Detecting Objects, Shadows and Ghosts in Video Streams by Exploiting Color and Motion Information

R. Cucchiara[1], C. Grana[1], M. Piccardi[2], A. Prati[1]
[1] *D.S.I. - University of Modena and Reggio Emilia, via Vignolese 905 - 41100 Modena, Italy*
Email: rita.cucchiara@unimo.it,{grana, prati}@dsi.unimo.it
[2] *Dip. Ingegneria - University of Ferrara,*
via Saragat 1 - 44100 Ferrara, Italy mpiccardi@ing.unife.it

Abstract

Many approaches to moving object detection for traffic monitoring and video surveillance proposed in the literature are based on background suppression methods. How to correctly and efficiently update the background model and how to deal with shadows are two of the more distinguishing and challenging features of such approaches. This work presents a general-purpose method for segmentation of moving visual objects (MVOs) based on an object-level classification in MVOs, ghosts and shadows. Background suppression needs a background model to be estimated and updated: we use motion and shadow information to selectively exclude from the background model MVOs and their shadows, while retaining ghosts. The color information (in the HSV color space) is exploited to shadow suppression and, consequently, to enhance both MVOs segmentation and background update.

1. Introduction

Detection of moving objects in video streams is the first relevant step of information extraction in many computer vision applications, including traffic monitoring, automated remote video surveillance, and people tracking. A common goal to these applications is robust tracking of objects in the scene, requiring to be based on a reliable and effective moving object detection.

This work proposes a novel approach for detection of moving objects in video streams (moving visual objects, or MVOs for short hereafter) in unconstrained indoor and outdoor video scenes. The approach is meant to be general purpose and is based on three assumptions:

- the background and the camera are assumed to be stationary;
- background changes are due to two factors: a) light condition variations (e.g. clouds covering the sun in outdoor scenes or lights turned on or off in indoor scenes); b) objects that modify their status from stopped to moving or vice versa;
- object segmentation is assumed model-independent: thus, we deal with different object classes (vehicles, pedestrians, bicycles, and so on) whatever their motion, trajectory and speed. Thus, our approach cannot be based on frame difference, where the frame rate must be carefully tuned in dependency on the object speed. At the same time, motion models cannot be exploited.

Starting from these assumptions, we focussed our attention on moving object detection based on background suppression. By this approach, an estimate of the background (also called a background model) is computed and evolved frame by frame: moving objects in the scene are detected by the difference between the current frame and the background model.

A typical problem arises from the changing nature of the background, as stated by assumption 2. First, there must be a trade-off between high responsiveness to changes and reliable background model computation. Second, the model must deal with erroneous "ghost" detection: when objects belonging to the background start to be in motion, they will be displaced with respect to their original position and the background subtraction will detect relevant differences in two areas: the area where the object is currently located, and the area where the objects was originally. This second area is commonly referred to as a ghost (see for instance [2][3]), since it does not correspond to any real moving object.

Another problem arising in object segmentation is related to shadows. Indeed, we would like the moving object detection not to classify shadows as belonging to foreground objects. Unfortunately, points of objects and associated shadows share two important visual features: motion and detectability. Thus, whatever the background update, often the moving points of both objects and shadows are detected at the same time and grouped together. As a consequence, the appearance and geometrical properties of the object are distorted. This problem affects many subsequent tasks, such as object classification and the assessment of moving object position (normally accounted as the shape centroid), as, for instance, in traffic control systems that must evaluate trajectories of vehicles and people on a road. Moreover, the probability of object undersegmentation (i.e. object merging) increases due to connectivity via shadows between different objects [4]. In order to eliminate these drawbacks, we have defined an approach for shadow detection and suppression based on a color analysis in the HSV space.

This work proposes a novel approach that fully exploits both motion and color information to detect moving objects, shadows and ghosts and exploit their knowledge for good detection and good background update.

The remainder of this paper is organized as follows. Section 2 briefly reviews some related work on the topic of moving object detection in video streams. In Section 3, an overview of the approach and its ability to recognize moving objects, shadows and ghosts are provided. The novelties in the background update and shadow detection methods are detailed in Sections 4 and 5, respectively. Finally, a system prototype and its possible applications are outlined and discussed in Section 6.

