Distributed Video Networks for Incident Detection and Management

Mohan M. Trivedi, Ivana Mikić, and Greg Kogut
Computer Vision and Robotics Research Laboratory
University of California at San Diego
La Jolla, CA 92093-0407

Abstract

In this paper we describe a novel architecture for developing distributed video networks for incident detection and management. The networks utilize both rectilinear and omnidirectional cameras. It is recognized that robust and reliable segmentation of automobiles and shadows is critical in our application. We describe new segmentation procedure and present experimental results to support the basic feasibility and utility of the algorithms.

Keywords: Machine vision, traffic flow analysis, incident detection, video image analysis, segmentation

1. Introduction: Research Context

An incident is defined as "*an event that causes blockage of traffic lanes or any kind of restriction of the free movement of traffic*" [Ozbay 1999]. Examples of incidents include a stalled vehicle, accidents, debris, or chemical spill blocking a lane. All of them can have a disruptive impact on the normal, smooth flow of traffic leading to delays as well as secondary incidents. A report published by the Institute of Transportation Engineers, estimates the staggering costs associated with the congestion caused by incidents. The report estimates that, in year 2005, the "user delay costs" will climb to $50.5 billion (up from $9.2 billion in 1984); the wasted fuel in estimates are 7.3 million gallons (up from 1.4 million gallons in 1984); and the total user delay will add up to 6.9 million vehicle-hours (up from 1.3 million vehicle-hours in 1984. Desire to control and curtail such "costs" provides the main motivation and underlines the significance of our research efforts. The main goal of the overall research is help in the realization of a powerful and integrated traffic-incident detection, monitoring and recovery system. The system will have direct impact on reducing congestion on the highways. It will make travel safer, smoother, and more economical and will reduce wasted fuel and pollution. This framework offers several novel features to significantly improve the ability of existing technologies and algorithms to handle a wide range of traffic scenes. These are fusion of multiple sensors and sensor modalities, cooperation between sensor clusters, analysis of individual and group behaviors and data archiving. The overall research project can be described with the help of a futuristic scenario showing how an incident will be detected, verified and managed in the future. We can anticipate that in the future transportation infrastructures will utilize novel camera clusters, microphone arrays, mobile platforms, high-bandwidth wireless communication networks, and powerful computers. The objective of our research effort is to detect an incident, inform the relevant authority of the event while continuously monitoring the event, and provide arbitrary views and interface to the remote operators for decision-making.

2. Vision-Based Traffic Monitoring: Related Research

Vision-based traffic monitoring systems have been previously developed by several research groups [Koller 1991, Koller 1993, Huang 1993, Ferryman 1995, Betke 1996]. Some of these systems are robust to the problems of real-life, real-time tracking, including problems with occlusion, varying lighting conditions, and noisy video data. Also, some systems include high-level description of both cars and their behaviors. This allows classification of car types based on visual features and classification of basic car behaviors, such as lane changing or braking, based on trajectory information.

However, there has been little research in using multiple sensors and sensor modalities in traffic scenes. Most systems use single rectilinear CCD cameras, and use simple linear transforms to map from image to world coordinates. This limits the area over which objects can be tracked, and the accuracy of tracking. Also, cars are tracked from a single, fixed perspective, while the best perspective with which to view the scene may change with time of day or traffic density. Also, no current system provides a robust database system that allows historical or standing semantic queries of traffic data or viewing of historical scenes based on such queries. Current systems also use single, dedicated processors to analyze and record data, and don't provide the ability to distribute processing, select among an array of available sensors, or access to real-time or archived data at multiple remote locations [Bhonsle 1999].

