Behavior Interpretation from Traffic Video Streams

Pankaj Kumar
E-mail: kumar@i2r.a-star.edu.sg
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613

Surendra Ranganath
E-mail: elesr@nus.edu.sg
National University of Singapore
4 Engineering Drive 3
Singapore 117576

Kuntal Sengupta
E-mail: kuntal.sengupta@authentec.com
AuthenTec, Inc.
Post Office Box 2719
Melbourne, Florida 32902-2719

Abstract

This paper considers video surveillance applied to traffic video streams. We present a framework for analyzing and recognizing different traffic behaviors from image sequences acquired from a fixed camera. Two types of interactions have been mainly considered. In one there is interaction between two or more mobile objects in the Field of View (FOV) of the camera. The other is interaction between mobile objects and static objects in the environment. The framework is based on two types of a priori information: (1) the contextual information of the camera's FOV, in terms of the description of the different static objects of the scene and (2) sets of predefined behavior scenarios, which need to be analyzed in different contexts. At present the system is designed to recognize behavior from stored videos and retrieve the frames in which the specific behaviors took place. We demonstrate successful behavior recognition results for pedestrian-vehicle interaction and vehicle-checkpost interactions.

1 Introduction

Video stream based surveillance systems detect mobile objects, tracks them, analyzes their behavior and advanced systems take decision based on the analysis. Behavior understanding in image sequences requires establishing a relationship between low level image features of targets with high level symbolic descriptions of activities. For example we may want to detect that a person is trying to "access a restricted area". Such a behavior can be recognized by position and motion features of the target in the context of the restricted area. If the target is near the restricted area and is moving towards it, and subsequent to this event if the target is inside the restricted area, then the behavior of "accessing a restricted area" is recognized. Hence a behavior can be analyzed in terms of a temporal sequence of events. Events are usually spatio-temporal relationships between a target and the context. For example the event 'near the restricted area' can be defined in terms of the position of the target in relation to a polygon representing the restricted area.

Behavior recognition is dependent on the spatial context of the target. Most of the previous works have defined context in terms of the interaction of targets with static objects in the environment [15]. An elaborate and detailed discussion on the role of context in behavior analysis and video understanding can be found in [1, 2]. A methodology for recognizing multi agents behavior in the game of American football was presented by Intille and Bobick in [6, 7]. They present a probabilistic framework for representing and visually recognizing complex multi-agent action. Medioni *et al.* [3, 12, 13] have shown event detection and behavior recognition in videos taken from a single moving camera. The event recognition involves humans and vehicles and relies on optical flow to segment the mobile object from the background. Herzog (VITRA) [5] proposed a system to dynamically describe scenes with humans. The novelty of this work is in its application environment: a soccer stadium. The inference method is based on time interval logic, to describe temporal sequence of events, which are computed and typed separately. Galton in [4] generated complex descriptions of human actions based on a set of generic basic spatial and temporal propositions. In [14], Neumann states that symbolic descriptions must be linked with properties defined at the image level. An automatic surveillance system which performs labelling of events and interaction of humans and cars in an open car park was presented by Ivanov *et al.* in [8].

Our work is similar in approach to [13, 15]. The system is designed for stationary cameras. We have added extra features to the description of context so as to include other mobile objects in the description of the context. This facilitates the analysis of interaction between two or more moving targets. There is a clear link between high-level event description and low-level target features. The robustness of event and behavior detection has been improved by consistency check.