2. Related work

A large literature exists concerning moving object detection in video streams, typically conceived as the first step of applications such as traffic control and video-surveillance. Many different approaches have been proposed including frame difference [5], double frame difference [6][7], and background suppression. The background suppression approach requires a computationally expensive background update, but is more general, and thus we focus on it in the following. Many works propose to perform the background update by using statistics functions on a sequence of the most recent sampled frames: for instance in [8] a mean function is used, in [9] the mode, in [3] multiple Gaussians. However, in order to correctly estimate the background model, a rather large frame sequence must be used. Other works propose to combine the statistics on the frame sequence with previous values of the background model (we will call these proposals adaptive methods for the sake of briefness): in [10] the use of a Kalman filter is suggested, while in several other papers a weighted average of the previously computed background and the current frame(s) (such as in [11]) is used, requiring a limited computational load.

Since MVOs are not part of the background, their inclusion in the background update function leads to errors. However, most of the aforementioned methods considers indifferently pixels belonging to MVOs and other pixels. In order to solve this deficiency, some methods propose to exclude from the background update pixels detected as moving points; we call these methods selective background update [3][1]. However, the use of selectivity carries a further problem, associated with ghosts: if ghosts are excluded from the background update, the background will never be correctly estimated, and ghosts will be permanently detected. This problem is referred to as deadlock[12]. To solve this problem, in [3] a verification step is introduced to check if pixels are really points in motion. In this work, we propose to perform this verification on the whole object containing the pixel, since the information on the whole object seems more reliable from the experiments performed.

Methods for detecting shadows have been proposed in a number of recent papers. In [13], an approach is proposed for extracting shadows from still images, which could be applied also to video streams, by analyzing each single frame separately. Other works, instead, propose methods for shadow detection based on the difference between the current frame and a reference frame. In [14], a method is proposed for traffic scenes: first, the background is suppressed from the current frame, then shadows are separated by looking either for vertical or horizontal edges, depending on the road, date, and day time. Therefore, the approach may not be easy to apply where this information is not available, like in indoor scenes. In [15], the authors propose to compute the ratio of the luminance between the current frame and the previous frame; a point is marked as shadow if the local variance of this ratio is small (this criterion is then followed by further validation). In [3], too, the ratio of the luminance the current frame and the background model is exploited. An improvement of this method is proposed in [1] based on the observation that shadows are semitransparent, retaining features of the covered surface such as patterns, color, textures; therefore, the authors propose an analysis of the chromaticity in the {R,G,B} color space. In this work a novel shadow detection approach is presented, similar to [1], but based on an analysis in the {H,S,V} color space, which seems more intuitive.

3. Detecting objects, shadows and ghosts

The goal of the process we propose is twofold. The first aim is to achieve good detection, meaning that we want to detect real moving objects correctly, separating them from their cast shadow. The separation is not carried out in terms of model-based object recognition, nor it is eased by assumptions on the light source position and light direction. Instead, shadow detection is based only on a syntactic discrimination between the "appearance" of shadows and objects in terms of both luminance and color. The second aim is to detect moving objects only, without confusing them with apparently moving areas, static objects, or noise, with the maximum responsiveness possible. Since we segment objects by means of background subtraction, this means a good background model and update, i.e., the definition and the dynamic modification of the background should be accurate and quickly updated.

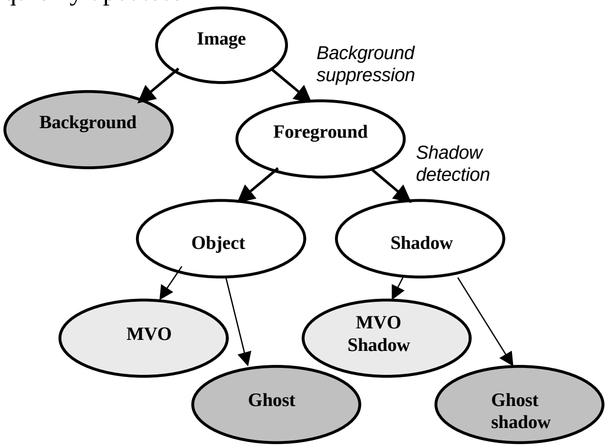


Figure 1. Object classification

To these aims, according with Fig. 1 we give the definitions of:

Moving visual object (MVO): is a target object that can be obtained by an ideal segmentation, i.e. the set of the connected points belonging to an object currently characterized by non null motion and a visual appearance different from the background. Conversely,

Background: is the set of scene points currently not in motion.