3. Distributed Video Networks: Architecture

We envision a system that covers the highways and intersections with many sensor clusters (Figure 1.) that communicate with each other. Each cluster would include microphones, rectilinear and omni-view CCD cameras, infrared cameras and real-time range sensing cameras. Fusion of information from the sensors within each cluster and between different clusters would allow for monitoring of the traffic, recognition of individual behaviors and group behaviors (single car speeding vs. multiple cars involved in a high-speed chase), incident detection and intervention management. In addition to triggering appropriate responses, results from such analysis would be stored in a database. This would allow statistical analysis of past events and addition of standing queries for behaviors that were not defined at the time the system was designed. Having multiple calibrated cameras provides the ability to resolve some types of occlusion, and to provide more information about the 3D structure of the cars than is possible with a single camera. Omni-view cameras can provide a much broader range of coverage than standard rectilinear cameras, and could be used either in place of, or in conjunction with standard rectilinear cameras. Range scanning cameras may be very useful in resolving problems with shadows and occlusion. Infrared cameras could aid in tracking during nighttime, and also solve many occlusion problems by identifying the engine in each car. Mobile or zoom-pan-title cameras could be used to focus attention on a problem area, where extra resolution is needed, or to capture, and possibly read, the license plate number of a specific car [Hermida 1997]. The use of multi-sensor architectures [Weil 1998] can assure a proper coverage of all possible conditions of operation, thus satisfying the desired requirements in terms of system's global performance. Sensory information will be made available to the system over wireless networks.

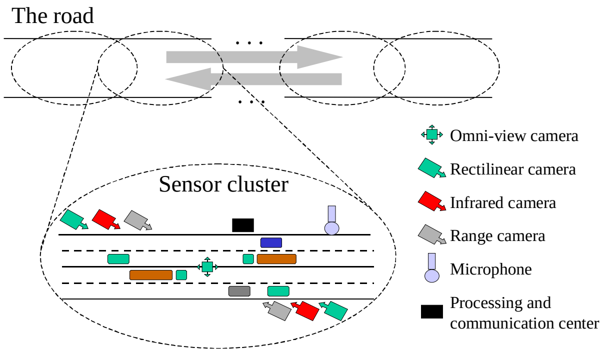


Figure 1. Components of a sensor cluster which provides multimodal sensory information useful for incident detection and management

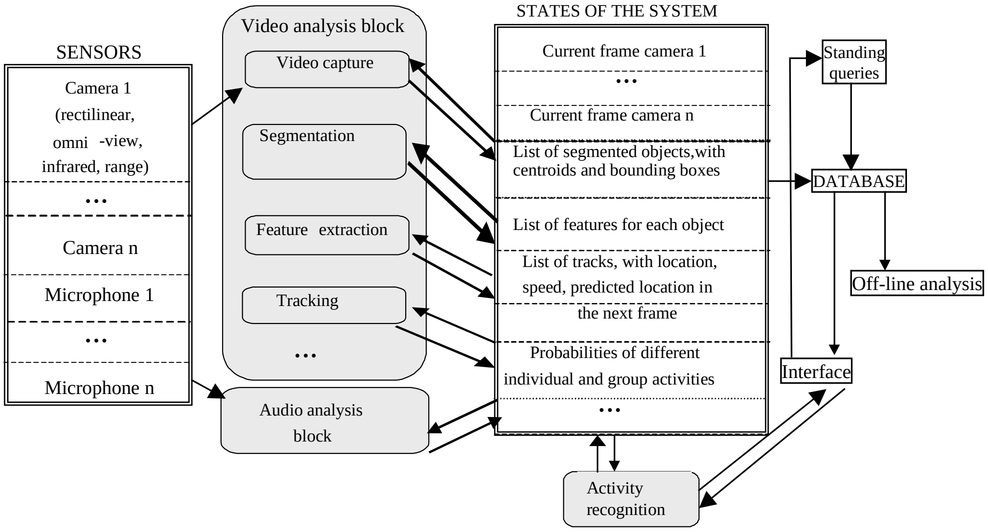


Figure 2. Sensory cluster block diagram. The five main components include: *sensors, processing layers, system states, database and the interface*

