IEEE International Workshop on Applications of Computer Vision, February 2007

Homography-based Analysis of People and Vehicle Activities in Crowded Scenes

Sangho Park
Computer Vision and Robotics Research Lab.
University of California, San Diego
La Jolla, CA 92037
parks@ucsd.edu

Mohan M. Trivedi
Computer Vision and Robotics Research Lab.
University of California, San Diego
La Jolla, CA 92037
mtrivedi@ucsd.edu

Abstract

This paper presents an new framework for homography-based analysis of pedestrian-vehicle activity in crowded scenes. Planar homography constraint is exploited to extract view-invariant object features including footage area and velocity of objects on the ground plane. Spatio-temporal relationships between people- and vehicle- tracks are represented by a semantic event. Context awareness of the situation is achieved by the estimated density distribution of objects and the anticipation of possible directions of near-future tracks using piecewise velocity history. Single-view and multi-view based homography mapping options are compared. Our framework can be used to enhance situational awareness for disaster prevention, human interactions in structured environments, and crowd movement analysis at wide regions.

1. Introduction

There has been a growing interest in making distributed sensor-based systems for the safety and efficiency of human inhabited environments. It is critically important to understand how persons and objects interact with each other and to extract salient semantics for the development of automated situational awareness system [9]. In addition to the issue of automated situational awareness, privacy protection is another important issue in video surveillance. It is often desirable for a surveillance system to satisfy two seemingly contradictory goals: i.e., recognizing human activity and blocking identification. In this paper, we present a methodology for multi-perspective analysis and query of pedestrian-vehicle activity in homography domain that meets the two goals. This is an extension of our previous work [8] on pedestrian safety enhancement in intelligent transportation infrastructure.

Several research issues are related to the current paper in the context of vision-based behavior analysis. A review of distributed surveillance systems can be found in [12]. Compared to indoor monitoring, outdoor surveillance has to deal with a lot of environmental variations such as weather change, time shift from morning to evening, moving backgrounds, etc. Robustness is still an issue in outdoor surveillance. Exemplar surveillance systems have been either based on track analysis [13] or body analysis [5]. 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 [6]. Other researchers have applied more detailed representation of a human body such as a moving region or a blob [13]. Body-level analysis usually focuses on more detailed activity analysis of individual persons in narrower field of view. Distributed sensor networks with pan-tilt-zoom (PTZ) cameras or omni-directional cameras have been studied to overcome the limitation [3, 11].

Most of the research mentioned above focuses mainly on trajectory analysis to recognize what is happening at the moment. Inclusion of human intent into computer vision has not been actively addressed, even though estimation of intent may enhance anticipation of what is about to happen in near future. This paper proposes a new framework for privacy-secure analysis and query of pedestrian-vehicle activity in crowded scenes by combining multiple homography maps.

The rest of the paper is organized as follows: Section 2 summarizes our approach that uses homography constraints to detect and track moving objects. Section 3 shows the representation of event on the homography domain. Experimental results and concluding remarks follow in Sections 4 and 5, respectively.

2. Objects in Homography Domain

Image appearance of the same object varies significantly according to camera perspectives. Reliable classification of object types may benefit from the estimation of the view-invariant size of the object. It is observed that the approxi-

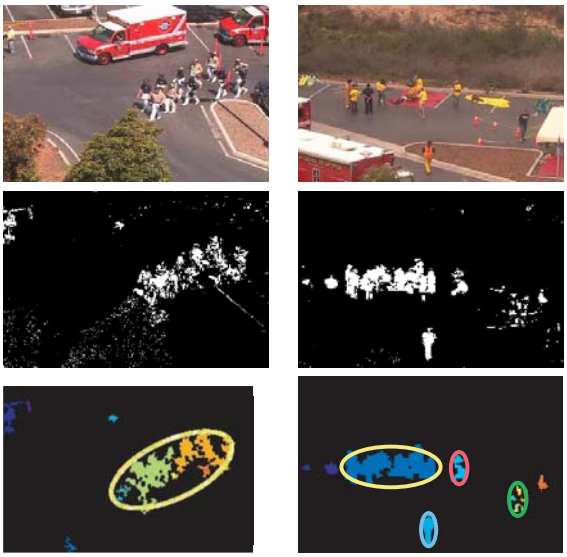

Figure 1. Single view-based estimation of group density.

mate size of the object's footage area is invariant to translation and rotation, unless the object falls or flips over. Planar homography [4] can be used to estimate view-invariant size and location of object on the world coordinate system, (1) if an orthographic camera model is valid from one view, or (2) if weak perspective camera models are valid from synchronized multiple views. The first case holds for a wide field of view (FoV) from a distant camera, while the second case holds for near FoV from synchronized network cameras.

