 provides the reconstruction error analysis. Section 6 presents the multiclass vehicle classification method. Experimental results for both error analysis and vehicle classification are shown in Section 7. Some ongoing work on multimodal vehicle classification is discussed in Section 8. Conclusions are provided in Section 9.

2. Related Work

The main usage of panoramic view images (PVIs) in [10, 11] was for scene understanding. A 1D slit scanning approach was used to construct route panoramas when a camera was mounted on a moving vehicle. In these works, the resulted PVIs do not require inter-frame matching of video images. Epipolar plane images (EPIs), combined PVIs, are used to track the motion of vehicles [12, 13] for the purpose of automatic traffic monitoring. However, they are not interested in real time vehicle reconstruction to improve the classification. Also, all these papers only deal with video data. In our paper we also capture and process acoustic data of moving vehicles using a laser Doppler Vibrometer (LDV) to reduce false target detection in the visual modality to limit the reconstruction only on vehicle candidates, and to facilitate multimodal vehicle classification. Works using both audio and video for surveillance can be found in [14, 15]. In their approaches, the full video images are processed, which are sometimes computationally expensive but unnecessary. The synchronized, manual labeling of the audio and video data for training classifiers could be very tedious. Also, the environmental variations, such as the changes of the entire background or presence of other stationary objects, could not be handled properly. These heavily affect the classification accuracy. In [16], the authors used entropy as an underlying measurement to calculate traffic flows and vehicles speed. However, these approaches cannot further classify vehicles to more detailed types for proving more accurate information. Therefore, it is important not only to represent multimodal information but also to perform vehicle reconstruction in order to improve vehicle classification accuracy.

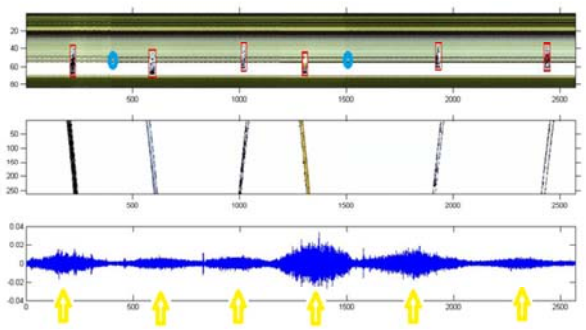

Figure 1. An example of the multimodal temporal panorama (MTP) synopsis layer that consists of a PVI, an EPI, and an audio wave scroll (from top to bottom). Red boxes show correct targets. Blue ovals show false targets. Yellow arrows point to meaningful sounding targets. False targets can be eliminated since they are considered as background from

3. Multimodal Data Processing

3.1. Data Representation

We capture both visual and audio data simultaneously and represent them in the multimodal temporal panorama (MTP) [8]. The MTP has two layers: a temporal synopsis layer (Fig. 1) and a layer of snapshots for individual vehicles. In the synopsis, there are mainly three synchronized panoramas: a 2D spatial-temporal *panoramic view image* (PVI) concatenated from 1D vertical detection lines from a selected column location of all video frames; a 2D spatial-temporal *epipolar plane image* (EPI) concatenated from 1D horizontal epipolar lines along the direction of vehicles' motion; and an *audio wave scroll* for visualizing vehicles' sounds. The PVI and the EPI are used for the target detection and the motion estimation of vehicles, respectively. The temporal energy from a window of audio signals can be calculated to indicate a silent background period or a sounding target, thus, can reduce false target detection from previous two panoramas. Then in the snapshot layer, the occlusion-free, motion-blur-free, and view-invariant *visual reconstruction* of each vehicle (with both shape and motion information) and the *acoustic signatures* (e.g. spectrogram) are recorded. The MTP facilitates the synchronization and integration of the information across the three modalities, both for automatic and interactive vehicle and traffic analysis, thus providing more succinct and reliable information for tasks like moving vehicle detection and classification using visual, motion, and audio information. The 1D scanning technique for generating multimodal