Figure 2 shows the block diagram that illustrates the design of the system. It has five major parts: *sensors, processing layers, system states, database and the interface*. System states contain data produced by processing layers, such as segments, tracks, video frames, etc. Each layer takes input from sensors or from the system states (outputs from other layers). Each layer produces results, which are included in the system states. For example, the video capture layer takes input from a camera and produces video frames. The segmentation layer takes video frames as inputs (and possibly predicted object positions that can be produced by the tracker) and outputs list of segmented objects with their centroids and bounding boxes. The activity recognition layer can take as input results from tracking and feature extraction layers as well as results from audio analysis block. It can output probabilities of certain activities, which are also considered states of the system. Some of the system states are stored in the database to enable detection of events of interest as well as off-line analysis of past events. A user interface would be designed that would simplify definition of events of interest and also include alarms and other triggers.

Such modular design provides great flexibility. New processing layers can easily be added and connected to appropriate existing layers without affecting other parts of the system. Different algorithms for the same task could be tested without any difficulty in a plug-and-play manner. For example, different trackers could be experimented with. One algorithm would use only calibration data and segmentation results, while a different algorithm may use color features of segments as well. Incorporating such an algorithm would involve adding feature extraction layer that would use video frames and segmentation results as inputs. Output of this layer would be used by the new tracker in addition to inputs used by the old tracker. The segmentation layer or the activity recognition layer would not have to change at all. This architecture is very convenient for dealing with multiple sensors. Some layers would operate on results that come from only one sensor (segmentation for example), while others would be responsible for integrating information from multiple sensors (3D tracking).

Both individual and group behaviors lend themselves to statistical analysis and classification techniques, due to the inherent unpredictability of driver behavior and the error associated with vision data. Techniques such as HMM-based classification, Bayesian inference, and statistical clustering have worked well with other computer vision applications and will likely make good tools with which to analyze traffic scenes. Statistical classifiers such as HMMs or Bayes nets can be trained to recognize specific behaviors, such as lane-changing, or deviations from a one of several "standard" behaviors. Clustering techniques could be used to analyze large amounts of data to determine use patterns of a section of freeway, and to recognize possible inefficiencies in traffic flow.

4. Novel Video Imaging Cameras: ODVS

Video networks in our approach utilize both the conventional rectilinear cameras as well as the omnidirectional vision sensors (ODVS) which offer a unique advantage of 360-degree coverage. ODVS consist of a hyperbolic mirror mounted above the lens of a rectilinear camera. This configuration provides a 360-degree field of view. The increased coverage reduces the number of necessary sensors.

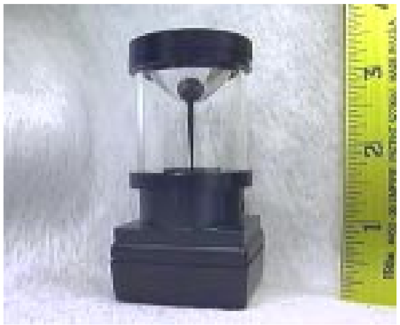


Figure 3. A compact Omnidirectional Vision Sensor (ODVS) for capturing full 360 degree views.

We have developed robust algorithms for efficient and accurate analysis of information acquired by a network of these ODVS [Ng et al 1999]. ODVS with their unique properties of *optical flow field* and *periodicity* make an ideal sensor for our application. ODVS have two flow fields, namely focus of expansion and focus of contraction. These are 180 degrees apart flowing in opposite direction. We use these unique properties for human tracking, vision modeling, and view synthesis [Ishiguro, trivedi1999, Ng et al 1999].