Among background points, we further distinguish:

Ghost: is a set of connected points, detected as in motion but not corresponding to any real moving object.

Shadow: is a set of connected background points modified by a shadow cast over them by a moving object (note that we do not consider static cast shadow, i.e. points shadowed by fixed objects, that are included in the generic set of background points).

Eventually, shadow can further be classified as MVO shadow, that is a shadow connected with an MVO, thus sharing its same motion, and Ghost shadow, i.e. a shadow not connected with any real MVO.

For instance, in the indoor scene of Fig. 2 a person passes away, after having opened a cabinet door. By simple background subtraction, all points in the right image are detected. The rightmost blob corresponds to the correct MVO (in grey) and its shadow (in black); the blob in the middle (in grey) is a ghost, with its shadow (in black) detected because the door of the cabinet is open in the current frame but still closed in the background model

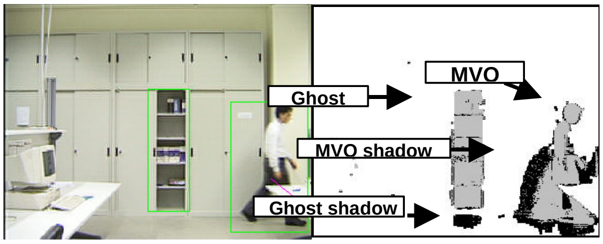


Figure 2. Example of image point classification in an indoor scene

Why have we outlined this classification? Because we state that both good detection and good background update can be achieved only if knowledge of all five categories is exploited explicitly.

For good detection, the need for shadow detection is obvious. The lack of separation between MVOs and MVO shadows causes two possible errors: the first one, unavoidable, is that the object shape is distorted and all geometric proprieties associated with MVOs are affected by errors; the second is that, due to shapes, more separate objects can be merged in a single blob, causing errors in further identification and tracking steps. Moreover, a good detection must be able to discriminate MVOs from “false positives”, such as ghosts. Therefore, discriminating shadows, ghosts, and MVOs is needed for good detection. It is useful also for robust background update, as will be outlined in the next section.

4. Background modeling and update

According with most of the recent literature, we adopt a background model computed at every new frame as a statistical combination of a sequence of previous frames and the previously computed background. We assume the background points to be those image points more frequently (in a statistical sense) observed as still. The statistical function chosen is the median: in [1] we compared median with the mean and mode functions and we experimentally proved that median performs well even with limited length of the sequence of previous frames. This method has some weak points: if the observation time window is limited, points of moving objects could be included in the background; on the contrary, due to the non null time window, the update process is slow and many false positives arise. In order to limit this effect, some authors propose to use selectivity [3][1], by excluding points detected as moving from the background update. However, wrong selection may lead to deadlock problems. Therefore, the approach we propose is the following:

$$B_{t+\Delta t}(x,y) = \begin{cases} B_t(x,y) \quad if \quad (x,y) \in \{MVO\} \cup \{MVO \ \ shadow\} \\ f\,(I_t(x,y), I_{t-\Delta t}(x,y), ..., I_{t-n\Delta t}(x,y), w_b B_t(x,y)) \\ \quad if \quad (x,y) \in \{BKG\} \cup \{ghost\} \cup \{ghost \ \ shadow\} \end{cases} \quad (1)$$

In order to exploit selectivity, we do not include in the background update process those points belonging to both MVOs and their MVO shadows. Instead, points belonging to ghosts or ghost shadows are correctly included in the background update by means of the function f, which computes, for each (x,y) point, the median value between values in previous frames and in the current background.

The complete process is described in Fig. 3. After an initial camera motion correction, the system selects foreground points by means of background suppression. These points are candidate to belong to MVOs since they are different from the current background. In order to improve detection, background suppression is computed by taking into account point chromaticity and not in gray levels only [17]. We compute the difference with the background DBt[1] as the distance in the RGB color space as:

$$DB_t(\mathrm{x,y}) \ = \ \mathrm{Distance}\ (I_t\ (\mathrm{x,y}),\ B_t\ (\mathrm{x,y}))$$
$$= \ \max\ (\ \left|I_t(\mathrm{x,y}).\mathrm{c} - B_t(\mathrm{x,y}).\mathrm{c}\right|\ )\ \ \mathrm{c} = \mathrm{R,G,B} \qquad (2)$$

On the difference image DBt, the selection of the initial set of foreground points is then carried out by thresholding with an adequately low threshold TL. Among the selected points, some are discarded as noise and included in the background, by applying morphological opening operators. Then, the shadow detection process is applied and other points are labeled as shadow points and separated from the set of foreground points. A region-based labeling is performed to compute the connected blobs of candidate moving points (by means of 8-connectivity). Finally, blob analysis validates detected blobs as either moving visual objects or ghosts.