Hence forth the paper is organized as follows: Section 2 gives and overview of the system and its different modules. Section 3 briefly describes the mobile object detection and tracking algorithm. The features like position, velocity etc. are translated to the world co-ordinate system by using camera calibration. This is discussed

0-7803-8125-4/03/$17.00

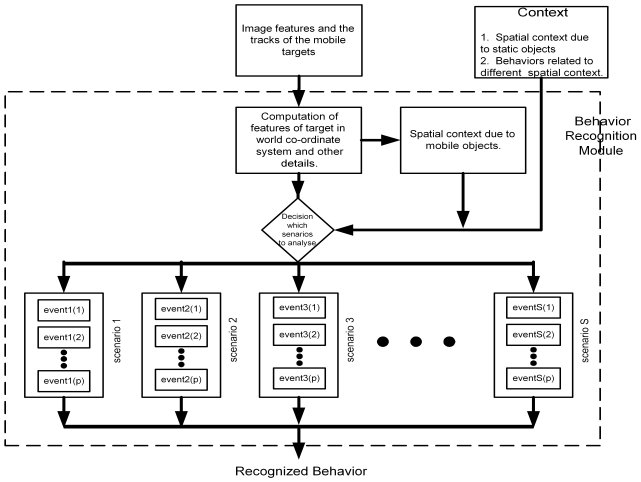


Figure 1. A schematic diagram of the behavior recognition system. The inputs to it are the features of targets, spatial context and the different behavior scenarios for different spatial contexts. The output is the recognized behavior and the frames in which it occurred.

in Section 4. Section 5 discusses some of the target features used in representation of the targets. Section 6 explains how the contextual information are programmed and used. Events and Behavior are discussed in Sections 7 and 8. In Section 9 we show the successful working of our system on example video streams and finally Section 10 concludes the paper.

2 System Overview

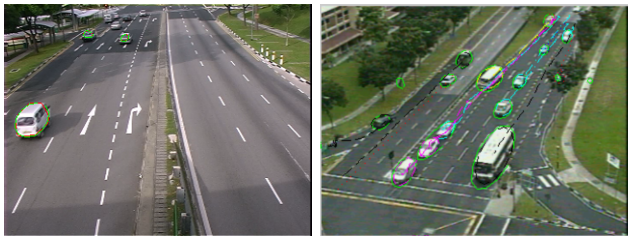
Figure 1 gives a schematic diagram of the various components of the behavior recognition module and the flow of data. The behavior recognition system takes two types of inputs:

1. The shape, position, motion, and track of the targets obtained from the motion detection and tracking module.
2. The spatial context of the various objects present in the FOV obtained from the user.

A priori knowledge of the relation between context and behaviors and description of behaviors in terms of events is programmed into the system. Based on contextual information a decision is made about which scenarios to analyze. For a given context only a subset of scenarios are analyzed because we do not expect all behaviors to occur in a context. Once the context is known the different types of behaviors that need to be considered is significantly reduced. The output of the behavior recognition module is the recognized behavior and the frames in which the specific behavior took place.

3 Motion Detection and Tracking

The moving objects in an image sequence can be detected in real-time using background subtraction. Our technique is capable of modelling the background even in presence of foreground objects in the images and can detect and remove shadows. Some details of this algorithm can be found in [11, 9]. After segmentation the foreground pixels are grouped to form 8-connected blobs. The convex hull of these blobs are then approximated by an ellipse. We use Kalman filter and a dynamic programming based pattern matching technique to achieve robust tracking [10]. Figure 2 shows shadow detection and multi-body tracking results. Our main focus is behavior analysis, therefore we dispense with the details of background modelling, segmentation, feature extraction, and tracking. These can be found in [10].



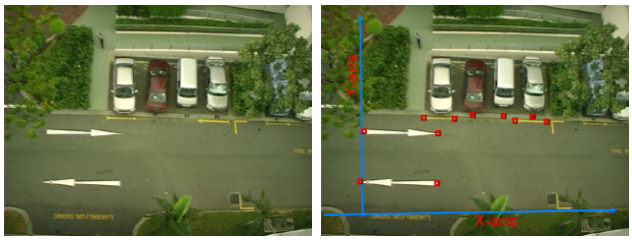
(a) (b)

Figure 2. Image (a) shows the result of shadow suppression and foreground extraction algorithms and (b) shows the results of the multi-body tracking algorithm used in our system.