2.1. Single-view Homography

A homography matrix H maps corresponding points between image coordinate systems [4, 10]. If we denote H_2^1 as the homography from view 2 to 1, we can register multiple cameras by the series of concatenated homographies given in Eqn. 1.

$$H_m^n = H_{n+1}^n H_{n+2}^{n+1} \dots H_{m-1}^{m-2} H_m^{m-1} \quad (1)$$

Fig. 1 shows an example of estimating crowd sizes at different places, A and B, captured by a distant camera located at C in Fig. 5. The distance from the camera to the sites is very far and optical zoom was used to capture the video of a disaster drill activity. (More detailed explanation is in the following Experiments Section.) In this situation, the size of an unoccluded object does not vary within the field of view, and orthographic camera model is valid.

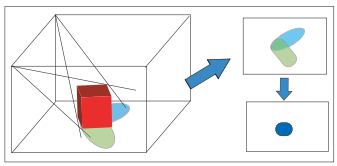

Figure 2. Schematic diagrams for footage area estimation using multiple planar homography.

2.2. Region-based Binding of Multi-view Homography

The single view-based homography suffers from occlusion between objects. If multiple views are available, then conjunction of multiple homography maps provides very useful image invariant: footage area. Multiple views of the same object are projected on to a common planar homography map, and the intersection of the projected images can be used as the object's footage region on the ground.

Schematic diagram in Fig. 2 shows the mapping process from view-1 or view-2 to the virtual top-down view using planar homography.

The process of estimating the footage area is depicted in Fig. 2. We use the 4-point algorithm [4] to compute the homography matrix H. The 4 points are selected from shared image corners or by having a person to walk around the shared region.

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, as follows:

$$P_1^v = H_1^v P_1, P_2^v = H_2^v P_2$$

Planar homography constraint 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. 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. By intersecting multiple projection maps of the same object, we can estimate the object's common footage region that observes the assumption.

Foreground moving objects on the homography plane are detected and segmented by frame differencing over multiple frames with a moving window. The motivation of using multi-frame differencing is that the mis-detection rate with the planar homography constraint is quite high. Therefore we want to reduce mis-detection and raising false alarm first. Then the raised false-alarm rate is effectively reduced by combining multiple homography constraints from multiple views.

Figure 3. Example of homography mapping. Two perspective views are registered in a common virtual top-down view.

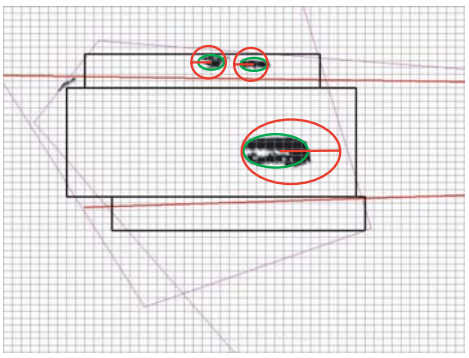

Figure 4. Grid-based representation of moving persons (in small circles) and a vehicle (in large circle) on homography domain. Position, size, and velocity of each object are estimated by the Kalman filter.

Tracking of the detected object is performed by data association between consecutive frames in the homography domain. The multi-object tracking uses 2D Gaussian ellipse representations of foreground regions. Tracking is performed on the homography domain by using a variant of the modified PDAF tracking algorithm [2] combined with a Kalman filter.

Fig. 3 shows an example of generating the virtual top-down view from two perspective images, which can be registered onto an available satellite map. The overlayed two horizontal thick (red) lines in Fig. 3 distinguish the borders of the driveway in the middle. We call the driveway as 'Driveway-Middle' and the two walkways above and below the driveway as 'Walkway-North' and 'Walkway-South', respectively, in the following sections.