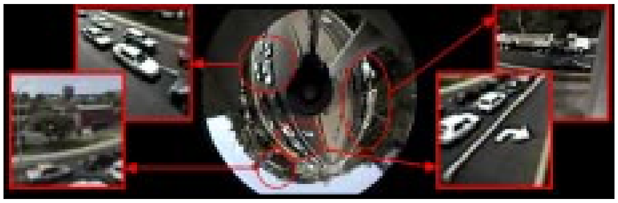


Figure 4. Four separate perspective views of a traffic scene from a single omnidirectional image. The perspective mappings can be used for any arbitrary view desired by an observer**.**

By applying 3D transformations both panoramic and perspective views can be generated. As an illustration, Figure 4 shows four separate perspective views of traffic flow instant directly from a single omnidirectional image. With multiple ODVS placed at close proximity, a virtual walk-through can be simulated by interpolation between camera views Generally, 3-D reconstruction by stereo is not sufficiently stable for practical use. On the other hand, recent progress of vision devices and the related computer interfaces enable us to use many cameras simultaneously. Multiple camera stereo using many cameras compensates the problem of matching between images and provides robust and stable range information.

5. Video Segmentation Module

Without using scene and object models, we can identify three sources of information that can help in detecting objects and shadows. The first is local, based on the appearance of the individual pixels. A point covered by a shadow gets darker compared to its appearance when illuminated. The second source of information is spatial: objects and shadows inhabit compact regions in the image, and the third is temporal: object and shadow positions can be predicted from previous frames.

We have found the diagonal model of pixel color change under shadow satisfactory (for details, see [Mikić 2000]):

$$\mu_{SH}^{i} = \mu_{IL}^{i} d_i$$
$$\sigma_{SH}^{i} = \sigma_{IL}^{i} d_i,\ i \in \{R, G, B\}$$

where $\mu_{SH}^{i}, \mu_{IL}^{i}, \sigma_{SH}^{i}$ and σ_{IL}^{i} are means and variances of shadowed and illuminated pixels for red, green and blue color components. This numbers d_i are constant for a given pixel and are determined from example data.

A fading memory estimator calculates background mean and variance for all pixel locations. Using the rules presented in the previous section, we derive statistics for same pixels when shadowed. Gaussian distributions are assumed for background and shadow pixels, and uniform distribution is assumed for foreground.

We start the segmentation by comparing the feature vector for each pixel (a three-dimensional vector of red, green and blue color components) to the mean at that location in the background model. If not significantly different, the pixel is classified into the background class. Otherwise, we assign to that location the a priori probabilities p_{BG}, p_{SH}, and p_{FG} of belonging to background, shadow and foreground classes, respectively. Then, we classify each pixel by maximizing the a posteriori probability of the class membership (C_1 = background, C_2 = shadow and C_3 = foreground):

$$p(C_i/\mathbf{v}) = \frac{p(\mathbf{v}/C_i)p(C_i)}{\sum_{j=1,2,3} p(\mathbf{v}/C_j)p(C_j)}$$

where **v** is the feature vector for a given pixel, $p(C_i)$ the a priori probability of occurrence of the i-th class at that location and $p(\mathbf{v}/C_i)$ the probability of the observed feature values given that the pixel belongs to the i-th class.

The majority of the pixels are classified correctly by the described appearance-based algorithm (73%, when compared to the hand-segmented images). However, object and shadow regions are very noisy due to misclassified pixels (See Figure 6a). The results can be significantly improved by imposing spatial smoothness. We investigated two approaches. First is simple post-processing by spatial filtering of the segmented images. We eliminate small gaps in foreground regions by performing one vertical and then one horizontal scan and assigning an encountered small line segment of non-foreground pixels to foreground if it is surrounded by foreground pixels in the direction of the scan. This is followed by morphological opening.

The second approach we investigated was performing an iterative probabilistic relaxation to propagate neighborhood information. In the first step, the a posteriori probability computations based on color are performed for all pixels. This is a local, appearance based computation. In the second step, we perform spatial propagation where the new class membership probabilities are computed for each pixel based on the results of the first step on the neighboring pixels. These are then used for a new computation of a posteriori probabilities in the first step and so on (Figure 5). The scheme converges quickly, and there is no noticeable change beyond the second iteration.

