Handling Occlusions in Monocular Surveillance Systems

Prithwijit Guha, Electrical Engg., IIT Kanpur
Nisarg Vyas, Robotics, IIT Kanpur
Amitabha Mukerjee, Computer Sc. & Engg., IIT Kanpur
K. S. Venkatesh, Electrical Engg., IIT Kanpur
{pguha,nisarg,amit,venkats}@iitk.ac.in

Abstract

Recent advances in computing machines and the availability of inexpensive vision sensors have paved the way for development of real-time imaging systems. Smart systems with a single fixed camera are often deployed for the task of outdoor surveillance. However, such systems are challenged by occlusions caused by interactions of foreground and background objects in the scene. In this paper, we propose an effective scheme for disambiguating such cases of occlusions and for the detection of entry and exit of objects using list based intelligent reasoning. The proposed methodology employs a statistical background model to identify foreground regions followed by smart tracking of the objects by using their color distribution and motion history. The present implementation runs at 7.5Hz while operating on color images of 320x240 resolution.

1. Introduction

Automated Video Surveillance deals with real time observation of people or vehicles in busy or restricted environments, leading to tracking and activity analysis of the subjects in the field of view. Potential applications range from monitoring in a car parking lot to military surveillance systems. Employing people for such a round-the-clock job is both expensive and prone to human error. These issues have led researchers to concentrate on Automated Video Surveillance [1, 4, 8, 9]. The first step of video surveillance consists of the detection and identification, as well as the real time process of determination of the trajectories (tracking) of the different foreground objects. The second step consists of activity analysis, which deals with the parsing of temporal sequences of detection and tracking observations to generate high-level action descriptions.

The detection of objects in the scene from a fixed camera requires foreground extraction. Elgammal et al. [7] have proposed a non-parametric background model for detecting intruders as the scene changes. Oren et al. [14] and Heisele et al. [10], on the other hand, adopt a different approach and identify pedestrians (intruders) by using wavelet templates and motion patterns respectively. Tracking algorithms estimate the trajectories of the intruders detected by foreground extraction, and constitute the second stage of system functionality. Computer Vision researchers have proposed a number of approaches for visual tracking - prominently, the Mean-Shift Algorithm [5], Kalman Filters [8], Optical Flow based methods [2] and Particle Filtering [11]. The spatio-temporal features thus extracted from the intruder images are further analyzed for recognizing the activities of the same.

In any surveillance activity, it is first necessary to monitor the entry/exit of subjects into/from the scene, in addition to tracking their movements when in the view. A monocular surveillance system is often seriously challenged in this task by foreground-foreground or foreground-background occlusions. In this paper, we extend the models for subject detection, monitoring of entry/exit, and tracking to the case of multiple intruders even in the presence of occlusions. This is achieved by a sequence of actions consisting of intruder detection by background subtraction, tracking through a robust combination of the the mean-shift algorithm and trajectory extrapolation based on the subjects' motion history, and disambiguation of occlusions through a process of intelligent reasoning. The presented results of experiments performed on a complicated motion pattern of the subjects indicate successful tracking even when the subjects merge and circle around each other.

This paper presents our work through the following sections. Section 2 describes the naive approach to foreground extraction along with shadow removal. In Section 3, we discuss the mean-shift algorithm used in tracking the subjects. Section 4 details the proposed methodology for intelligent reasoning to disambiguate occlusions. The results are presented in Section 5. Finally, we conclude our work in Section 6 and discuss possible future extensions and improvements.

