Multi-perspective Video Analysis of Persons and Vehicles for Enhanced Situational Awareness*

Sangho Park and Mohan M. Trivedi

Computer Vision and Robotics Research Laboratory
University of California at San Diego, La Jolla, CA, USA
{parks, mtrivedi}@ucsd.edu

Abstract. This paper presents a multi-perspective vision-based analysis of people and vehicle activities for the enhancement of situational awareness. Multiple perspectives provide a useful invariant feature of object in image, i.e., the footage area on the ground. Moving objects are detected in image domain, and tracking results of the objects are represented in projection domain using planar homography. Spatio-temporal relationships between human and vehicle tracks are categorized to safe or unsafe situation depending on site context such as walkway and driveway locations. Semantic-level information of the situation is achieved with the anticipation of possible directions of near-future tracks using piecewise velocity history. Crowd density is estimated from the footage in homography plane. Experimental data show promising results. Our framework can be applied to broad range of situational awareness for emergency response, disaster prevention, human interactions in structured environments, and crowd movement analysis in wide-view areas.

1 Introduction and Motivation

There has been a growing interest in the society and industry to make sensor-based systems enhance the safety and efficiency of human inhabited environments. Enhanced situational awareness is one of the key issues in developing intelligent infrastructures for safer environments. In order to develop automatic situational awareness system, it is important to understand how people interact with each other and with the environment itself. It will be useful to detect, represent and estimate what kind of events are occurring or about to occur in the monitored site. Pedestrian safety and crowd behavior analysis are good examples.

In this paper, we present a methodology for multi-perspective vision-based analysis of human interactivity with other persons and vehicles for enhanced situational awareness. It belongs to more general research problems of analyzing and recognizing human behavior in active environments. We present our methodology in the context of pedestrian safety and crowd monitoring domain.

* This research was supported in part by the NSF RESCUE ITR-Project and US DoD Technical Support Work Group (TSWG).

S. Mehrotra et al. (Eds.): ISI 2006, LNCS 3975, pp. 440–451, 2006.

Several research issues have been addressed in the context of behavior analysis when visual modality is used as the main source of information. First of all, the vision-based system is required to distinguish pedestrians versus vehicles and their typical movement patterns, respectively. Detection of invariant features from raw data is critical for the purpose. It is also desirable to locate each moving objects (i.e., persons or vehicles) and to effectively map them on the world coordinate system of the site of interest. Extraction and formation of semantic information from raw video signal is at the heart of the *situational awareness* of the system.

There has been active research effort for vision-based analysis of human activity in computer vision including video surveillance, human-computer interaction, virtual reality, choreography, and medicine. Reviews of general research on vision-based understanding of human motion can be found in [1, 3].

Most of outdoor human monitoring systems have been developed under certain specific environmental contexts: i.e., specific time, place, and activity scenarios involved in the situation [4, 5, 9]. Examplar surveillance systems have been either based on track analysis [6, 8, 10] or body analysis [4]. Track-level analysis is usually applied to wide-area surveillance of multiple moving vehicles / pedestrians in open space such as a parking lot or a pedestrian plaza. In some wide-area surveillance situations, coarse representation of human body in terms of a moving bounding box or an ellipse may be enough for tracking [6]. Other researchers have applied more detailed representation of a human body such as a moving region or a blob [8, 10]. Velastin et al. [10] estimated optical flow to compute the motion direction of pedestrians in subway environments. Body-level analysis usually focuses on more detailed activity analysis of individual persons.

Another important categorization of examplar systems is related to indoor vs. outdoor setup. Comparing to indoor environments, outdoor environments have a lot of environmental variations such as weather change, time shift from morning to evening, moving backgrounds, etc. Outdoor surveillance systems have to deal with those variations, and robustness is still an issue in outdoor surveillance. Most of the outdoor survillance systems apply track analysis due to the limited image resolution, because the wide field of view (FOV) for outdoor surveillance usually limits the resolution of person appearance to relatively low-resolution images. One of the recent developments in video surveillance is the usage of distributed system to cover multiple monitored scenes with various FOV's. Most of the research mentioned above mainly focuses on recognition of human activity, i.e., human-human interactions. Recognition of human-machine interaction in outdoor environments such as human-vehicle interaction has not been actively addressed. This paper presents a new framework to analyze human activity and interaction with vehicles as well as other humans for the enhancement of automatic situational awareness.