Fig. 4 shows an example of object tracking on the homography plane. Each detected object is represented on the virtual grid of 10×10 pixels by a tightly surrounding ellipse, and the Kalman-filter based tracking parameters provide the velocity estimation, which is represented by a bar. The three rectangles depict the 'Walkway-North', 'Driveway-Middle' and 'Walkway-South' from top to bottom, respectively.

Moving object's true velocity (i.e., speed and direction) in 3D world coordinate system is estimated at the projection (i.e., homography) plane. The velocity of a moving object determines the object's reaching boundary in a given time. 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.

3. Event Representation

The interaction between persons and vehicles can be recognized if we identify the objects' categories and their relative motions. The object category is efficiently recognized by estimating the view-invariant footage area of the object in the homography domain. The motion characteristics and its semantic connotation depend on the object type and situational context. We represent person-vehicle interaction as an event in terms of an agent's (i.e., person or vehicle's) meaningful motion toward an optional target. Following a linguistic theory of *verb argument structure*, we have proposed an activity representation framework using the *operation triplet* defined in terms of <agent-motion-target> [7].

In the current paper, the *agent* term includes a single pedestrian, crowd of people, small vehicle, and large vehicle. The *motion* term includes stay, move-left, move-right, move-up, move-down on the homography plane at low / high speeds. The homography mapping provides a very reliable estimation of the objects' true size and speed. A modified version of the HMM-based classifiers [7] is used to recognize the motion descriptions of the tracked objects to form the 'motion' term of the operation triplets associated with the objects. Each object is associated with its own activity status represented by the operation triplet The *target* term includes the same vocabulary as the agent term, and the validity of a target is determined by the proximity defined by any nearby agent.

Event representation may involve *interaction* between multiple sub-events. We represent the interaction of two events with respect to their spatial, temporal, and logical relations. We adopt a predicate calculus to represent these relations as follows; We adopt Allens interval temporal logic [1] to represent temporal relations between two

events: before, meet, overlap, start, during, and finish. Each predicate takes two time intervals as event entities, and decides whether they are true or false. The possible relations between two time intervals $a = (a_{start}, a_{end})$ and $b = (b_{start}, b_{end})$ are define as follow:

before (a,b) $\equiv a_{end} < b_{start}$
meets (a,b) $\equiv a_{end} = b_{start}$
overlaps (a,b) $\equiv a_{start} < b_{start} < a_{end}$
starts (a,b) $\equiv a_{start} = b_{start}$ and $a_{end} < b_{end}$
during (a,b) $\equiv a_{start} > b_{start}$ and $a_{end} < b_{end}$
finish (a,b) $\equiv a_{end} = b_{end}$ and $a_{start} > b_{start}$

Spatial relation of two events is represented by the spatial proximity of the events. The threshold of proximity between two objects is adaptively determined by the velocity of each object. Logical relations represent useful domain knowledge and site context. For example, a large vehicle is not likely to go over a walkway, but pedestrians and vehicles may go over a driveway frequently. We use *and, or, not* predicate to represent a constraint of an event [7].

4. Experiments

We tested our system to apply it to a real situational-awareness task; a multiple agencies-coordinated disaster drill event that involved police department, fire department, SWAT team, medical support team of the City of San Diego.

Fig. 5 depicts the overall setup of the event site that occurred at a university campus' parking lot. Site modeling is achieved by incorporating the a-priori knowledge of the drill plan, which includes different agecies' locations and their role models in the drill. The results of the single view-based homography mapping from Fig. 1 are shown in Fig. 6.

We have also tested our system with video data captured from two synchronized network cameras at 10AM, 2:30PM, and 5PM on different days. Total image frames are 25,665 frames (30 minutes long.) Fig. 7 shows the screen capture of our system.

We conducted experiments on dynamic density estimation in natural settings where unobtrusive busy traffic flows of multiple pedestrians and vehicles were involved. Fig. 8 summarizes the dynamic density (i.e., the evolution of grid density patterns) captured at 2:50pm. The plot shows the grid-cell numbers in each ROI defined in Fig. 4: 'Walkway-North', 'Walkway-South', and 'Driveway-Middle'.