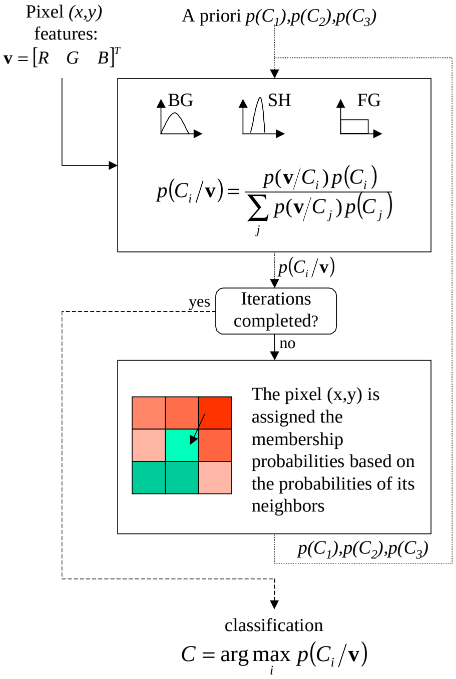
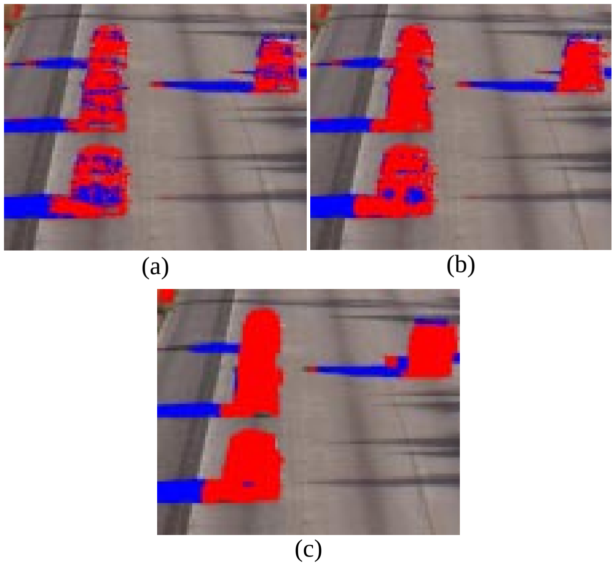

Pixel (x,y)
features:
$\mathbf{v} = [R\ \ G\ \ B]^T$
A priori $p(C_1),p(C_2),p(C_3)$
BG
SH
FG
$p(C_i/\mathbf{v}) = \frac{p(\mathbf{v}/C_i)p(C_i)}{\sum_j p(\mathbf{v}/C_j)p(C_j)}$
$p(C_i/\mathbf{v})$
yes
Iterations completed?
no
The pixel (x,y) is assigned the membership probabilities based on the probabilities of its neighbors
$p(C_1),p(C_2),p(C_3)$
classification
$C = \arg\max_i p(C_i/\mathbf{v})$

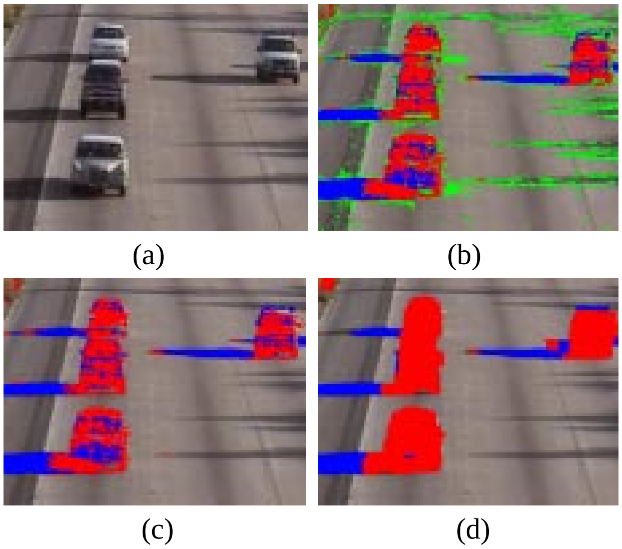
Figure 5. Iterative procedure that integrates appearance based and spatial information

We found that the results are slightly improved (78% of pixels correctly classified – see Figure 6b). However, there is still a need for post-processing that is of similar complexity to the post-processing described in the previous paragraph, which we used on original segmentation results. The final result is very similar (around 90% of pixels classified correctly – see Figure 6c). Also, performing these iterations reduces the speed and increases the memory requirements. We therefore conclude that the spatial smoothness is imposed most efficiently by a simple post-processing.