2. Foreground Extraction

Consider the intensity value of a pixel over time in a completely static scene (i.e., with no background motion). In such a case, the only possible change that may occur to this pixel value is assumed to be caused by the zero mean Gaussian noise. Thus, it is quite legitimate to assume the intensity value P_{xy} of the $(x,y)^{th}$ pixel as a Gaussian $[N\{\mu_{xy}, \sigma^2_{xy}\}]$ random variable, where the mean value (μ_{xy}) and variance (σ^2_{xy}) are estimated from a time indexed training set of intensity values of that particular pixel. Similarly, one can assume such a set of distributions for the whole background image and hence derive the statistical model of the background image. A pixel of value P_{xy} is thus classified to be in the foreground region, if it has very less belongingness in the background pixel intensity distribution. Hence, the set of F foreground pixels is mathematically expressed as,

$$F = \{P_{xy} : |P_{xy} - \mu_{xy}| > k\sigma_{xy}\} \quad (1)$$

The parameter k decides the allowable limit of belongingness of background pixels and is typically assigned a value between 3 and 5. In practical cases, minor movements always persist among the background elements, thus defying the (perfect) static background assumption and this leads to background pixel classification error. A pixel neighborhood modeling approach performs much better. Elgammal et al. [7] have suggested a spatio-temporal modeling of background pixels through a non-parametric approach. Rudolf Mester et al. [13] have proposed an illumination invariant method for background classification that defines a new statistic instead of the traditional pixel intensity value. They have shown this statistic to follow a Gaussian distribution and used it to classify background pixels. However, both these methods are quite computation intensive. However, for surveillance applications, we only need to know the foreground region approximately for our analysis and would thus like to devote less time for foreground extraction. Our system thus opts for the standard (and simpler) method for background classification. However, this method often gives rise to single pixel misclassifications, which are removed by morphological operations, typically an erosion followed by a dilation. An example of foreground extraction by the simple background subtraction procedure is shown in Figure 1.

It is evident that such a method of foreground extraction often gives rise to classification error due to the shadow cast by the foreground object on the background region. This error occurs due to change in lighting conditions in certain backgrond pixels, a shadow being a special case of reduced illumination. A number of papers have addressed the issue of disambiguating the shadowed regions [3, 6, 12]. Javed et al. [12] have suggested the use of gradient direction information to achieve illumination invariance. Cucchiara et al. [6], on the other hand, use the hue-saturation components to suppress the shadowed regions. The video sequences used for our work are captured by a simple web cam and hence are very noisy. The approaches proposed by Javed et al. and Cucchiara et al. have shown poor performances on account of the noise. We have found that the scheme proposed by Branca et al. [3] provides far better results as compared to the other two while dealing with noisy sequences. Branca et al. assume the illumination change to be a slowly varying function and use the fact that the illumination changes are merely modulations of the pixel intensity values. Thus, if the intensity value P of a pixel changes to a new value S due to a cast shadow, then S is merely a multiple of P and can be mathematically expressed as,

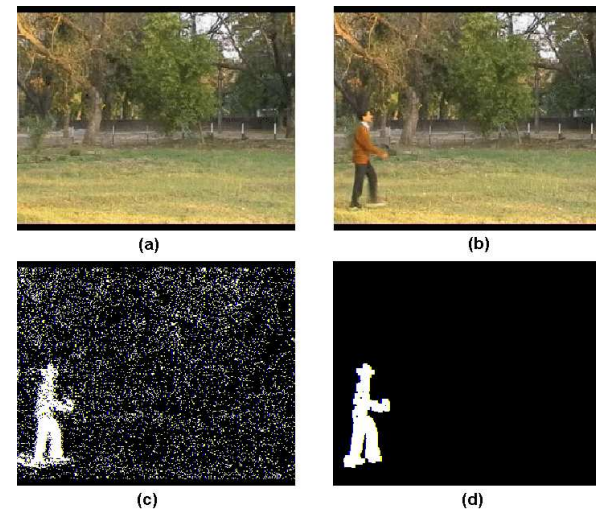

Figure 1. Results of Background Subtraction. (a) The background scene (b) An Intruder in the field of view causing a scene change (c) Background Subtraction Result (d) Refined foreground extraction after Morphological Operations

$$S = KP \quad and \quad K < 1 \quad (2)$$

In case of colored images, we need to compute the modulation factors in shadowed regions in the three different channels (Red, Green and Blue). Thus, if the image pixel intensity values are subscripted by *img* and that of the mean values of the corresponding learned background model pixels by *bg*, then the modulation factors for the different color components are computed as,

$$K_r = \frac{R_{img}}{R_{bg}}, \quad K_g = \frac{G_{img}}{G_{bg}}, \quad K_b = \frac{B_{img}}{B_{bg}} \quad (3)$$