4 Camera Calibration

Working in world co-ordinates is better than image co-ordinates as many ambiguities can be resolved. For example, perspective foreshortening gives an erroneous perception of target motion in the image plane. The targets closer to the camera appear to move faster than the targets further from the camera even if their ground speeds are similar. To translate the measurements in image co-ordinates to measurements in world co-ordinates we need to know the camera parameters. We assume that the camera is placed sufficiently high i.e. at least ten times higher than the height of the targets. This assumption allows us to consider the moving targets as flat moving patches on the ground plane. We apply the geometry of the planar world, and use measurements from the 3D world for camera calibration. The world co-ordinate system is placed so that the Z axis is aligned with the ground surface normal. Thus, any points on the ground plane would have co-ordinate values $[X_w, Y_w, 0, 1]^T$. The perspective transformation equation for a pin hole camera model can now be written as

$$\begin{bmatrix} x_i \\ y_i \\ \lambda \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ 0 \\ 1 \end{bmatrix} \quad (1)$$



(a) (b)

Figure 3. Image (a) shows a frame from an image sequence used for behavior analysis. Image (b) shows the same frame with X and Y axis aligned with the ground plane in blue. The points marked red are used for camera calibration. The world co-ordinates of these red marks are obtained by measurements on the ground.

which simplifies to

$$\begin{bmatrix} x_i \\ y_i \\ \lambda \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{14} \\ p_{21} & p_{22} & p_{24} \\ p_{31} & p_{32} & p_{34} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix} = P \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix} \quad (2)$$

The 3×3 matrix P is obtained by using manually selected points in the image and their corresponding measurements in the world co-ordinate system. Figure 3(a) shows a frame from a test sequence of images and 3(b) shows the X and Y axis of the world co-ordinate system. The points marked red are used in computation of the P matrix. To find P we need a least four points. We pick more than the minimum number of points and use a least squares estimate to solve the over-constrained linear equations and filter out noise due to errors in measurements. Using this technique of converting measurements from image co-ordinates to world co-ordinates we are able to detect vehicle speeds within an error of $\pm 5\%$. This error range is obtained by comparing the speed measurement from the speedometer of the vehicle as ground truth and the speed measured from the tracking system as observation. This high accuracy of speed detection makes it possible to detect acceleration and deceleration of the targets.

5 Target Features

The targets have a set of features described by their spatial and temporal parameters in 2D image space and also in 3D world co-ordinate space. Some of the features used in representation of targets are:

1. Size: the major and minor axis of the ellipse approximating the convex hull of the target.
2. Position: the centroid of the target.
3. Velocity of the target obtained from the Kalman filter tracking the target centroid.
4. Target type: at present we have four types of targets (a) pedestrians (b) motorbikes (c) cars (d) heavy vehicles (buses and trucks). This classification is done based on the size, velocity, and position of the targets. During classification size is given the highest weightage followed by velocity and finally position.
5. Track of the target.
6. Color information of the target in the form of histogram of the different color channels.

The target properties are used for event detection and to define context for interaction with other mobile objects. Figure 4 shows different levels of target properties. The numerical descriptors of higher level properties are computed from lower level image features. The world co-ordinate velocity and acceleration can be computed using the temporal information of the frames, camera calibration and the Kalman filter estimates of target velocity in image plane.

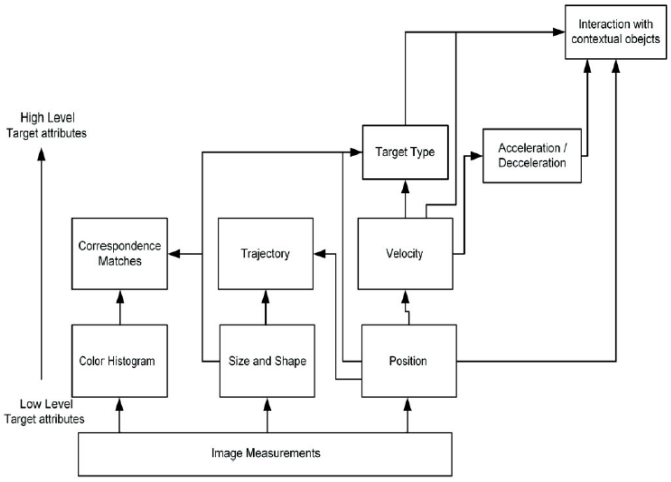


Figure 4. This figure shows the different levels of target attributes. The high level target attributes are obtained from low level image measurements, tracking, and target classification.