Table 1 shows the confusion matrix of the object classification into 3 classes: *large vehicle* (LV), *small vehicle* (SV), and *pedestrian(s)* (Ps). The dynamic density plot values in Fig. 8 are used as features for classification. A single sample in the table represents the whole span of a peak from any region of the ROIs. *Recall* is defined as the fraction of the total number of objects in a particular class that are

Figure 5. Aerial map of a disaster drill site showing ad hoc configuration of the camera.

	LV	SV	Ps	recall
LV	39/41	2/41	0	0.95
SV	3/55	52/55	0	0.95
Ps	0	4/65	61/65	0.94
precision	0.93	0.90	1	

Table 1. Confusion matrix of object classification.

classified correctly by the system for that class. *Precision* is defined as the fraction of objects recognized for a particular class that are actually correct.

Table 2 shows the speed estimation of different object types captured by the current system. Note that the skateboarding involves fast motion of a person, which may result in a dangerous situation. The connotation of specific motion depends on situational contexts, and it motivates us to categorizes interaction patterns as safe or unsafe interaction.

Fig. 9 shows the estimation of interaction patterns between a pedestrian and a small vehicle captured on a cloudy

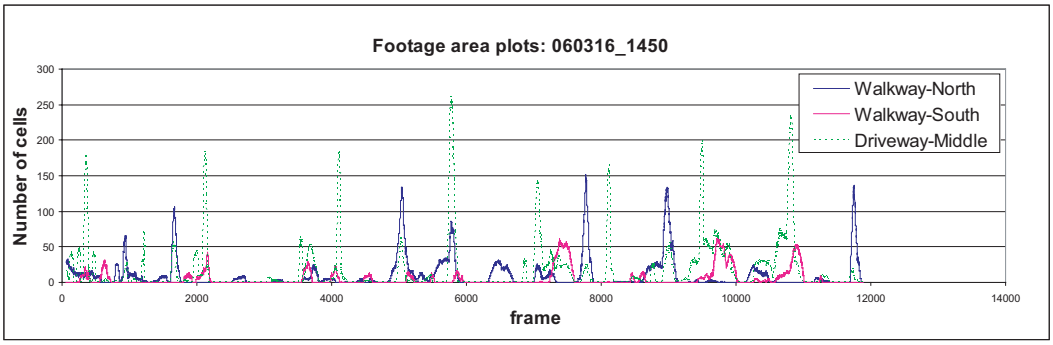

Figure 8. Dynamic density estimation of crowds and moving vehicles represented in terms of number of grid cells vs. frame. See the text for details.

Figure 6. Visualization of the crowd motions in Fig. 1 on the ground plane using the single-view homography. Compare it with Fig. 4.

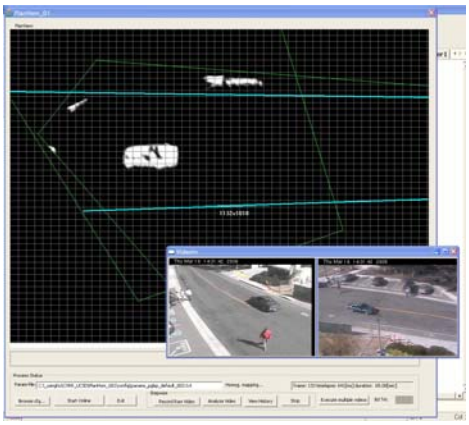

Figure 7. Screen capture of our proposed system that detects and tracks moving foreground objects on the homography domain. The inlet shows the multi-perspective input frames.

day. The first row in Fig. 9 shows the two views of a walking person and a car. The second row shows two simultaneous tracks: a walking person's track in green and a moving vehicle's track in red. The lengths of the arrows depict the different speeds. The reachable boundary in a given period (called *interaction potential*) is depicted by the overlayed circles. The interaction pattern is analyzed using the tracks on the projection plane. HMM-based event recognizers were built for training and testing of events in the give scene. The person's proximity to the big interaction potentials (denoted by gray circles) of the vehicle properly indicates the danger of a possible hit.

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.

Speed [m/sec]		
Bus	9.4	$+/-$ 2.31
Sedan	10.8	$+/-$ 3.43
Pedestrian	2.0	$+/-$ 0.36
Skateboarding	7.5	

Table 2. Speed estimation of different object types: bus, sedan, pedestrian, and skateboarding person on driveway.