Let, $K_N(x,y)$ be the set of the modulation factors (in all color channels) computed over a small neighborhood (typically 3x3) of the $(x,y)^{th}$ pixel. This particular pixel is

declared to be a shadowed one if all the modulation factors belonging to the set $K_N(x,y)$ are less than unity and the variance computed over the set elements is less than or equal to some small quantity ϵ. Thus, the conditions for the $(x,y)^{th}$ pixel to belong to a shadowed region can be mathematically expressed in the following manner.

$$\forall K \in K_N(x,y),\ \ [K \in (0,1)] \wedge [Var\{K\} \leq \epsilon] \quad (4)$$

In our case, only the pixels classified as foreground are reprocessed for shadow detection. Figure 2 illustrates the results of shadow removal after foreground extraction. Finally, the processed image is subjected to connected component analysis for identifying the separated blobs. Connected component analysis works by scanning an image, pixel-by-pixel (from top to bottom and left to right) in order to identify connected regions of adjacent pixels, which share the same set of intensity values. For our application, the image is binary (classified as either foreground or background) and an *8-connectivity* is assumed for separation of blobs.

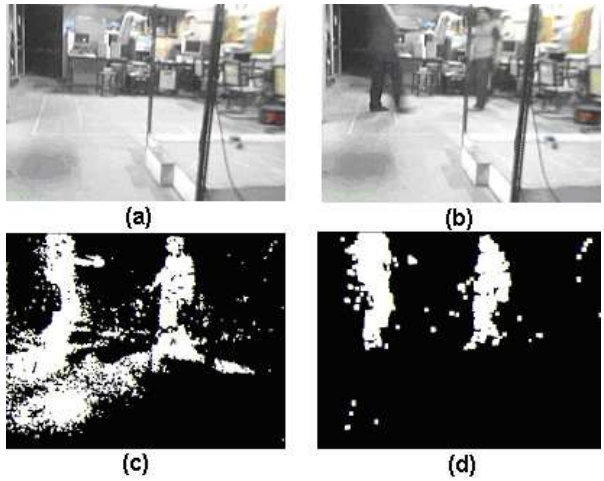

Figure 2. Results of Shadow Removal. (a) Background Image (b) Original Scene (c) Result of Background Subtraction, (d) Shadow Removal Applied on result in (c).

3. Foreground Region Tracking

Once extracted, the foreground regions need to be tracked for further analysis. In this work, we emphasize on maintaining an approximate estimate of the position and motion information of the intruder depicted as the foreground region. The computer vision community has used several algorithms for tracking objects through image sequences, the most noticeable ones being based on Kalman filters [8] and the CONDENSATION algorithm [11]. The Kalman filter based methods fail in many practical applications due to its assumptions of unimodality and linearity in motion and measurement. The CONDENSATION algorithm based approach (also known as particle filtering) overcomes these difficulties, but at the cost of high computation time. Recently, Comaniciu et al. [5] have proposed a novel method for tracking non-rigid objects based on the Mean-shift algorithm. They have proposed a color histogram based representation of the target model and a similarity measure based on the Bhattacharya coefficient for comparison of subject and target regions. The proposed algorithm typically assumes a monotonically radially decreasing kernel profile, going from unity magnitude at the center to zero value at the periphery and outside the target region. This provides us with a weighting function that gives maximum importance to the central pixels and assigns minimum belief to the peripheral ones. Such a weight assignment is also useful as the peripheral pixels are less trustworthy due to higher chances of occlusions and belongingness to background. The proposed algorithm has typically assumed the Epanechnikov kernel with an elliptical support over the target region. Thus, if the target region is centered at C, then the weight W_i of the pixel X_i will be given by,

$$W_i = \begin{cases} 1 - \|X_i - C\|^2; & \|X_i - C\| \leq 1 \\ 0; & \text{otherwise} \end{cases} \quad (5)$$