The rest of the paper is organized as follows: Section 2 summarizes our approach to the problem of vision-based transportation safety. Section 3 articulates new concepts and methods for spatio-temporal analysis of tracks. Section 4 explains the site modeling and sensor distribution in a real-world environment.

Section 5 describes the method to track multiple moving objects. Section 6 shows the representation of tracked objects in a world coordinate systems. Experimental results and concluding remarks follow in Sections 7 and 8, respectively.

2 System Overview

Our system uses multiple cameras with different perspectives and analyzes the visual information at multiple levels. At gross level, we represent each moving object as a trajectory point of the center of gravity of the object. Track of the moving object is formed along the video sequence. At detailed level, we represent the object in terms of its footage area on the ground in order to estimate the invariant size of the object. We observe that the approximate size of the object's footage area is invariant to translation and rotation, unless the object falls or flips over. Planar homography is used to locate the object's footage position on the world coordinate system. At semantic level, the interaction among persons and vehicles is analyzed. Contextual information including site model and activity scenario is integrated at the semantic level. The concepts of *spatio-temporal interaction boundary* and *time to collide* are introduced to represent and predict various interaction patterns among moving objects.

Foreground moving objects are detected and segmented by background subtraction. Tracking of each object is performed by data association of foreground object appearances between consecutive frames using a constrained expectation-maximization (EM) learning method. Image appearance of the same object varies significantly according to camera perspectives as shown in Fig. 1. Therefore, even though the tracker keeps following the same object, it does not classify the object category into vehicle or person. For the reliable classification of object types, we need to estimate the invariant size of the object. We rectify the images and

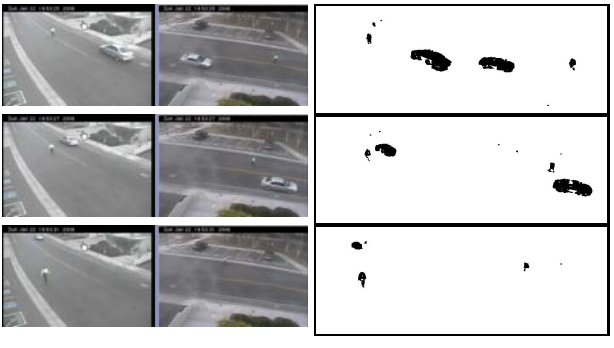


Fig. 1. Perspective effect on image appearance. The same object appears very different at different time frames.

map the objects to the world coordinate system using planar homography. We estimate the footage area and location of an object in the world coordinate system by using multiple-view geometry. Fig. 2 depicts the process of estimating

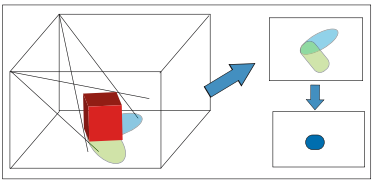


Fig. 2. Schematic diagrams for footage area estimation using multiple planar homography. Multiple views of the same object are projected by planar homography, and the intersection of the projected images are used as the object's footage region on the ground.

the footage area. Multiple views of the same object are transformed by planar homography and the intersection of the projected images are used as the footage region of the object on the ground.

Planar homography assumes all the pixels lie on the same plane (i.e., the ground plane in 3D world.) Pixels that violate this assumption result in mapping to a skewed location on the projection plane. By intersecting multiple projection maps of the same object, we can estimate the object's common footage region that observes the assumption. In order to reduce the false estimation of the footage region possibly caused by other adjacent objects, we compare the color histograms of the raw image regions in the multiple views using back-projection of the homography. Bhattacharyya distance measure is used to compare the histogram profiles.

Moving object's true velocity (i.e., speed and direction) in 3D world coordinate system is estimated at the projection plane. The velocity of a moving object determines the object's reaching boundary in a given time. This reaching boundary defines the *spatio-temporal interaction boundary* of the object. If there exists a foreign object at the vicinity of a moving object, the estimation of *time to collide* becomes important; the time of arrival or the time to collide has significant implication regarding safety in transportation systems. In the next section, we discuss the spatio-temporal analysis of tracks.