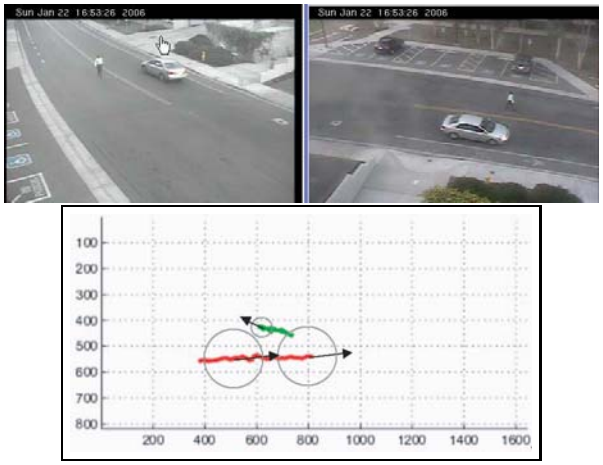

Figure 9. Estimation of interaction patterns of moving objects.

5. Conclusion

In this paper we have presented a framework for homography-based analysis of people and vehicle activities in crowded scenes for enhanced situational awareness. Planar homography constraint using a single or multiple views in a certain imaging condition provides invariant estimation of footage area and velocity of objects. Intent of a pedestrian or a driver is estimated by the direction and speed of motion on the view-independent homography plane. 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, which builds semantically meaningful situational awareness. A grid density-based indexing paradigm and an event grammar are combined to provide an efficient query method. We demonstrated experimental evaluation of our method using pedestrian safety and disaster anticipation applications. The proposed framework can be applied to broader domains including human interactions in structured environments, people behavior control, crowd movement analysis, and traffic control.

Acknowledgments

This research was supported in part by US DoD Technical Support Working Group (TSWG) and by NSF RESCUE ITR-Project. We thank our colleagues from the UCSD Computer Vision and Robotics Research Laboratory for their valuable support.

References

[1] J. F. Allen and G. Ferguson. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4(5):531–579, 1994.

[2] Y. Bar-Shalom and W. Blair. *Multitarget-multisensor tracking: applications and advances*, volume 3, pages 199–231. Norwood, MA, 2000.

[3] R. Brooks, P. Ramanathan, and A. Sayeed. Distributed target classication and tracking in sensor networks. *Proc. of the IEEE*, 91(8):1163–1171, 2003.

[4] A. Criminisi, I. Reid, and A. Zisserman. A plane measuring device. *Image and Vision Computing*, 17(8):625–634, 1999.

[5] 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.

[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 J.K. Aggarwal. Semantic-level understanding of human actions and interactions using event hierarchy. In *IEEE Workshop on Articulated and Nonrigid Motion*, 2004.

[8] S. Park and M. M. Trivedi. Multi-perspective video analysis of persons and vehicles for enhanced situational awareness. In *IEEE Int'l Conference on Intelligence and Security Informatics (ISI2006), LNCS 3975*, pages 440–451, San Diego, USA, 2006.

[9] S. Park and M. M. Trivedi. Multi-person interaction and activity analysis: A synergistic track- and body- level analysis framework. *Machine Vision and Applications Journal: Special Issue on Novel Concepts and Challenges for the Generation of Video Surveillance Systems*, to appear.

[10] C. Stauffer, K. Tieu, and L. Lee. Robust automated planar normalization of tracking data. In *Proc. IEEE Intl Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Nice, France, 2003.

[11] M. M. Trivedi, K. S. Huang, T. Gandhi, B. Hall, and K. Harlow. Distributed omni-video arrays and digital tele-viewer for customized viewing, event detection and notification. In *IEEE Int Conf on Information Technology: Coding and Computing (ITCC'04)*, pages 669–674, 2004.

[12] M. Valera and S. Velastin. Intelligent distributed surveillance systems: a review. *IEEE Proceedings Vision, Image and Signal Processing*, 152(2):192–204, 2005.

[13] S. Velastin, B. Boghossian, B. Lo, J. Sun, and M. Vicencio-Silva. Prismatica: Toward ambient intelligence in public transport environments. *IEEE Transactions on Systems, Man, and Cybernetics -Part A*, 35(1):164–182, 2005.