These weights are used while constructing the normalized target histogram H with m color bins $b_1, \ldots, b_m$. Such a representation suggests that the probability of occurrence of b_j in the target region is H_j. More so, the probability p_i that a particular pixel X_i with color value Q_i belongs to the target region is given by,

$$p_i = H_j, Q_i \in b_j \quad (6)$$

The mean-shift iterations start from an initial region centered at C_0 and gradually converges to the desired target region centered at C^*. The center-update rule in the k^{th} iteration is given as follows,

$$C_{k+1} = \frac{\sum_{i=1}^{N} X_i(k) p_i(k)}{\sum_{i=1}^{N} p_i(k)} \quad (7)$$

where $X_i(k)$ are the pixels in the current elliptical region (consisting of N pixels) centered at C_k and $p_i(k)$ are computed from target histogram H as given in equation 6. Comaniciu et al. [5] have used the Bhattacharya coefficient $B(H_k, H)$ as a measure of comparison of the normalized color distribution H_k (at the k^{th} step) and the target histogram H and is given by,

$$B(H_k, H) = \sum_{j=1}^{m} \sqrt{H_k(j), H(j)} \quad (8)$$

This algorithm is proved to converge [5] subject to proper choice of kernel function and is shown to maximize the

Bhattacharya coefficient at each step of the iterations. The current value of the Bhattacharya coefficient is used as a termination criteria and algorithm is seen to converge within 2 to 3 iterations in most of the cases. However, the mean-shift tracking algorithm assumes the target regions to be overlapping in subsequent frames and hence fails in cases of high acceleration. Thus, we also maintain a simple velocity-acceleration based dynamic model of the foreground region to account for its position information. Thus, the object's position $X(n)$ in the n^{th} frame is predicted using the position $X(n-1)$, velocity $V(n-1)$ and acceleration $A(n-1)$ information from the $(n-1)^{th}$ instant as,

$$X(n) = X(n-1) + V(n-1) + 0.5A(n-1) \quad (9)$$

4. List based Intelligent Reasoning

Monocular tracking of multiple subjects is very often challenged by partial visibility or sudden invisibility of the targets due to occlusions. A very common example of occlusion is the phenomenon of crossing of two persons. This event can be sub-divided into three different temporal stages. First, the two different persons approach each other and merge into a single foreground region: at this stage, one of them is fully visible and the other is only partially detectable. Second, complete occlusion, where one is fully visible and the other disappears completely. Finally, when they again tend to separate, the person who has disappeared previously, gradually becomes fully visible

Our system works on the above mentioned model of the occlusion phenomena. We also use the facts that the mean-shift algorithm can track objects under partial visibility and that motion history can predict current position of an object even if it is temporarily invisible. Thus, our scheme switches between a dynamic-model-based tracking and the mean-shift algorithm for disambiguating occlusions. Each subject in the field of view is represented by his color histogram and his motion history, which jointly constitute the signature of the target. The system typically maintains a list of such signatures and operates on the same to infer about phenomena like entry, exit, merge, split etc. In this section, we present the list based reasoning system and introduce the adopted nomenclature to perform the same. Let, the process of background subtraction detect n foreground regions and let there be m subject signatures in the list at the t^{th} frame. Let, $F_i(t)$ and $L_j(t)$ denote the i^{th} foreground region and the motion predicted minimum rectangular bounding region of the j^{th} subject in the list respectively. We define the extent of overlap between $L_j(t)$ and $F_i(t)$ by the ji^{th} element of the overlap matrix $O(t)$, given by,

$$Overlap[L_j(t), F_i(t)] = O_{ji}(t) = \frac{|L_j(t) \cap F_i(t)|}{|L_j(t)|} \quad (10)$$