3 Track Analysis in Spatio-temporal Domain

The spatio-temporal characteristics of the interaction boundary provides a useful tool to analyze human-human interactions as well as human-vehicle interactions in terms of time, velocity, and distance as described below.

It is observed that the distance x that can be reached within time t is proportional to velocity v according to dynamics as formulated in Eqn. 1. This implies that, with higher velocity, the range of impact of interaction can reach farther within a given time period.

$$x = v \times t \tag{1}$$

where v has directional component.

Humans are subconsciously aware of this fact, and anticipate the consequence of speed with respect to safety. In the case of human movement, the direction of

motion is ambiguous due to the possibility of agile body motion. Therefore, we make the directionality broader, resulting in a circular interaction boundary. It means that we model the interaction boundary as a circular shape with radius proportional to track velocity. In circular interaction boundary, the velocity v is replaced by speed $|v|$ and the reaching distance is represented in terms of distance $|x|$. In the case of vehicle movement, the direction of motion would be more deterministic depending on the driver's intent. Therefore, it would be more realistic to shape the interaction boundary of a vehicle more directional depending on velocity. In the current paper, we assume the circular boundary for both humans and vehicles for simplicity.

The spatio-temporal interaction boundary can be categorized into *interaction potential* from *interaction region*. Both concepts are expressed in terms of spatial boundary that surrounds a moving object, but the former is related to anticipatory interaction, while the latter indicates actual interaction.

We derive the effective radius of *interaction potential*, r_p, of a moving object:

$$r_p \approx |v| \times t \tag{2}$$

We model the radial shape of the interaction potential as a probability distribution function (PDF) in terms of a 2D Gaussian distribution, $R = N(\mu, \sigma)$, truncated by the circle of radius r_p. The actual parameters of the PDF can be learned with training data. A similar formulation of pedestrian's moving directionality was proposed by Antonini and Bierlairein [2]. However, their method using manual tracking is computed on image plane from a single perspective and is perspective-dependent, whereas our approach is computed on projection plane using planar homography and is view-independent.

As seen above, the spatial and temporal analysis of tracks are highly correlated. We will present more details about the spatio-temporal analysis at track-level modeling of human/vehicle activity in later sections in this paper. The significance and connotation of a specific human track pattern depends on the site context: driveway, walkway, crowded area, etc. The relation between human track patterns and site context is mediated by policy, by which we mean which activity needs to be regulated/monitored and which activity is allowed. In this paper, we are interested in the combination of spatial and temporal relations in the site context as summarized in Table 1. The table on the left shows the *spatial* site context of human activity, while the table on the right shows the *temporal* site context of interactivity between two objects. A person may stay, walk, or run at different sites such as walkway, driveway, or specific region of interest (ROI) at a bus-stop area or a building entrance zone. ◯, △, and × denote normal, cautious, and abnormal track patterns, respectively. Cautious or abnormal pattern at a specific ROI depends on the duration of stay and the site context. *Interaction region* is the actual boundary in which interaction between two objects occurs. We define the interaction region between two objects (i.e., person or vehicle) to be the intersection of the two interaction potentials.

Diagrams for the track-level analysis of human activity and interactivity are shown in Fig. 3. The figures from the left to the right show a track in 3D spatio-temporal space in xyt dimensions, the track's interaction potential boundary in

Table 1. Track vs. site context. ◯, △, and × denote normal, cautious, and abnormal track patterns, respectively. Cautious or abnormal pattern at a specific ROI depends on duration of stay and site context.