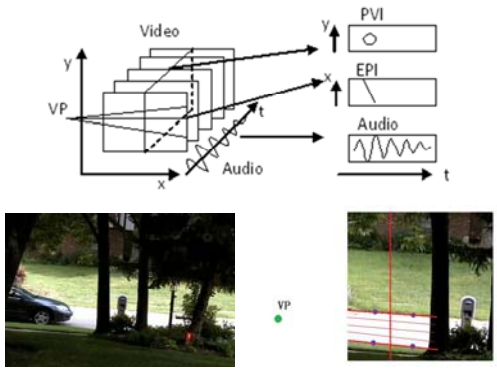

Figure 2. MTP representation(top), a scene with many occlusions (bottom left), and initial selection of an area with least occluded vertical detection line and edges of the road for determine the multiple epipolar lines (bottom right).

temporal panorama does not require the whole body of a vehicle in the field of view. Instead a stationary camera monitoring an area of interest is sufficient to detect a passing-by moving vehicle that can only be partially viewed from the camera. With this simple technique, there are three main advantages. First, there will be less or no occlusions. Second, the projections of all vehicles have the same standard side view. Third, motion blur is removed after the vehicle is reconstructed using the PVI and EPI. Furthermore, the acoustic signals of a vehicle that passes the checkpoint are also recorded using a LDV from a large distance and aligned with video panoramas in the time axis.

Fig. 2 shows the creation of the MTP. The PVI is generated by consisting of a vertical slit of single pixel cross all frames. The least occluded line in the scene is selected initially. Then, an epipolar line on the horizontal direction is selected to track the motion of a vehicle on the road that forms the EPI. The epipolar line, which indicates the moving path of a vehicle, is determined once the vehicle is detected via PVI. It is a line that connects a point on the vehicle to the vanishing point of two parallel lines on the road sides. The audio information is obtained to remove some false target detection before perform vehicle reconstruction.

3.2. Real Time Target Detection

Adaptive Gaussian mixture model for background subtraction [17, 18] are applied on both PVI and EPI during the generation of these two spatio-temporal images that produce a panoramic detection image (PDI) and a motion detection image (MDI), respectively. Only a small window of background containing a few liens needs to be trained initially, and then new incoming lines are accumulated to update the model. It is much faster than performing the subtraction on the whole frames. Also the result is more consistent since there is little variation in

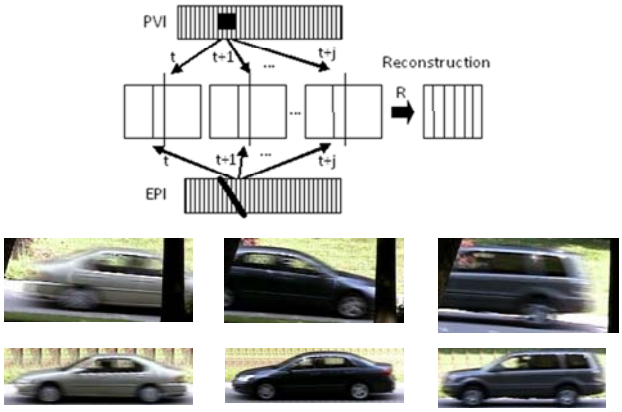

Figure 3. Reconstruction procedure (top) and samples of original images (middle) and reconstruction results (bottom) for Nissan Altima, Honda Accord and Honda Pilot

consecutive background lines over time in the video sequence. The process is performed online for every new frame, so the detection is done in real time. The total spectral energy of the acoustic signals obtained from the LDV is calculated in the audio detection "image" (ADI) from a starting time t_s to a finishing time t_f when a possible object is detected in the PVI. Note that the LDV has advantage that can capture the acoustic signals of the target (vehicle) by pointing its laser beam from a large distance to a retro-reflective surface that is right at the check point where the vehicle passes. The results in the ADI combined with PDI and MDI can determine the likelihood ψ of a target g at *ith* frame as:

$$\psi(g_i) = p(g_i|PDI_i, MDI_i, ADI_i), for\ i \in [t_s, t_f] \quad (1)$$

Thus, the addition information from the ADI can reduce some false targets that are not moving vehicles.

The original frame shot of the object, which is used to compare with reconstructed result, can be retrieved by measuring the boundary of detected object in the PVI. The EPI is used for acquiring column pixel locations of the vehicle in the horizontal direction, thus can be used to estimate the moving direction and speed of a vehicle. Together, the reconstruction can be performed.