This overlap matrix is further thresholded by a quantity α, to generate a new matrix T, whose elements are given by,

$$T_{ji}(t) = \begin{cases} 1; & O_{ji}(t) \geq \alpha \\ 0; & \text{otherwise} \end{cases} \quad (11)$$

Using this representation of T, we can further define two important quantities, viz. the number of subjects overlapped (denoted by $M_F(i,t)$) in the i^{th} foreground region and the number of foreground regions overlapped (denoted by $M_L(j,t)$) with the j^{th} subject. These quantities can be computed from the threshold overlap matrix T in the following manner.

$$M_F(i,t) = \sum_{j=1}^{m} T_{ji}(t) \;\; and \;\; M_L(j,t) = \sum_{i=1}^{n} T_{ji}(t) \quad (12)$$

The quantities computed in equation 12 are used to perform the following set of measurements,

$$E_1(j,t) = \begin{cases} 1; & Overlap(L_(j,t), W) < \eta \\ 0; & \text{otherwise} \end{cases} \quad (13)$$

$$E_2(j,t) = \begin{cases} 1; & M_L(j,t) > 0 \\ 0; & M_L(j,t) = 0 \end{cases} \quad (14)$$

$$E_3(j,t) = \begin{cases} 1; & \exists i : [T_{ji}(t) = 1] \wedge [M_F(i,t) > 1] \\ 0; & \text{otherwise} \end{cases} \quad (15)$$

$E_1(j,t)$ merely measures whether or not the j^{th} subject is within the reasoning region W, a subset of the image region presently under consideration. $E_1(j,t)$ is set to one if the extent of overlap between $L_j(t)$ and W falls below a certain level η. $E_2(j,t)$ indicates whether or not the j^{th} subject has shown up as a foreground region. If the j^{th} subject doesn't overlap with any of the extracted foreground regions, $E_2(j,t)$ is set to zero. $E_3(j,t)$ signifies the condition of merging. It is set to one if other subjects than the j^{th} subject have a significant overlap with a common foreground region. At this stage, we identify that each subject could be in one of the four following states,

$\mathbf{S}_0(j)$: j^{th} *subject exits the scene*

$\mathbf{S}_1(j)$: j^{th} *subject is isolated from the others*

$\mathbf{S}_2(j)$: j^{th} *subject in foreground-foreground ambiguity*

$\mathbf{S}_3(j)$: j^{th} *subject in foreground-background ambiguity*

$E_1(j,t)$ detects the *exit* state. For the state of being *isolated* , the j^{th} subject should have significant overlap with a foreground region within the scene and $E_3(j,t) = 0$. *Foreground-foreground ambiguities* occur when more than one subject merge in the same foreground region thereby causing difficulties in tracking each of them. A subject assumes this state when $E_3(j,t) = 1$ and $E_1(j,t) = 0$. A

subject is in *foreground-background ambiguity* when some portion of the background occludes it. A typical example could be the occlusion caused by a tree in an outdoor surveillance scenario. In this case, although the motion history predicts the subject to be within the image region, it remains invisible due to occlusion. This situation is indicated by $E_1(j,t) = 0$ and $E_2(j,t) = 0$. Table 1 summarizes the state decisions based on the composition of the values of the three measurements $E_l(j,t); l = 1,2,3$. Once the states of the individual subjects are decided, we proceed towards the actions to be taken for keeping track of the subjects. For this, we define the following four actions,

$\mathbf{A}_0(j,t)$: *j^{th} subject removed from list in t^{th} frame*

$\mathbf{A}_1(j,t)$: *update j^{th} subjects motion and color histogram information in t^{th} frame*

$\mathbf{A}_2(j,t)$: *update j^{th} subjects motion history using only mean shift tracking in t^{th} frame*

$\mathbf{A}_3(j,t)$: *update j^{th} subjects motion history using only position prediction in t^{th} frame*