person / site	stay	walk	run
walkway	◯	◯	×
driveway	×	◯	◯
ROI	◯, ×	◯	×

object 1 / object 2	stay	slow	fast
stay	◯	◯	△
slow	◯	△	×
fast	△	×	×

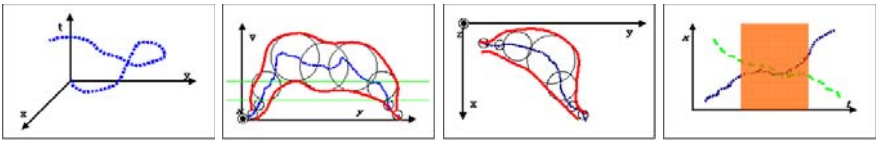


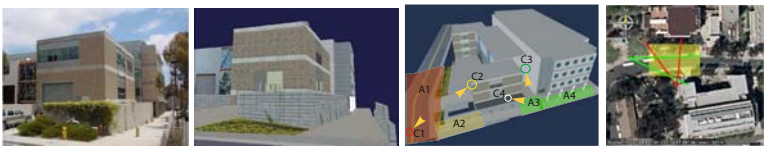
Fig. 3. Schematic diagrams for trajectory analysis in spatio-temporal space. Circles represent interaction potential boundaries at a given space/time. Red curves represent the envelopes of the interaction boundary along tracks.

velocity (v) vs. spatial (y) dimension, planar view of the track and interaction potential in space (x vs. y axis), and the interaction duration between two tracks depicted by the rectangle along a time line, respectively.

The main focus on moving-person interactions in this paper is regarding the macro-level concepts such as *approach, pass by, depart*, etc. This kind of interactions is characterized by short duration of the interaction period.

4 Distributed Sensor Placements and Site Model

Environmental context, especially spatial environment, can be represented by site modeling. Various approaches are possible depending on the available site information. If the 3D structural information is available, we can build a 3D CAD model of the site, which is useful for representing important structures such as buildings and roads. The merit of 3D CAD model is that it provides actual 3D world coordinate systems for the site. But this modeling usually requires multiple cameras and accurate camera calibration. If the site is mainly composed of a flat ground plane, then we can build a planar homography. The advantage of homography-based modeling is that it provides perspective-compensated plan view of the site. It may require multiple cameras with overlapped field of view (FOV). If the site is arbitrarily complex or spatial configuration is ambiguous from camera view, we can still manually assign region of interest (ROI) for specific interest regions. Most of the single camera-based 2D site modeling falls in this category. The advantage of 2D site modeling is that it is flexible and



(a) (b) (c) (d)

Fig. 4. The real testbed of the current system: Actual building (a),its 3D site model (b), camera placements (C1-C4) with viewing directions to areas (A1-A4) (c), and Area-1 (A1) in yellow in the satellite image (d), respectively

simple. Some ambiguity is inevitable due to occlusion, perspective distortion of the view, etc.

We have built an intelligent infrastructure (called 'smart space') with the combination of the above modeling options to generate a heterogeneous site model of an actual building in Fig. 4 (a) which is located in the satellite image in Fig. 4 (d). A 3D CAD model is made and texture-mapped based on architectural data about the building structure and floor plans (Fig. 4 (a), (b).) Four cameras are mounted on specific locations of the building to cover surrounding roads (Fig. 4 (c)). Camera placements are indicated by (C1-C4) with viewing directions and the corresponding view areas (A1-A4). Cameras 1 and 2 view Area-1, Camera-3 views Area-3, and Camera-4 views Area-4, respectively. This paper focuses on Area-1 viewed from Cameras 1 and 2. Area-1 viewed from C1 and C2 is shown in yellow in Fig. 4 (d); straight lines depict the camera fields of view, and the yellow rectangular region corresponds to the planar homography result in Fig. 5 (c). (See the upper right panel in Fig. 6 for another example of the homography mapping.)

5 Tracking of Multiple Objects

Vision-based tracking of multiple objects starts from the processing of foreground segmentation. We use the frame differencing technique with posterior morphological operation for the segmented foreground.

At track-level, the multi-object tracking uses bounding boxes and 2D Gaussian representations of foreground regions. As the objects translate, the Gaussian parameters are updated along the sequence in a frame-by-frame manner. Updating these Gaussian parameters along the sequence amounts to tracking the objects moving in the 2D image frames. The expectation maximization (EM) based updating of the 2D Gaussian representation effectively keeps track of grouping and splitting between objects. However, the usual EM-based update is not reliable under severe occlusions or long time grouping and it can be caught in a local minimum in the parameter space. We overcome this problem by constraining the EM process according to each objects' track history and velocity limitations. At detaile level, the color distribution of the foreground regions are represented by Gaussian mixture model and trained by Expectation-Maximization (EM) learning algorithm. Multiple coherent image patches (called blobs) are formed within

the segmented foreground regions and are represented by the Gaussian mixture model. The details of the tracking algorithm is explained in our previous paper in [7].