6 Context

Context plays a very important role in detection of events. The contextual information of the scene is provided by the operator and needs to be done once for a given surveillance setup. Context is defined by the spatial-temporal properties of static objects in the environment and by the zone of influence (ZOI) of mobile objects. Contextual information governs the different types of predefined scenarios of events that needs to be analyzed for different behavior recognition. An example is the context of a checkpost, which checks the entrance of unauthorized vehicles. The behavior of interest would be improper access to the restricted area, or detection of a malfunctioning checkpost.

Static objects which form a part of the context are defined geometrically by polygons and attributes like *name, function, time of normal interaction, status* etc. Table 1 gives the attributes of checkpost 1 as shown in Figure 5.

This figure shows an example of a scene with some static contextual objects, which form the context for recognition of behaviors at a checkpost. The objects are:

i. Checkpost 1 for vehicles entering the restricted area and checkpost 2 for exiting vehicles.
ii. Areas for interaction 3 and 4 with checkposts 1 and 2, respectively.
iii. Cash card machine 5.

When a vehicle enters the area for interaction (AFI) of a checkpost, the system analyzes the scenarios of events related to the context of checkpost. There are different possible behaviors in this context and each is defined by a temporal sequence of events.

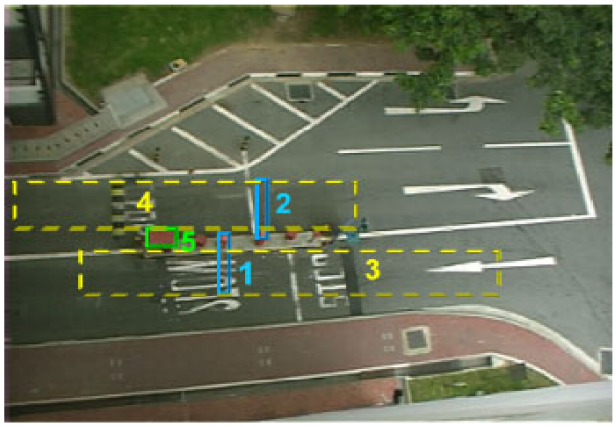


Figure 5. This figure highlights the different contextual elements in camera's FOV for description of static context that underlies behavior recognition. 3 and 4 are AFI of checkposts 1 and 2. The checkposts are represented by thin rectangular regions. Element 5 is a cash card machine used for paying the parking cost.

name	Checkpost
function	to temporarily stop vehicles
normal interaction time	5 seconds
geometry	rectangle [(115 137), (116 160)]

Table 1. This table gives the attributes of a static contextual object, the checkpost 1, as seen in Figure 5.

To recognize behaviors which involve interaction of two or more mobile objects, we define context which arises when two or more targets come in proximity with each other. Proximity of targets is determined by the normalized area of overlap of ZOI of the targets. ZOI of a target is defined as the outer ellipse whose center and orientation is the same as the target's but whose major and minor axes are 1.6 times that of the approximating ellipse. This value is heuristically chosen after some experimentation. When two or more targets are close to each other then we look for events where their relative velocities are dangerously high. The relative velocity of the targets is obtained by vector subtraction of the measured velocity of interacting targets.

7 Events

Events are usually described by the spatio-temporal relationship between targets and contextual elements or with other targets. Events can also be due to some properties of the high level target attributes. For example, if we want to detect 'speeding of cars'then its measured speed is compared with the upper limit of speed provided by the operator. If the measured speed is greater than the speed limit provided by the user then the event 'car is speeding'is detected.