MVOs are validated with rules on area, saliency and motion. First the area must be large enough (greater than a TA threshold depending on the scene and on the signal-to-noise ratio of the acquisition system); then, the blob must be a "salient" foreground. In practice, we use a double threshold for foreground points: the previously defined TL select many candidates, while a higher threshold, TH, confirms only “strong” foreground and points connected, discriminating fortuitous noise aggregation. Finally, the third rule for MVO validation is the average blob motion. To measure motion, for each pixel belonging to an object we compute the spatio-temporal differential equations for optical flow approximation. The average optical flow computed over all the pixels of an MVO is the measure we use to discriminate between MVOs and ghosts: in fact, ghosts have a near-to-zero optical flow, since their motion is only apparent, resulting from erroneous background values. Optical flow is very time consuming; however, we compute it only when and where necessary, i.e. only on the blobs resulting from background suppression (a small percentage of image points, as reported in Table 1). This allows us to achieve real time performance with standard PC hardware and, at the same time, to obtain a significant measure of motion in the scene (useful for example to compute the average object speeds).

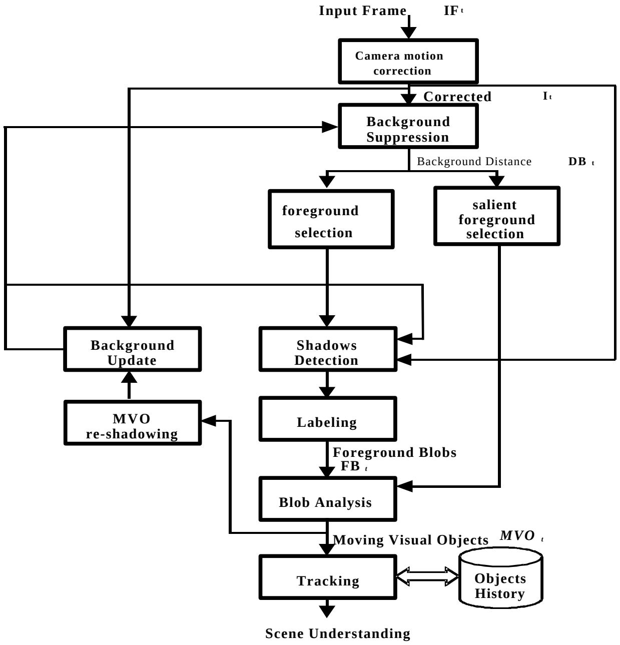


Figure 3. The control flow path

Validated MVOs are excluded from background update. The same validation should be carried out for shadow points, too, in order to select those corresponding to moving shadows, that must similarly discarded from background update. However, computing the optical flow is not reliable on uniform areas such as shadows, which typically do not exhibit high average optical flow. In fact, the spatial difference in the equation is nearly nullified because shadow smoothes and uniforms the luminance values of underlying background. Therefore, in order to discriminate MVO shadows from ghost shadows we use information about connectivity between objects and shadows: shadows 8-connected to MVOs are classified as moving and finally associated with foreground points;

[1] Bold notation means vector variable in the RGB color space; each value referred on a single color band is indicated by a dotted suffix.

remaining shadows are reintroduced in the background update in the step named re-shadowing in Fig. 3.

The approach we proposed is independent from any a-priori knowledge of the scene, in the sense that it works at pixel level without exploiting any model-based assumption on the scene, aiming to be general purpose. By this approach, a stopped MVO is included in the background after the update time. However, if the application requires it, it is straightforward to prevent inclusion of stopped objects in the background, even if they stay still for a high number of frames: objects identified as MVO, can be considered MVO in successive frames even if their intrinsic motion is null with an arbitrary timeout.