6 Planar Homography Mapping

The geometric registration of a camera viewpoint is performed using a homography mapping H from a set of 4 matching points between image coordinate system (i.e., $[x_i, y_i], i \in \{1, 4\}$) and the world coordinate system (i.e., $[x_i^{'}, y_i^{'}], i \in \{1, 4\}$). The perspective parameters correspond to a null space of the matrix A (given in Eqn. 3) and are estimated using SVD of A.

$$AH = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1^{'}x_1 & -x_1^{'}y_1 & -x_1^{'} \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -y_1^{'}x_1 & -y_1^{'}y_1 & -y_1^{'} \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2^{'}x_2 & -x_2^{'}y_2 & -x_2^{'} \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -y_2^{'}x_2 & -y_2^{'}y_2 & -y_2^{'} \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -x_3^{'}x_3 & -x_3^{'}y_3 & -x_3^{'} \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -y_3^{'}x_3 & -y_3^{'}y_3 & -y_3^{'} \\ x_4 & y_4 & 1 & 0 & 0 & 0 & -x_4^{'}x_4 & -x_4^{'}y_4 & -x_4^{'} \\ 0 & 0 & 0 & x_4 & y_4 & 1 & -y_4^{'}x_4 & -y_4^{'}y_4 & -y_4^{'} \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \\ h_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{3}$$

If we denote H_2^1 as the homography from view 2 to 1, we can register multiple cameras by series of concatenated homographies given in Eqn. 4.

$$H_m^n = H_{n+1}^n H_{n+2}^{n+1} \dots H_{m-1}^{m-2} H_m^{m-1} \tag{4}$$

In the current system, we map points in view 1, P_1, and points in view 2, P_2, to a common corresponding point in the virtual view, P_v, by homography matrices H_1^v and H_2^v, respectively. The coordinate system of the virtual view is specified by the 3D CAD model in Fig. 4.

$$P_1^v = H_1^v P_1 \tag{5}$$

$$P_2^v = H_2^v H_2 \tag{6}$$

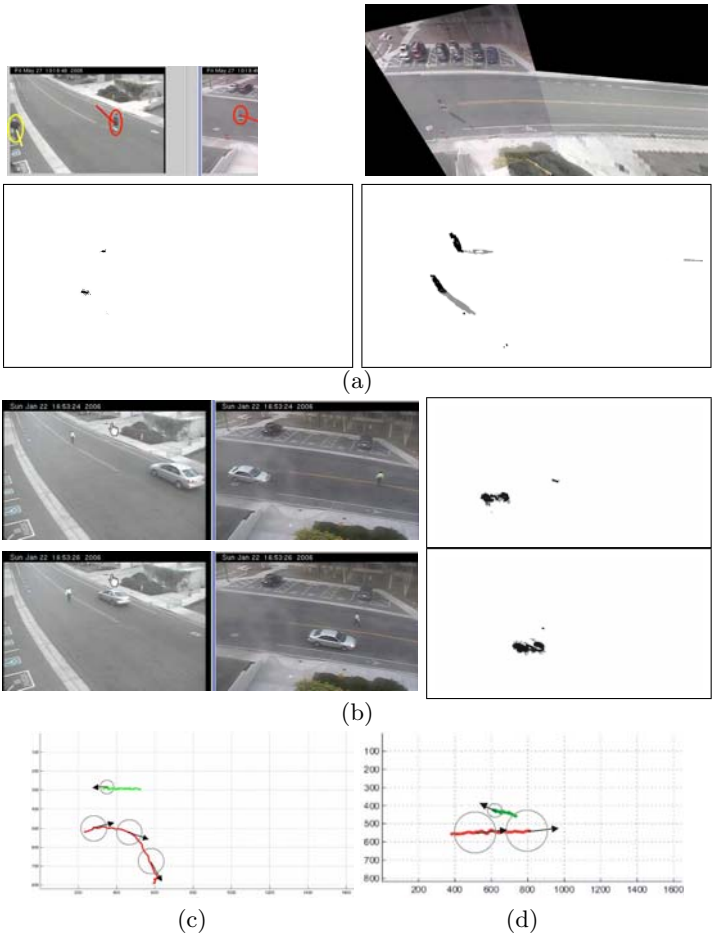
P_1^v and P_2^v are then averaged.