(a) (b)

(c)

Figure 6. Imposing spatial smoothness. (a) result of the color based segmentation. (b) result of adding a smoothing component to the iteration loop. (c) Result of post-processing of (a).


(a) (b)

(c) (d)

Figure 7. Moving shadow and object detection. (a) the original image frame. (b) Classification results after the second iteration. Red pixels are classified as foreground, blue as shadow and green as background. (c) Same as in (b), with background pixels not shown. (d) final result after post-processing by a spatial filter

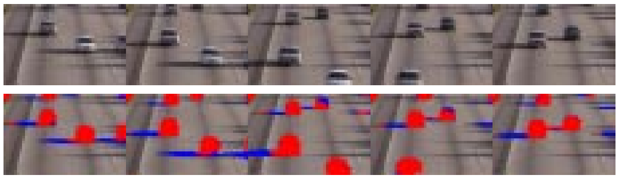


Figure 8. Five frames from the video of a traffic scene. Top row shows the raw video data and the bottom row shows the results of the algorithm

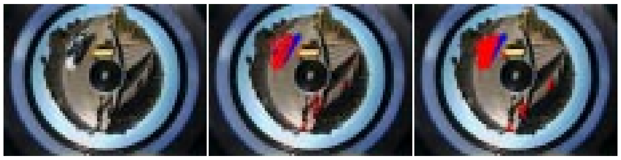


Figure 9. Results of the algorithm on the omniview image sequence. Left: original image, Middle: results based on pixel appearance, Right: results after postprocessing

6. Video Segmentation: Experimental Results

Figure 7 shows segmentation results for one frame from the video of a traffic scene. By correctly classifying shadows and flickering background pixels that simple background subtraction would classify as foreground, the accuracy of the calculated object locations is greatly improved, especially in scenes with long shadows. Note that static shadows are considered to be part of the background. Segmented shadows also provide an important clue for separating objects that are so close that they are segmented as one object. Often in those cases, the shadows of such objects will be distinct and help us separate the objects (see Figure 7d). Figure 8 shows results on several video frames. Figure 9 shows the result of the same algorithm applied to the omniview image sequence. The only difference is that the postprocessing step looks at radial rather than vertical lines.

There are a number of activities that we are currently pursuing to enhance the capabilities and performance of the above system. First, including temporal information could significantly improve the performance of the algorithm without much speed degradation. We could use predicted object locations to select a priori probabilities in the current frame.

Another important direction of future work is analysis of the relationship between the scene illumination and the parameters of the model that is used to derive shadow statistics given the statistics of a point when it is illuminated. As the algorithm adapts background statistics to the slow changes in the scene conditions, it could also collect statistics for shadow pixels it identified with high confidence and modify the parameters of the change rules (d_i-s). By independently measuring illumination at the scene, we should be able to build a database of these parameters indexed by the scene illumination and use it to recover from sudden changes in scene conditions.

7. Concluding Remarks

The main goal of the overall research is help in the realization of a powerful and integrated traffic-incident detection, monitoring and recovery system. It will make travel safer, smoother, and more economical and will reduce wasted fuel and pollution. Installing multiple sensors introduces several new issues into the system design, including handoff schemes for passing tracked objects between sensors and clusters, methods for determining the "best view" given the context of the traffic scene, and sensor fusion algorithms to best employ the strengths of a given sensor or sensor modality. Archiving some intermediate results of the analysis in a database system allows further analysis about the behavior of traffic and groups of cars in a variety of traffic conditions, as well as allow offline analysis of any incidents captured by the system and statistics on the observed properties of the traffic. [Bhonsle 1999] The Internet allows remote visualization of data without the need for specialized processing or digitization hardware. Robust user interfaces may be constructed which bring all the functionality of a traffic-monitoring system to a large class of users.