4. Reconstruction Algorithm

Vehicle reconstruction is necessary since the vehicles may be occluded by other stationary objects, such as bushes, trees, parked vehicles or others. The motion blur is mostly caused by the interlacing of the camera, and can be removed after reconstruction. In addition, reconstructed vehicles all have the same views, which should improve the recognition and classification performance while keeping classifiers simple.

The general idea of reconstruction is demonstrated in Fig. 3. Each vertical line in the detected region in the PVI indicates a particular time frame I_t in the original video.

The slope m of the vehicle's locus at the corresponding time t in the EPI shows the relative speed v_t of a moving vehicle as:

$$v_t = m = \frac{\partial x}{\partial t} \quad (2)$$

In other words, it equals to the number of pixels in the motion direction in the original image that need to be extracted. The speed of the vehicle is used to accurately align the even and odd fields of the image pieces into a single image piece of a frame so that the original image resolution in the vertical direction is restored. The sign of the slope indicates the direction the vehicle moves to. So, if the vehicle moves from left to right, the image piece to the left of the vertical detection line (here defined as referenced line *rl*,) is extracted. If the vehicle moves from right to left, the image piece to the right of the *rl* is used. This is because the concatenation of PVI is in the left-to-right (or time increasing order). Then the image slice S_t at time t is:

$$S_t = I_t^J,\ J = \{j | j \in (rl, rl + v_t)\} \quad (3)$$

where J is the number of columns in the original time frame need to be selected. If the number is not an integer, then interpolation between two consecutive frames is applied. Although, the camera does not need to be perpendicular to the moving path of the vehicle, the segmented image pieces cannot be horizontal aligned smoothly if there is a rolling angle of the camera. An affine transformation is used to rectify those image pieces:

$$S_t \mapsto A_\gamma S_t + b \quad (4)$$

where A_γ is the rotation matrix has rolling angle γ, and b is translation vector. If the true rolling angle is not known in advance, it can still be calculate from the initial image shot as:

$$\gamma = \tan^{-1} \frac{Ep_y - Vp_y}{Ep_x - Vp_x} \quad (5)$$

where (Ep_x, Ep_y) is the intersection point of the referenced vertical line and selected epipolar line; and (Vp_x, Vp_y) is the vanishing point of any two parallel lines showing the roads structure. Then the reconstructed image I_R for a vehicle is the integration of all image pieces from starting time t_s to finishing time t_f when the vehicle is observed through the reference vertical detection line:

$$I_R = \bigcup_{i=t_s}^{t_f} A_\gamma I_t^J \quad (6)$$

5. Error Analysis

To show the accuracy of reconstructed image results, we perform error analysis under two cases depending on whether the true sizes of vehicles are known or not. For the first case, giving the true length L and the true height H of a vehicle, the relative errors of a vehicle in the length ε_L and in the height ε_L are:

$$\varepsilon_L = \frac{|L - L'|}{L}, L' = \frac{I_L D_m}{f_L} \quad (7)$$

$$\varepsilon_H = \frac{|H - H'|}{H}, H' = \frac{I_H D_m}{f_H} \quad (8)$$

where L' and H' are the length and the height in reconstructed result, respectively. I_L and I_H are the width and the height of the reconstructed vehicle image in pixels. f_L is the focal length in horizontal direction, and f_H is the focal length in vertical direction. D_m is the distance of a vehicle at the *mth* lane. We also perform a theoretical error analysis in order to compare with the actual errors calculated with Eq. (7) and Eq. (8). The theoretical relative errors in length ε'_L and in height ε'_L are:

$$\varepsilon'_L = |\frac{\delta L}{L}|, \delta L = \frac{D_m}{f_L} \delta I_L \quad (9)$$

$$\varepsilon'_H = |\frac{\delta H}{H}|, \delta H = \frac{D_m}{f_H} \delta I_H \quad (10)$$

where δI_H and δI_L are the measurement errors in the height and length directions of the image of a vehicle (in pixels).