Measurements from visual sensors are usually erroneous; therefore the system should be robust to errors. To do this we look for temporal consistency in detected events. Temporal consistency is measured by confidence factor κ. In a given context all events that can take place are associated with the target with an initial confidence factor of zero. When a specific event is detected then its confidence factor is increased by 0.2 and κ for other events is decremented by 0.2. The confidence factor has a floor value of 0 and maximum value of 1, i.e, once κ reaches a value 1 or 0 then it is not further incremented or decremented. Following are some examples of events we considered in our experiments:

1. **Moving towards checkpost**: this event is detected when the current distance between the target and checkpost is greater than the distance between the target and checkpost in the next frame.
2. **Stopped in front of the checkpost**: target is in the AFI of a checkpost and the speed of the target is less than a threshold.
3. **Crossing the checkpost**: the distance between target and checkpost is almost zero but the speed is above a threshold.
4. **Moves away from the checkpost on the other side of the checkpost** : the direction of velocity is same as before but the distance between the target and checkpost is increasing.
5. **Moves away from the checkpost on the same side of the checkpost**: the direction of velocity is reversed and the distance between the target and checkpost is increasing.
6. **Moves out of the AFI of a checkpost**: the current position of the target is within the AFI of a checkpost but the velocity is directed towards moving out of the AFI of the checkpost.
7. **Crosses the checkpost outside the AFI of the checkpost**: the target is outside the AFI of a checkpost and is crossing the checkpost. The protocol for recognition of the event of crossing the checkpost is the same as 3.

8 Behavior Analysis

A behavior is defined as a sequence of events, with or without temporal constraints on the order of event occurrence. Behavior analysis can be as simple as detection of a single event, e.g. a car is speeding or can be a complex sequence of multiple events, e.g. a car is entering a restricted area violating the checkpost norms. Given the context of the vehicle different behaviors are analyzed. We do a case study to illustrate how the whole system works. We consider the example of a vehicle entering AFI of a checkpost. In this context, the following behaviors were analyzed (Each of these behaviors is defined by sequence of events as follows):

1. **Normal crossing of checkpost**
   (a) Target moves towards the checkpost
   (b) Target stops in front of the checkpost
   (c) Target moves towards the checkpost
   (d) Target crosses the checkpost
   (e) Target moves away from the checkpost on the other side of the checkpost
   (f) Target leaves the AFI of the checkpost.
2. **Breakdown of checkpost or breakdown of a vehicle in front of a checkpost**
   (a) Target moves towards the checkpost
   (b) Target stops in front of the checkpost for more than the normal time of interaction with the checkpost
   (c) There are more vehicles stopping in the AFI of the checkpost.
3. **The target avoids the checkpost and backs off**
   (a) Target moves towards the checkpost
   (b) Target stopped before the checkpost
   (c) Target moves away from the checkpost on the same side of the checkpost
   (d) Target leaves the AFI of the checkpost.
4. **Vehicle trying to gain illegal access to the restricted area by moving on pedestrian's area**
   (a) Target moves towards the checkpost
   (b) Target moves out of the AFI of the checkpost i.e. outside the road region onto pedestrian's path
   (c) Target crosses the checkpost outside the AFI of the checkpost
   (d) Target moves away from the checkpost on the other side of the checkpost.

For behavior recognition we a compute recognition factor Φ, for each scenario, which is the sum of the confidence factor, κ of each event indexed by i in the behavior j, divided by N_j the total number of events in that behavior

$$\Phi_j = \frac{\sum_{i=1}^{i=N_j} \kappa_i}{N_j} \tag{3}$$

The behavior which yields the highest value of Φ is considered the recognized behavior. To increase the discrimination of behavior recognition a higher weightage can be given to more crucial events and lower weights to the less significant events. For example in the case of the behavior of a 'vehicle avoiding the checkpost and backing off' the most crucial event is, 'target moves away from the checkpost on the same side of the checkpost'. An example of a common and hence less significant event is 'target moves towards the checkpost', this event is common to all behaviors in the context of a checkpost.