Acknowledgements

Our research is supported in part by the California Digital Media Innovation Program in partnership with the California Department of Transportation (Caltrans). We wish to thank our colleagues from the Computer Vision and Robotics Research Laboratory who are also involved in related research activities. We would specifically like to thank Dr. Kim Ng and Rick Capella for their contributions with Omnidirectional images.

References

[Betke 1996] Betke, M.; Haritaoglu, E.; Davis, L.S. Multiple vehicle detection and tracking. Proceedings of the SPIE vol. 2692, (25th AIPR Workshop. Emerging Applications of Computer Vision, Washington, DC, 16-18 Oct. 1996.

[Bhonsle 1999] S. K. Bhonsle et. al., “Complex visual activity recognition using a temporally ordered database”, 3rd International Conference on Visual Information Systems, Amsterdam, June 1999.

[Ferryman 1995] Ferryman, J.M.; Worrall, A.D.; Sullivan, G.D.; Baker, K.D. Visual surveillance using deformable modals of vehicles. Robotics and Autonomous Systems, vol. 19, (no.3-4), (Third International Symposium on Intelligent Robotic Systems. SIRS '95, Pisa, Italy, 10-14 July 1995.

[Hermida 1997] X. F. Hermida, F. M. Rodriguez, J.L.F. Lijo, F.P. Sande, and M. P. Iglesias, A system for the automatic and real time recognition of VLPs (vehicle license plate), Proceedings of International Conference on Image Analysis and Processing, Florence, Italy, 17-19 Sept. 1997, vol.2, pp. 552-9.

[Huang 1993] T. Huang, G. Ogasawa, and S. Russell, “ Symbolic traffic scene analysis using dynamic belief networks,” In AAAI *Workshop on AI in IVHS*, Washington D.C., 1993..

[H. Ishiguroa and M. M. Trivedi] H. Ishiguroa and M. M. Trivedi, "Integrating a Percpetual Information Infrastructure with Robotic Avatars: A Framework for Tele-Existance," *IEEE/RSJ Intelligent Robotic Systems Conference*, Korea, Oct. 1999.

[Koller 1991] D. Koller, N. Heinze, and H.-H Nagel, “Algorithmic characterization of vehicle trajectories from image sequences by motion verbs,” In *IEEE Conf. on Computer Vision and Pattern Recognition,* pages 90-95, Lahaina, Maui, Hawaii, June 3-6, 1991.

[Koller 1993] Dieter Koller, Joseph Weber, and Jitendra Malik, “Robust multiple car tracking with occlusion reasoning,” Technical Report UCB/CSD 93/780, Computer Science Division (EECS), University of California, Berkeley, 1993.

[Mikić 2000] I. Mikić, P. Cosman, G. Kogut, M. Trivedi “Moving shadow and object detection in traffic scenes”, ICPR 2000, Barcelona, Spain, September 3-8, 2000

[Ng 1999] K. C. Ng, H. Ishiguro, M. M. Trivedi, and T. Sogo, "Monitoring Dynamically Changing Environments by Ubiquitous Vision System," *IEEE Int. Workshop on Visual Surveillance,* June 1999.

[Ozbay 1999] K. Ozbay and P. Kachroo, *Incident Management in Intelligent Transportation Systems*, Artech House, 1999.

[Weil 1998] Weil, R., Wooton, J., and Garcia-Ortiz, A., “Traffic incident detection: sensors and algorithms,” *Mathematical and Computer Modelling,* vol.27, no.9-11, Elsevier, May-June 1998.