In order to give evidence of the performance that can be achieved with this approach, we have defined the two performance metrics DR (Detection Rate) = TP/(TP+FN) and FAR (False Alarm Rate) = FP/(TP+FP), where TP is the number of the correctly detected MVO pixels, FN the missed MVO pixels, and FP the background pixels incorrectly detected as MVO pixels. In the ideal case, DR should be 1 and FAR 0. We have implemented the approach described and tested on several different sequences. In particular, we have devised a ground-truth benchmark where a car is moving along a complex trajectory, starting and stopping several times; on this benchmark, we have measured a very high detection rate DR = 0.988, with FAR = 0.019 only.

5. Shadow detection

Shadow detection and suppression from the set of foreground points aim to prevent moving cast shadows being misclassified as moving objects, thus improving object detection and limiting the risk of undersegmentation. Detecting shadows is not trivial in general. In fact, as deeply detailed in [4], the two classes of points belonging to objects and shadows may have similar visual appearance in many cases. This is quite true, especially if working with grey level only. In [4], we proved that the discrimination between shadows and objects can be improved by adding color information. Also in previous works in the literature (see for instance [2]) the chrominance information was exploited. However, in [4] we presented a novel approach based on the exploitation of the HSV space to better distinguishing shadows from objects, and reported results on detection improvement. Our algorithm is based on the following equation:

$$SP_t(x,y) = \begin{cases} 1 \quad if \quad a \le \dfrac{I_t(x,y).V}{B_t(x,y).V} \le b \quad \wedge \quad \left| I_t(x,y).H - B_t(x,y).H \right| \le t_H \\ \qquad \wedge \quad I_t(x,y).S - B_t(x,y).S \le t_S \\ 0 \quad otherwise \end{cases} \qquad (3)$$

where SPt(x,y) is set to 1 if point It(x,y) is classified as shadow, 0 otherwise.

Eq. 3 states that a point (x,y) is classified as shadow if three properties hold: i) the ratio of the V component (i.e., the lightness) of It(x,y) and Bt(x,y) respects both a lower and a upper bound; ii, iii) the differences of the H and S components (i.e., the chromaticity) are limited.The rationale of the equation comes from the observation that when an area is covered by a shadow, this often results in a significant change in lightness without a great modification of the color information. Thus, we upper-bound the hue and saturation differences with a threshold each, which values are deducted by many experiments, and we impose the lightness ratio to be a value bound by two thresholds α and β (with $0 < \alpha < \beta < 1$): the first one takes into account the “power” of the shadow (the lower the α value, the more the shadows are assumed to darken the covered objects), while the second is used to increase the robustness to noise (the lightness of the current frame cannot be too similar to that of the background).

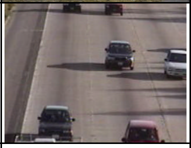
In order to assess performance of shadow detection, we have segmented all the ground-truth MVOs from a video sequence with strong shadows, and compared against those extracted without and with shadow suppression. On the sequence, we have measured two parameters which could be exploited in the tracking phase, namely the MVO area (in pixels) and centroid position. The MVO area results 52.9.% greater than ground truth on average without shadow suppression, while 10,6% only by using shadow suppression. The centroid position is about 4.6 pixels distant from that of ground-truth MVO on average without shadow suppression, and 1.9 only with shadow suppression, thus proving the efficacy of the proposed approach.

6. Application

The method for detecting moving objects, shadows and ghosts in video streams presented in this work is part of a visual tracking system, that we called SaKbOT (Statistic and Knowledge-based Object Tracker) system. The system is composed of two main modules, one for object detection described in the previous sections and the other for object tracking. The information on the detected MVOs (without shadow) extracted by the object detection module is passed to a higher level module that implements tracking using object-level history. The object-level matching between objects in the scene and objects in the past history is done by using a set of rules working on a symbolic representation of objects as feature vectors. Further details on the tracking module are reported in [1]. This object-level abstraction allows to reduce the computational load and to compensate for errors of the detection module by assessing object consistency during time.

The SaKbOT system has been tested in a wide range of different environments and applications: from video-surveillance of the campus of University of Modena (Italy), to traffic monitoring at intersections, aimed to optimize traffic lights timings in a project of sustainable mobility of the city of Bologna (Italy), to highway incident detection at University of California, San Diego[2].

Table 1 shows a variety of applications where our MVO detection method has been experimented, including traffic monitoring of urban areas and highways, surveillance of parking zones, and indoor people detection and tracking. These applications differ significantly in terms of light conditions, and, as reported in Table 1, density of objects in the scene, object size, number of frames of object presence, thus proving that the method can be rightly considered as general-purpose.