If a vehicle's size is not known, we manually measure the length L'' and the height H'' of the vehicle in the original image corresponding to the reconstructed image at time t_m, where t_m is the time the vehicle half way passes through the detection line. Note that the vehicle may be partially occluded at the front or the rear part, or cannot be fully displayed in individual image frames. Therefore we combine the image frames at time t_s or t_f that have the vehicle partially displayed so that the correct length and height can be measured. Here t_s and t_f are the starting and finishing time the vehicle is detected. Then, the calculation of the relative errors for the unknown vehicle is just a matter of substituting L'' and H'' for L and H in Eqs. (7) and (8), respectively.

6. Audio and Visual Vehicle Features

Various visual and audio features could be extracted from the multimodal data and integrated for vehicle classification. Based on the reconstruction, vehicles' visual images are invariant to perspective views, so the size information of vehicles could be useful to distinguish small vehicles (sedans) and large vehicles (buses). For multiclass classification, histograms of oriented gradients (HOG) feature, which keeps the texture and local structure statistically, is used. We apply the HOG feature for both reconstructed vehicle images and their corresponding original images for comparison of classification. The audio information can provide complementary acoustic signatures in addition to the visual features. The Mel-frequency cepstral coefficients (MFCC) are used for audio features. They are commonly used to perceptually represent the frequency band responses of the human auditory system. They are good at characterizing the spectral variations and sharpness of acoustic signatures. In experimental results, we will show that the combination of

both audio and visual features could provide improvement than using individual features in vehicle classification.

To perform multiclass vehicle classification, we use a linear based support vector machine (SVM) [19] for the small dataset. One-against-one classifiers are used to choose the best class. Other techniques, such as, one-versus-all, or constructing multi-class SVMs could also be used but are not described here.

7. Experimental Results

In our experiment, we used a Canon 50i PTZ camera to capture video sequences at a two-lane road from a distance of about 25 meters. The road side contains a lot of trees and other objects such that the views of the moving vehicles are always partially occluded. Also, the PTZ camera captures analog signals, so we have to handle the interlaced scanning to reduce motion blur as well. The resolution of the camera is 720x480 pixels with a frame rate of 30 frames/second. In order to improve the detection rate by removing false targets via reconstruction, we also recorded vehicles' acoustic signals using the mono sound track of a sound card at 22.5 KHz sample rare and 16 bit resolution from a laser Doppler vibrometer (LDV) when its laser beam was pointed to a retro-reflective surface (e.g. a traffic sign) close to the "check point" for vehicles. The LDV, acted as a long range remote microphone, is basically a non-contact long range vibration sensor that is able to capture acoustic and vibrational signals of vehicles that vibrate the retro-reflective surface. The data were collected at different times over a period of two weeks.

7.1. Reconstruction Error Analysis

The road has two lanes of opposite directions, with one is about 24.8 meters and the other is about 26.8 meters to the platform. The focal length of the camera is 10.5 mm under 15x zoom. We used three known vehicles, Nissan Altima, Honda Accord, and Honda Pilot to evaluate the accuracy of vehicle reconstruction, each passing through the check point for 10 times. Some of the reconstruction results as well as their corresponding frame shots are shown in (Fig. 3 bottom). The actual relative error and theoretical relative error results are shown in Table 1. The theoretical errors are obtained by assuming the measurement errors in the height and length directions of the image of a vehicle (δI_H and δI_L, in Eqs. 9 and 10) are both one pixel. The actual reconstruction errors are comparable to the corresponding theoretical errors, and the average reconstruction error in both length and height is about 4%.

Table1.Reconstruction error analysis for vehicles of known types

Types	Nissan Altima	Honda Accord	Honda Pilot	Total Avg. Err.
True L(mm)	4661	4811	4849	-
Act. err. in L	3.87%	5.31%	3.86%	4.34%
Theo. err. in L	3.87%	5.14%	3.70%	4.24%
True H (mm)	1420	1445	1847	-
Act. err in H	4.64%	5.37%	1.68%	3.90%
Theo. err in H	4.46%	5.28%	1.29%	3.70%

7.2. Classification of Reconstructed Images

In our current dataset, there are 140 vehicle images reconstructed; their corresponding original images are also captured for comparison. We divide and label them in four groups: sedan, van, pickup truck and bus. There are 89 samples for training and 51 samples for testing. The image features extracted are histograms of gradients (HOGs). Each vehicle image is divided into 6x3 grids and each grid has 9 bins so that the result HOG feature vector for a vehicle image has 162 dimensions. Linear SVMs are used as the baseline classifier throughout the comparison. We use LIBSVM 3.1 [20] to solve the multiclass problem using the one-against-one technique. The testing accuracy using reconstructed results is about 80.39% and the confusion matrix is shown in Table 2.