The j^{th} subject is removed from the list $[\mathbf{A}_0(j)]$ once it is detected to be in the *Exit* state. However, if the subject is isolated from others, both its motion and color signatures are updated $[\mathbf{A}_1(j)]$ from the position and color histogram data of the foreground region with which it has the highest overlap. Since, the condition of *foreground-foreground ambiguity* makes it difficult to track subjects under partial visibility, we go for tracking each of them by the mean-shift algorithm and only the motion signature is updated $[\mathbf{A}_2(j)]$ according to the position information obtained thereby. The color signature is not updated, as it is tough to locate the exact contour of the subject in the common foreground where others have also merged. Finally, we come to the case of *foreground-background ambiguity* , where the subject is practically invisible while giving an approximate cue of its present position as predicted by the associated motion tracker. In this case, we rely on the predicted position and update the motion tracker based on the same. Table 1 summarizes the actions to be taken according to the present state of the subject under consideration.

After settling the issues related to tracking individual subjects and maintaining the list, we proceed to check the entry of new subjects. For this, we re-compute the overlap matrix and threshold it to obtain $M_F(i,t)$ while using the corrected (after reasoning and tracking) positions of the subjects. However, if there exists some i^{th} foreground region, which does not have any overlap with the listed subjects and it is within the scene, then it is either a new subject, or an old one for which we lost the track. Thus, we perform a search through the list to compare color histogram of the new region with the listed subjects. This region is declared new if no matches are found and is added to the list.

Table 1. State Decision-Action Table

State	*Observation*	*Action*
$\mathbf{S}_0(j,t)$	$E_1(j,t)$	$\mathbf{A}_0(j,t)$
$\mathbf{S}_1(j,t)$	$\overline{E_1(j,t)} \wedge E_2(j,t) \wedge \overline{E_3(j,t)}$	$\mathbf{A}_1(j,t)$
$\mathbf{S}_2(j,t)$	$\overline{E_1(j,t)} \wedge E_3(j,t)$	$\mathbf{A}_2(j,t)$
$\mathbf{S}_3(j,t)$	$\overline{E_1(j,t)} \wedge \overline{E_2(j,t)}$	$\mathbf{A}_3(j,t)$

5. Results

The proposed methodology is tested offline on a set of image sequences obtained from an outdoor surveillance setting. Here, we have arranged for two subjects to enter from two different sides of the scene. They are tracked individually as they approach each other and merge. We have experimented with a very complex motion pattern where the two subjects circle around each other. Here, although the tracking fails at several instants due to the complexity of the case, intelligent reasoning provides robustness to the system by resuming the track of subjects.

The view under surveillance is assumed to be free from any intrusion for the first few (typically 100) frames as the system starts up. Whenever an intrusion occurs, the system detects a change in the background and pixels belonging to the intruders' images are extracted as foreground regions. Initially, for the first three frames, the foreground region is only tracked using mean-shift algorithm while the motion history is being acquired. Subsequently, the color distribution and motion history of each subject are updated and saved in a dynamic list. Intelligent reasoning processes this list to handle cases of merging and splitting.

The proposed methodology for monocular visual surveillance is implemented on a standard 1.6 GHz Pentium 4 processor. The current implementation operates on images of resolution 320x240 at 7.5 FPS. Figure 3 shows the results of tracking two intruders crossing each other. An online version of this system is also implemented and is found to function satisfactorily in real-time in the lab environment.

6. Conclusions

In this paper, we have proposed an algorithm for monocular tracking of multiple objects while disambiguating occlusions through intelligent reasoning. This paper also reports the results obtained by the implementation of the proposed methodology. An intruder is recognized as a change of the background (foreground region). The surveillance system maintains a list of different intruders, where the color distribution and motion history form the signature of each. Our algorithm processes a dynamic list of these intruder signatures in an intelligent manner to resolve ambiguities caused by occlusions. The present work only con-

