Efficient Occlusion Handling for Multiple Agent Tracking by Reasoning with Surveillance Event Primitives

Prithwijit Guha
Dept. of Electrical Engg.
IIT Kanpur, India
Email: pguha@iitk.ac.in

Amitabha Mukerjee
Dept. of Computer Sc. & Engg.
IIT Kanpur, India
Email: amit@cse.iitk.ac.in

K.S. Venkatesh
Dept. of Electrical Engg.
IIT Kanpur, India
Email: venkats@iitk.ac.in

ABSTRACT

Tracking multiple agents in a monocular visual surveillance system is often challenged by the phenomenon of occlusions. Agents entering the field of view can undergo two different forms of occlusions, either caused by crowding or due to obstructions by background objects at finite distances from the camera. The agents are primarily detected as foreground blobs and are characterized by their motion history and weighted color histograms. These features are further used for localizing them in subsequent frames through motion prediction assisted mean shift tracking. A number of Boolean predicates are evaluated based on the fractional overlaps between the localized regions and foreground blobs. We construct predicates describing a comprehensive set of possible surveillance event primitives including entry/exit, partial or complete occlusions by background objects, crowding, splitting of agents and algorithm failures resulting from track loss. Instantiation of these event primitives followed by selective feature updates enables us to develop an effective scheme for tracking multiple agents in relatively unconstrained environments.

I. INTRODUCTION

Automated video surveillance deals with real time observation of people or vehicles in busy environments, leading to the activity analysis of the subjects (agents) in the field of view. The process of detection, identification and trajectory tracking of different agents are of prime importance in this context, as they are parsed further to generate higher level scene activity descriptors. A good overview of the challenges and developments in this area can be found in [1].

This paper deals with multiple agent tracking through a static camera in a surveillance scenario. Usual approaches to this problem deals with tracking blobs obtained from the process of background subtraction [2]. However, such blobs do not necessarily correspond to individual agents, as the agents can form a group and get detected as a single blob or an agent can be detected as multiple blobs due to occlusions. The W4 system [3] differentiates people from other objects by shape and motion cues and tracks them under occlusions by constructing appearance models and detecting body parts. Several researchers [4], [5] have employed particle filtering along with prior shape and motion models for multi-person tracking in cluttered scenes. Recently, Zhao et al. [6] proposed a Bayesian approach for tracking multiple persons under occlusions by computing MCMC based MAP estimates with prior informations about camera model and human appearance along with a ground plane assumption. McKenna et al. [7], on the other hand, presents a color based tracking algorithm that performs in relatively unconstrained environments and works at three levels of abstraction, viz. regions, people and groups.

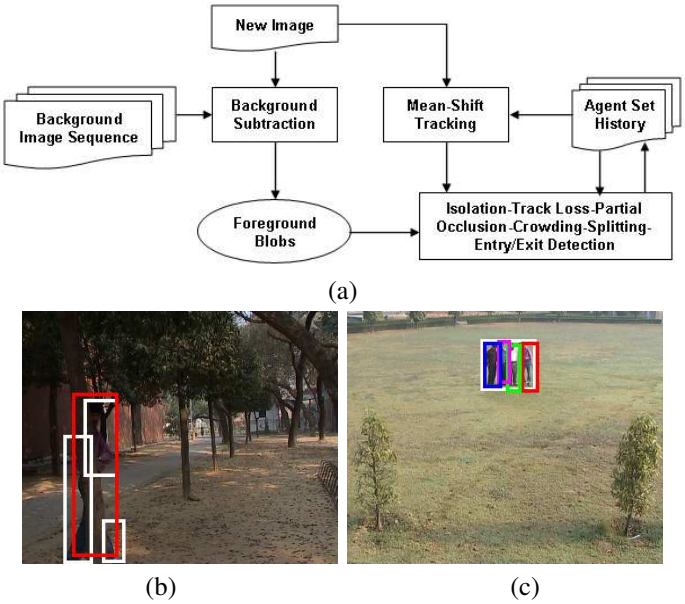

Fig. 1. Illustrating the proposed algorithm. (a) The flowchart of the proposed algorithm; (b) Partially occluded agent: the agent region is marked in red; (c) Four agents crowding together. The results of foreground detection in (b) and (c) are marked by white rectangles

The goal of this work is to develop a formal reasoning scheme for tracking multiple agents. The proposed approach is based on a non-evanescence assumption on the agents and that their deformations are continuous. The foreground blobs are initially detected by using a co-linearity statistic [8] based background subtraction methodology. The agents are localized in subsequent frames by motion prediction assisted mean-shift tracking [9]. The correspondences between the localized agents and detected foreground blobs are established by computing fractional overlaps which lead to the evaluation of certain Boolean predicates. The work aims at detecting and handling the following cases while tracking multiple agents under occlusions.