We also applied the same feature extraction method to the original images corresponding to the reconstructed results. Note that the original images may include partial occlusions, various side views and motion blur. The training size and testing size are as same as reconstructed samples. We also used the same classifier to be trained using the training set, but this time with the original images. The testing accuracy using the original images is about 70.59% on average and its confusion matrix is shown in Table 3.

The comparison shows that the classification on the reconstructed results has about 10% improvement over that on original images because the reconstruction removes a lot of noises such as occlusions, motion blur and view changes. By comparing the confusion tables, it can be seen that the reconstruction has more impact to the classification of the sedans and the trucks, most probably because they have more variations than the vans and the busses in our experiments.

7.3. Audio and Visual Classification

In processing the multimodal temporal panoramas (MTPs), the audio information can be used to reduce false target detection. More importantly, audio signatures of

Table 2. Performance improvement with reconstruction over original images (S-Sedans, V-Vans, T-Pickup Trucks, B-Buses). Expected label in columns, actual lable in rows.

Reconstruction results				
Accuracy: 80.39%				
	S	V	T	B
S	17	5	0	0
V	1	18	0	0
T	2	1	4	0
B	0	0	1	2

Original images				
Accuracy: 70.59%				
	S	V	T	B
S	13	9	0	0
V	1	18	0	0
T	3	1	3	0
B	0	0	1	2

Table 3. Classification peroformance using audio feautre only, visual feautre only, and combination of both.

Audio (MFCC) only				
Accuracy: 70.59%				
	S	V	T	B
S	20	2	0	0
V	6	13	0	0
T	1	4	2	0
B	0	2	0	1

Visual (HOG) only				
Accuracy: 80.39%				
	S	V	T	B
S	17	5	0	0
V	1	18	0	0
T	2	1	4	0
B	0	0	1	2

Audio + Visual				
Accuracy: 88.24%				
	S	V	T	B
S	20	2	0	0
V	1	18	0	0
T	1	1	5	0
B	0	0	1	2

vehicles can improve the classification performance by combining them with visual features extracted from the reconstructed vehicle images. We use the first 15 coefficients of MFCCs and calculate their means and standard deviations into a feature vector of 30 dimensions. Results are then scaled to the same range with visual HOG features in order to integrate them. The experiment results on the current dataset show that there is about 8% improvement when using the combined audio and visual features over that using only the visual modality (Table 3).

8. Conclusions

In this paper, we represent multimodal data into a multimodal temporal panorama for real time vehicle detection and reconstruction. The combination of detection and motion estimation are used for vehicle reconstruction to remove occlusions, motion blur and variations of views. The classification analysis with a small dataset shows that the reconstructed results can provide significantly better accuracy in vehicle classification than their original images. Multimodal classification also shows promising results. In the future, we will generate a larger dataset with more visual and audio features types, and perform more comprehensive multimodal classification in terms of classifier designs, feature selection and performance analysis.

Acknowledgments

This work is supported by the Air Force Office of Scientific Research (AFOSR) under the 2011 Air Force Summer Faculty Fellow Program (SFFP), by the National Collegiate Inventors and Innovators Alliance (NCIIA) under an E-TEAM grant (No. 6629–09), and by a PSC-CUNY Research Award. The work is also partially supported by NSF under award #EFRI-1137172.

References

[1] Y. Wang, Real-Time Moving Vehicle Detection with Cast Shadow Removal in Video Based on Conditional Random Field, *IEEE Transaction on Circuits and Systems for Video Technology*, vol 9, iss. 3, pp 437-441, March 2009.