9 Results

We show the results of our behavior recognition system in two different contexts. One is for interaction between two mobile objects and another is for interaction between mobile objects and static objects of the environment. In the results the different targets have been successfully classified into their respective classes and there behavior correctly annotated with textual remarks. Figure 6 shows correct detection of a dangerous interaction between a pedestrian and a vehicle. The targets are in close proximity with each other and their relative velocity is high. The system correctly analyzes this behavior to be dangerous and all such behaviors in the video stream were correctly detected. In Figure 7 we show the recognition of behaviors of vehicles at a checkpost. All the possible behaviors at the checkpost as discussed in Section 8 were correctly analyzed and classified by the recognition factor Φ. The figure captions give further details of the results.

10 Conclusions

In this paper we have described a behavior interpretation system for traffic video streams. The system is based on the analysis of 2D image features and 3D position and motion features. The a priori knowledge of context and predefined scenarios is used in behavior recognition. The problem of imprecision and uncertainty due to errors in signal processing and image feature measurements have been ameliorated by introducing a new parameter κ for confidence measure. This confidence factor is based on the temporal consistency in event detection. We have demonstrated successful, high accuracy and robust behavior recognition and object classification results. All the results are on real life traffic video streams. Almost 100% the behaviors recognition in two different contexts have been achieved for different behavior recognition.

References

[1] F. Bremond and M. Thonnat. Issues of representing context illustrated by video-surveillance applications. *International Journal of Human-Computer Studies Special Issue on Context*, 48:375–391, 1998.

[2] N. Chleq, F. Bremond, and M. Thonnat. *Video-based Advanced Systems Monitoring*, pages 106–116, Chapter: Image Understanding for Prevention of Vandalism in Subway Stations. Kluwer Academic Publishers, 1999.

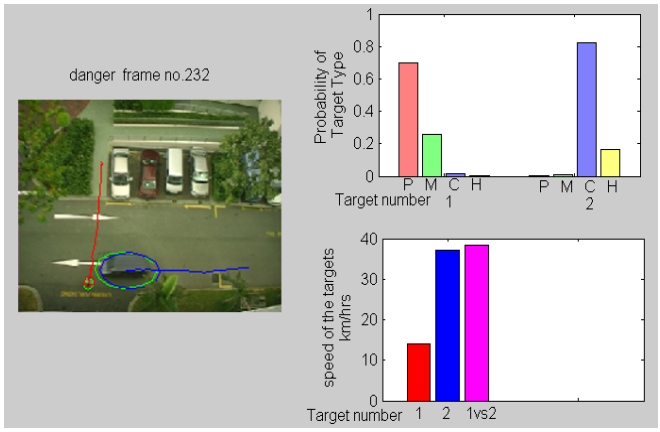


Figure 6. This figure shows the result of detecting dangerous behavior between two targets. The bar graph on the top right shows the degree match for classification of the target type. There are two targets in the FOV. Target #1 is a pedestrian and #2 is a van which has been classified to be of type car. The four classes of objects are Pedestrian (P), Motorbike (M), Car (C), and Heavy vehicle (H). The bar graph on the bottom right shows the measured speeds of targets 1, 2, and their relative speeds (1vs2), with colors red, blue, and pink, respectively. The relative speed of the targets is very high and they are in close proximity to each other. Therefore it has been detected as a dangerous interaction in frame number 232.