- Isolated, unoccluded and well tracked agents
- Losing track of agents
- Partial occlusions
- Crowding
- Identifying new regions
- Entry/Exit of agents
- Splitting of agents

The flowchart of the proposed algorithm along with a snapshot of the results for certain typical occlusion cases are shown in figure 1 for illustration purposes.

Our algorithm makes no constraining assumptions regarding the agent shape and motion models, ground plane etc. and demonstrates that intelligent reasoning applied on a set of easily computable predicates can provide a wealth of high level activity information. Some salient strengths of the proposed scheme are the following.

(a) Ability to distinguish a variety of problem situations.
(b) Using the above information to decide on the advisability of feature updates.
(c) Inherent ability of recognizing failure situations, should they occur.
(d) Ability to automatic track restoration at a later time
(e) No constraining assumptions related to agent shape and motion models, ground plane, etc.

This paper presents our work through the following sections. The co-linearity statistic based approach to background classification leading to the extraction of disjoint foreground blobs is discussed in section II. Section III briefly describes the features employed in characterizing the agents that facilitate localizing the same while tracking. The proposed scheme for tracking multiple agents by disambiguating occlusions using intelligent reasoning is explained in section IV. Section V presents the implementation details along with the experimental results for a few complex situations. Finally, we conclude our work in section VI and outline the possibilities of future extensions to the present work.

II. Foreground Blob Extraction

Efficient intrusion detection is the primary task of a smart surveillance system and is generally achieved by the process of background subtraction. Several techniques have been proposed in the domains of feature computations for background image representation and statistical modeling of the obtained feature vectors. The most commonly used approaches consider the intensity value in (normalized) RGB color space and/or image gradients [10] as the background image features and are statistically modeled through pixel-wise single Gaussian [2] or mixture of Gaussian [11] or the non-parametric approach with a Gaussian kernel function [12]. Alternate approaches through incremental PCA [13], MRF [14] and HMM [15] have also been proposed. Recently, Mester et al. [8] have proposed a novel approach to change detection under varying illumination. They have suggested the use of a co-linearity statistic built up on the assumption of a linear signal model. In this section, we briefly describe the co-linearity statistic based approach to foreground extraction.

The background classification system initializes by computing the reference image $\overline{\Omega}$ from the first T frames Ω_t $(t = 1, \ldots, T)$ converted to normalized RGB color space, which are assumed to be unintruded by any agent(s). Let, the rectangular neighborhood regions of the $(x, y)^{th}$ pixel position in $\overline{\Omega}$ and Ω_t be $\omega(x, y)$ and $\omega_t(x, y)$ respectively. Let, $\overline{v}_{xy}$ and $\overline{v}_{xyt}$ be the respective column vectors obtained by stacking the rows of $\omega(x, y)$ and $\omega_t(x, y)$. Now, if no structural changes occur in these rectangular windows, deviations from the reference vector can only happen due to multiplicative change in illumination (assumed to be equal over the rectangular region) and additive noise. However, it is to be noted that, neither the observed images and nor the reference vector provides us with the pure signal. Hence, both of them can be treated as an additive composition of the scalar modulation of the pure signal unit vector $\overline{u}_{xy}$ and white noise.

$$\overline{v}_{xy} = \kappa_{xy}\overline{u}_{xy} + \overline{\xi}_{xy}; \quad \overline{v}_{xyt} = \kappa_{xyt}\overline{u}_{xy} + \overline{\xi}_{xyt} \tag{1}$$

where, κ_{xy}, κ_{xyt} and $\overline{\xi}_{xy}$, $\overline{\xi}_{xyt}$ are the modulation and white noise components for $\overline{v}_{xy}$ and $\overline{v}_{xyt}$ respectively. Let us define the quantity d_{xyt} as the norm squared sum of the white noise components $\overline{\xi}_{xy}$, $\overline{\xi}_{xyt}$ and is given by,

$$d_{xyt} \stackrel{def}{=} \|\overline{\xi}_{xy}\|^2 + \|\overline{\xi}_{xyt}\|^2 \tag{2}$$