[2] J. Zhou, D. Gao, D. Zhang, Moving Vehicle Detection for Automatic Traffic Monitoring, *IEEE Transaction on Vehicular Technology*, vol. 56, iss. 1, pp 51-59, Jan. 2007

[3] P. Dickson, J. Li, Z. Zhu, A. R. Hanson, E. M. Riseman, H. Sabrin, H. Schultz, G. Whitten, Mosaic Generation for Under Vehicle Inspection, *IEEE WACV*, Dec. 2002.

[4] S. Gupte, O. Masoud, R. F. K. Matrin and N. P. Papanikolopoulos, Detection and Classfication of Vehicles, *IEEE Transactions on Intelligent Transportation System*, vol. 3, no. 1. Pp. 37-47, March 2002

[5] W. L. Hsu, S. H. Yu, Y. S. Chen and W. F. Hu, An Automatic Traffic Surveillance System for Vehicle Tracking and Classification, *IEEE Trans. on Intelligent Transportation Systems*, vol. 7, no. 2, 175-187, 2006

[6] S. Samadi, F. M. Kazemi, R. Mohamad, and T. Akbarzadeh, Vehicle Detection Using a Multi-agent Vision-based System, *Advances in Computer and Information Sciences and Engineering*, T. Sobh ed., 147-152, Springer, 2008

[7] Y. Iwasaki, A Method of Real-time Moving Vehicle Detection for Bad Environments Using Infrared Thermal Images, *Innovations and Advanced Techniques in System, Computer Science and Software Engineering*, K. Elleithy ed., 43-46, Springer, 2008.

[8] T. Wang, Z. Zhu, and C. N. Taylor, Multimodal Temporal Panorama for Moving Vehicle Detection and Reconstruction, *International Workshop on Video Panorama (IWVP)*, Dec. 2011.

[9] Y. Qu, T. Wang, and Z. Zhu, Vision-aided Laser Doppler Vibrometry for Remote Automatic Voice Detection, *IEEE/ASME Transactions on Mechatronics*, issue. 99, pp. 1-10, Nov. 2010.

[10] J. Y. Zheng, S. Tsuji, Panoramic Representation of Scenes for Route Understanding, *in Proc. 10-ICPR, IAPR*, 161-167June 1990.

[11] G. Flora, J. Y. Zheng, Adjusting Route Panoramas with Condensed Image Slices, *ACM Conf. Multimedia 07*, 815-818, Germany, 2007

[12] Z. Zhu, G. Xu, B. Yang, D. Shi, X. Lin, VISATRAM: A Real-time Vision System for Automatic Traffic Monitoring, *J. of Image and Vision Computing*, pp. 781-794, July 2000.

[13] J. Y. Zheng, X. Wang, Pervasive Views: Area Exploration and Guidance Using Extended Image Media, *ACM Multimedia Conference*, 05, 986-995, Singapore, 2005

[14] Y. Dedeoglu, B. U. Toreyin, U. Gudukbay, A. E. Cetin, Surveillance Using Both Video and Audio, *in Multimodal Processing and Interaction: Audio, Video, Text*, P. Maragos, A. Potamianos and P. Gros Eds., 143-156, 2008

[15] M. Cristani, M. Bicego, V. Murinon, Audio-visual Event Recognition in Surveillance Video Sequences, *IEEE Tran. on Multimedia*, vol. 9. no. 2, pp. 257-267, Feb., 2007.

[16] W. L. Hsu, H. Y. Liao, B. S. Jeng, and K. C. Fan, Real-time Traffic Parameter Extraction Using Entropy, *IEEE Proceedings-Vision, Image and Signal Processing*, vol.151, no.3, pp.194-202, June 2004.

[17] C. Stauffer, W.E.L. Grimson, Adaptive Background Mixture Models for Real-time Tracking", *CVPR*, Jun. 1999.

[18] Z. Zivkovic, F. van der Heijden, Efficient Adaptive Density Estimation Per Image Pixel for the Task of Background Subtraction, *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773-780, 2006

[19] C. Cortes and V. Vapnik, Support-vector network, *Machine Learning,* 273-297, 1995

[20] C-C. Chang and C-J. Lin, LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011