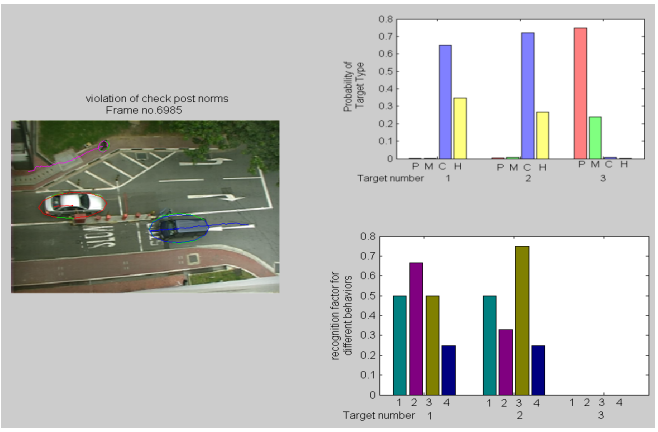


Figure 7. This figure shows the results of detecting different behaviors of vehicles at checkposts. Targets 1,2, and 3 shown in a frame by red, blue, and pink have been correctly classified as car, car, and pedestrian, respectively as shown by the top right bar graph in this figure. The results of behavior analysis of the targets is shown by the bottom right bar graph. The behavior types 1, 2, 3, and 4 are the same as discussed in Section 8. Target #1 indicated by the red ellipse has stopped at the check post for unusually long time. This is correctly detected by the system as the recognition factor for behavior #2 is highest of all. Target #2 indicated by the blue ellipse is avoiding the check post by backing off. This behavior has also been correctly detected as the recognition factor in this case is highest for behavior #3. Target #3 indicated by the pink ellipse is detected as a pedestrian and hence not analyzed for behavior recognition in the context of checkpost.

[3] I. Cohen and G. Medioni. Detecting and tracking moving objects for video surveillance. In *Computer Vsion and Pattern Recognition*, pages 2319–2325, Fort Collins, Colorado, 1999.

[4] A. Galton. Towards an integrated logic of space, time and motion. In *Proc. Int'l Joint Conf. Artificial Intelligence(IJCAI)*, pages 1550–1555, August 1993.

[5] G. Herzog. Utilizing interval-based event representation for incremental high-level scene analysis. In *4th International Workshop on Semantics of Time, Space, and Movement abd Spatio-Temporal Reasonning*, Chateau de Bonas, France, 1992.

[6] S. Intille and A. Bobick. A framework for recognizing multi-agent action from visual evidence. In *Proceedings of the National Conference on Arftificial Intelligence*, pages 518–525, July 1999.

[7] S. Intille, J. Davis, and A. Bobick. Real time closed-world tracking. *Computer Vision and Pattern Recognition*, June:697–703, 1997.

[8] Y. Ivanov, C. Stauffer, A. Bobick, and W. Grimson. Video surveillance of interactions. In *2nd International Workshop on Visual Surveillance*, pages 82–89, Fort Collins, Colorado, June 1999.

[9] P. Kumar, , S. Ranganath, and H. Weimin. Bayesian network based computer vision algorithm for traffic monitoring using video. In *Proccedings of The IEEE 6th International Conference on Intelligent Transportation Systems*, Shanghai, china, October 2003.

[10] P. Kumar, S. Ranganath, and K. Sengupta. An efficient scheme for robust multi-body tracking. NUS TECH. Report, February 2003.

[11] P. Kumar, K. Sengupta, A. Lee, and S. Ranganath. A comparative study of different color spaces for foreground and shadow detection for traffic monitoring system. In *Proccedings of The IEEE 5th International Conference on Intelligent Transportation Systems*, pages 100–105, Singapore, September 2002.

[12] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. In *DARPA Image Understanding Workshop*, Monterey, November 1998.

[13] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):873–889, August 2001.

[14] B. Neumann. *Semantic Structures: Advances in Natural Language Processing*, pages 167–206, chapter: 5. Hillsdale, N.J. Lawrence Erlbaum, 1989.

[15] N. Rota and M. Thonnat. Video sequence interpretation for visual surveillance. In *3rd IEEE International Workshop on Visual Surveillance*, pages 59–68, Dublin, Ireland, July 2000.