Application	Sample frame	Obj	Size	Pres
Shopping center **(2300 frames 352x288)**		140	1484	92
US highway **(440 frames 320x240)**		70	3241	19
Parking area **(500 frames 345x135)**		1	2456	490
Laboratory **(980 frames 320x240)**		2	7228	48

Obj: total number of objects detected
Size: average MVO size in pixels
Pres: average number of frames of MVO presence

Table 1. Examples of application.

Typical parameters used in SaKbOT are n = 7 for the length of the sequence of previous frames in the median computation, sub-sampled one every ten; the weight wb in Eq. 1 is posed equal to 2. All the thresholds are empirically tuned, but they proved stable under environment changes. The system runs on a standard PC with images up to 352x288 pixels and achieve average performance close to 10 frames per second. For most applications this can be considered real-time.

References

[1] Cucchiara, R. Grana, C., Piccardi, M., and Prati A., “Statistic and knowledge-based moving object detection in traffic scenes”, Proc. of ITSC2000, 2000, pp. 27-32.

[2] McKenna, S.J., Jabri, S., Duric, Z., and Rosenfeld, A. and Wechsler, H., “Tracking groups of people”, CVGIP, 2000, vol. 80, pp. 42-56

[3] Elgammal, A., Harwood, D., and Davis, L.S., “Non-parametric Model for Background Subtraction", Proc. of ICCV '99 FRAME-RATE Workshop, 1999.

[4] Cucchiara, R., Grana, C., Piccardi, M., and Prati, A., “Improving Shadow Suppression in Moving Object Detection with HSV Color Information”, to appear in Proc. of ITSC’01, 2001.

[5] A. Lipton, H. Fujiyoshi, and R. Patil, “Moving target classification and tracking from real-time video,” Proc. of WACV’98, 1998, pp. 8-14.

[6] Cucchiara, R., Piccardi, M., and Mello, P., “Image analysis and rule-based reasoning for a traffic monitoring system”, IEEE Trans. on Intelligent Transportation Systems, vol. 1, no. 2, June 2000, pp. 119-130

[7] Yoshinari, K., and Michihito, M., “A human motion estimation method using 3-successive video frames”, Proc. of Intl. Conf. on Virtual Systems and Multimedia, 1996, pp. 135-140.

[8] Dagless, E.L., Ali, A.T., and Bulas Cruz, J., “Visual road traffic monitoring and data collection”, Proc. of IEEE-IEE VNIS'93, 1993, pp. 146-149.

[9] Shio, A. and Sklansky, J., “Segmentation of People in Motion”, Proc. of IEEE Workshop on Visual Motion, 1991, pp. 325-332.

[10] Koller, D., Weber, J., Huang, T., Malik, J., Ogasawara, G., Rao, B., and Russell, S., “Towards Robust Automatic Traffic Scene Analysis in Real-time”, Proc. ICPR’94, November 1994, pp. 126-131

[11] Rota, N. and Thonnat, M., “Video sequence interpretation for visual surveillance”, Proc. of Third IEEE Int. Workshop on Visual Surveillance 2000, 2000, pp. 59-68.

[12] Karmann, K.P., and von Brant, A., "Moving object recognition using an adaptive background memory" in Time varying Image Processing and Moving Object Recognition, Elsevier Science B.V., 1990.

[13] Jiang,, C., and Ward, M.O., “Shadow identification”, Proc. of CVPR'92, pp. 606-612, 1992.

[14] Kilger, M., “A shadow handler in a video-based real-time traffic monitoring system”, Proc. of WACV'92, pp. 11-18, 1992.

[15] Stauder, J. and Mech, R. and Ostermann, J., “Detection of moving cast shadows for object segmentation”, IEEE Trans. on Multimedia, vol. 1 , n. 1, pp. 65-76, March 1999.

[16] Haritaoglu, I., Harwoord, D., and Davis, L.S., “W4: Real-time surveillance of people and their activities”, IEEE Trans. on Patt. Anal. and Machine Intell., vol. 22, no. 8, pp. 809-830, 2000.

[2] The authors wish to thank Prof. Mohan M. Trivedi and the staff of the CVRR (Computer Vision and Robotics Research) Lab, at UCSD for their support in testing the system.