7 Experiments

We have tested our system with video data captured at area A1 in Fig. 4 during different day times for several days. Two cameras C1 and C2 were used to capture the views. Images in Fig. 5 (a)(b) are example views from camera C1 and C2 with detected persons, respectively, and the homography-based registration result is shown in (c). The green and red regions in (c) are user-defined walkway and driveway for site contextualization. The ground truth for the image registration is obtained from satellite imagery in Fig. 4 (d).

(a) (b) (c)

Fig. 5. View registration of Area-1 (A1) in Fig. 4 using the planar homography



(a)

(b)

(c) (d)

Fig. 6. Estimation of interaction patterns of moving objects with different velocities using the homography-mapped footage regions in the projection planes. Walking person in green plot vs. skateboarding person in red plot (a)(c). Walking person in green plot vs. driving car in red plot (b)(d).

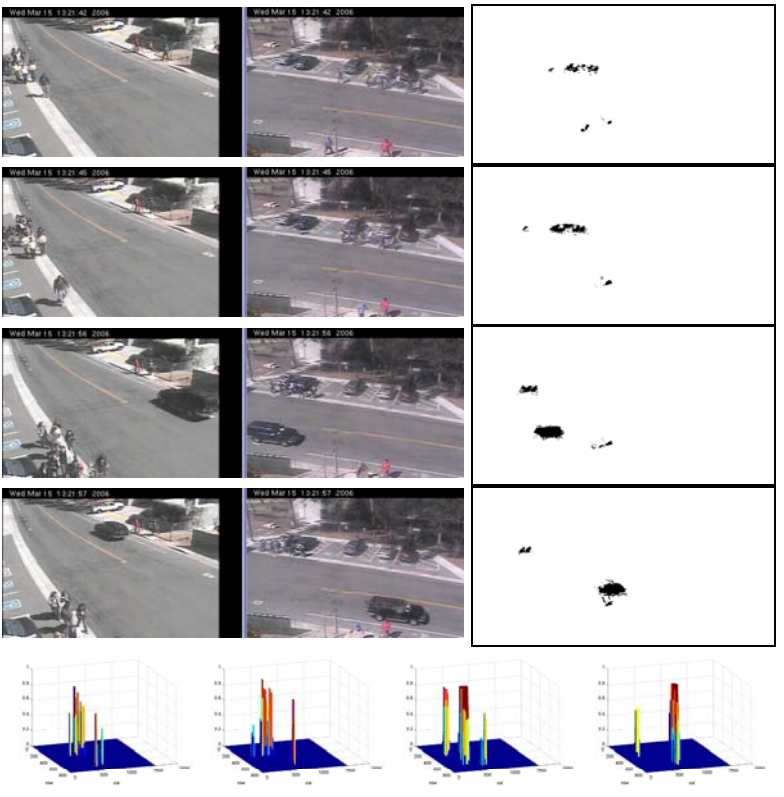


Fig. 7. Dynamic density estimation of crowds and moving vehicles captured at 0, 3, 14, and 15 seconds, respectively. PDFs correspond to the homography maps.

Fig. 6 shows the estimation of interaction patterns between different moving objects captured on a cloudy day. The interaction patterns are analyzed using the tracks of the homography-mapped footage regions in the projection plane corresponding to the region of interest in Fig. 4 (d). In Fig. 6 (a), the images starting clockwise from upper left panel show that multiple views of the site with two detected persons in circle, homography projection plane map of the two views, overlay of the two projections of foreground regions (not shadow!), and the footage areas of each person obtained from the overlay, respectively. Fig. 6 (c) shows two simultaneous tracks of (a): a walking person's track in green and a skateboarding person's track in red in different speeds. Piecewise velocities are represented by arrows, and spatio-temporal interaction potentials are denoted by gray circles at each moment. The absence of overlap between the two tracks' interaction potentials successfully indicates that the monitored scene is in safe situation. In Fig. 6 (b), the images on the upper and lower rows show the raw input views and the detected footage regions in projection plane at different moments.