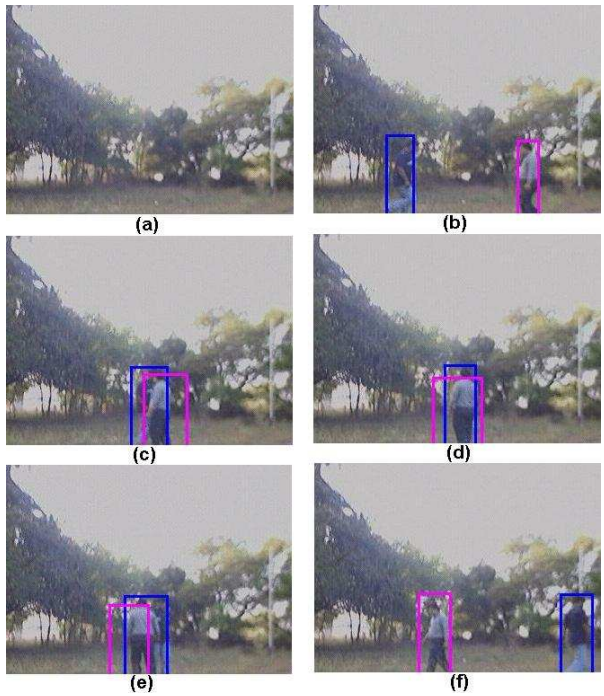

Figure 3. Results of Tracking Two Persons Crossing and Occluding each other. (a) The Background Image, (b) Two intruders approaching each other (Frame 155), (c) Intruders merge and one is partially visible (Frame 214), (d) One Intruder is Completely Occluded by the Other (Frame 235), (e) Intruders tend to separate and one is partially visible (Frame 261), (f) The Intruders have separated forming two different foreground regions (Frame 375). Bounding boxes of different colors (Pink and Blue) mark each intruder.

siders the color and motion information to distinguish foreground objects: including shape features will help differentiate multiple objects having similar color distribution.

This paper reports a significant part of our ongoing work on semantic analysis of surveillance videos. Future extensions to this work are aimed at analysis and recognition of intruder activities. Common actions like walking, standing, running and bending can be recognized from the analysis of foreground shape and motion information. Besides, we would like to identify foreground-background interactions like hiding (behind a tree, for example) and foreground-foreground interactions such as communication between two intruders, on the basis of occlusion information. The analysis of activities would help in scene understanding and will amount to a truly Smart Surveillance System.

References

[1] J. Batista, P. Peixoto, and H. Araujo. Real-time active visual surveillance by integrating peripheral motion detection with foveated tracking. In *Visual Surveillance*, pages 18–25, January 1998.

[2] A. Beghdadi, M. Auclair-Forteir, and J. Monteil. Tracking of image intensities based on optical flow: An evaluation of nonlinear diffusion process. In *Second IEEE International Symposium on Signal Processing and Information Technology*, pages 691–696, 2002.

[3] A. Branca, G. Attolico, and A. Distante. Cast shadow removing in foreground segmentation. In *International Conference on Pattern Recognition*, volume 1, pages 214–217, August 2002.

[4] 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.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition*, volume 2, pages 142–149, 2000.

[6] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti. Improving shadow suppression in moving object detection with hsv color information. In *Intelligent Transportation Systems*, pages 334–339, August 2001.

[7] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *Six'th European Conference on Computer Vision*, July 2000.

[8] J. M. Ferryman, S. J. Maybank, and A. D. Worrall. Visual surveillance for moving vehicles. *International Journal of Computer Vision*, 37:187–197, June 2000.

[9] I. Haritaoglu, D. Harwood, and L. Davis. W4 : Real time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:809–830, August 2000.

[10] B. Heisele and C. Woehler. Motion based recognition of pedestrians. In *International Conference on Pattern Recognition*, 1998.

[11] M. Isard and A. Blake. Condensation : Conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29:5–28, 1998.

[12] O. Javed, K. Shafique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *Motion and Video Computing*, pages 22–27, December 2002.

[13] R. Mester, T. Aach, and L. Dumbgen. Illumination-invariant change-detection using a statistical co linearity criterion. In *DAGM*, 2001.

[14] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Computer Vision and Pattern Recognition*, pages 193–199, June 1997.