Fig. 6 (d) shows two simultaneous tracks of (b): a walking person's track in green and a moving vehicle's track in red. The person's proximity to the big interaction potentials (denoted by gray circles) of the vehicle properly indicates the danger of a possible hit.

Our systems' site context information also includes various statistics computed on the fly for crowd density plot, pedestrian flow directions, vehicle traffic histogram, etc. In wide-view open area, counting individuals may not be possible or robust especially when the site is crowded. Therefore, it would be more useful to estimate the detected objects' density, range, and moving velocity in the world coordinate system using the footage areas for each group of objects. Fig. 7 shows our estimation of dynamic density patterns of crowds and moving vehicles observed on a sunny day. Each row in Fig. 7 shows multi-perspective image frames, and the moving objects' detected footage regions mapped on homography plane. The upper four rows show the scene change in terms of spatial distribution of moving objects. The last row shows the probability density functions (PDFs) of crowdiness of the upper four rows. The PDFs were estimated by dividing the homography plane inherently into grid regions and computing the density of the footage pixels in each grid cell. The density patterns and their dynamic changes provide the information about how each region of the monitored site is occupied by people or vehicles for how many frames and how they interact. The information provides enhanced situational awareness.

From the tested experimental site, it is observed that the driveway is sporadically occupied by fast moving high-density large blobs classified to vehicles, whereas the pedestrian walkways are frequently occupied by slow-moving sparse blobs classified to moving crowds. This empirical observation supports our framework for the spatio-temporal analysis of site context in Table 1.

8 Conclusion

In this paper we have presented a multi-perspective vision-based analysis framework to estimate human and vehicle activities for enhanced situational awareness. Planar homography using multiple perspectives provides invariant estimation of footage area of viewed objects for object classification. Moving objects' tracks are robustly estimated with the footage regions in the world coordinate system. Spatio-temporal interrelationship between human and vehicle tracks is capitalized in terms of different combinations of track vs. site context such as walkway and driveway. The concepts of *interaction boundary* and *time to collide* of each moving objects are introduced to build semanticcally meaningful situational awareness. We demonstrated experimental evaluation of our method using pedestrian safety and disaster anticipation applications. Our multi-perspective vision system and the multi-level analysis framework can be applied to broader domains including emergency response, human interactions in structured environments, and crowd movement analysis in wide-view sites.

References

1. J.K. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):295–304, 1999.
2. Antonini G and Bierlaire M. Capturing interactions in pedestrian walking behavior in a discrete choice framework. *Transportation Research Part B*, September 2005.
3. D. Gavrila. The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.
4. I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 22(8):797–808, August 2000.
5. Stephen J. McKenna, Sumer Jabri, Zoran Duric, and Harry Wechsler. Tracking interacting people. In *4th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2000)*, pages 348–353, 2000.
6. N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian Computer Vision System for Modeling Human Interactions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):831–843, August 2000.
7. S. Park and M. M. Trivedi. A track-based human movement analysis and privacy protection system adaptive to environmental contexts. In *IEEE International Conference on Advanced Video and Signal based Surveillance*, Como, Italy, 2005.
8. P. Remagnino, A.I. Shihab, and G.A. Jones. Distributed intelligence for multi-camera visual surveillance. *Pattern Recognition: Special Issue on Agent-based Computer Vision*, 37(4):675–689, 2004.
9. M. M. Trivedi, T. Gandhi, and K. Huang. Distributed interactive video arrays for event capture and enhanced situational awareness. *IEEE Intelligent Systems, Special Issue on Artificial Intelligence for Homeland Security*, September 2005.
10. S.A. Velastin, B.A. Boghossian, B. Lo, J. Sun, and M.A. Vicencio-Silva. Prismatica: Toward ambient intelligence in public transport environments. *IEEE Transactions on Systems, Man, and Cybernetics -Part A*, 35(1):164–182, 2005.