The co-linearity statistic c_{xyt} is defined as the minimum value of d_{xyt}, optimized with respect to $\overline{u}_{xy}$, and can be proved [8] to be the minimum eigen value $\lambda_{xyt}^{(min)}$ of the matrix $\mathbf{V}_{xyt}\mathbf{V}_{xyt}^T$, where the matrix $\mathbf{V}_{xyt}$ is defined as,

$$\mathbf{V}_{xyt} \stackrel{def}{=} \begin{pmatrix} \overline{v}_{xy} & \overline{v}_{xyt} \end{pmatrix}^T \tag{3}$$

The background model at the $(x, y)^{th}$ pixel position is thus learned by computing the mean $\mu_c(x, y)$ and the standard deviation $\sigma_c(x, y)$ of the co-linearity statistic c_{xyt} for the T training frames. The Chebyshev inequality [16] ensures that $(1 - \frac{1}{k^2})$ fraction of the values taken by the random variable lie within $(\mu_c(x, y) \pm k\sigma_c(x, y))$ irrespective of the distribution. Hence, we define the set of foreground pixels $\mathbf{F}_t$ for a new image Ω_t $(t > T)$ as,

$$\mathbf{F}_t = \{\Omega_t(x, y) : c_{xyt} \geq \mu_{xy} + k\sigma_{xy}\} \tag{4}$$

The value of $k = 5.0$ was empirically observed to provide the best performance. However, it is worth noting that the use of the co-linearity statistic doesn't ensure automatic shadow removal. A number of algorithms have addressed the issue of shadow removal. Javed et al. [10] have suggested the use of gradient direction information to achieve illumination invariance. Cucchiara et al. [17], on the other hand, use the hue-saturation components to suppress the shadowed regions. Recently, Branca et al. [18] have proposed an algorithm based on the assumptions of sub-unity photometric gain and the (almost) constancy of same in a small neighborhood of the shadowed area. This algorithm was found to provide comparatively better results and thus was adopted in our work. Shadow suppression

is only performed at the detected foreground pixels to achieve reduced computations and the resulting image is subjected to neighborhood voting corrections for classification noise removal. Figure 2 illustrates the results of co-linearity statistic based foreground extraction followed by the post-processing stages. Finally, the background-foreground segmented (binary) image at the t^{th} instant is subjected to connected component analysis (with 8-connectivity) to produce the set $\mathcal{F}_t$ consisting of n_t number of disjoint foreground blobs $F_i(t)$.

$$\mathcal{F}_t = \{F_i(t), i = 1, \ldots, n_t\} \tag{5}$$

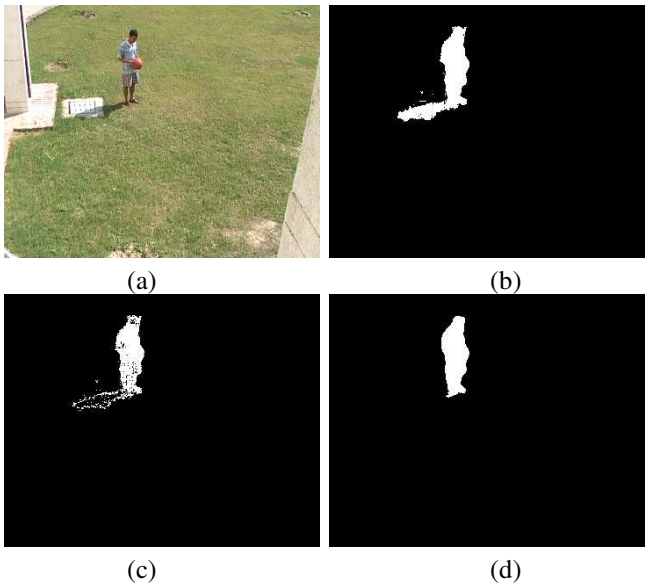

Fig. 2. Illustrating foreground blob extraction. (a) The original scene; (b) Result of co-linearity statistic based background subtraction; (c) Result of shadow suppression performed on the detected foreground pixels; (d) Foreground blobs after post-processing stages of neighborhood voting corrections.

III. Agent Characterization

The agents are primarily detected from the extracted foreground blobs and are initialized with the features computed from the blobs. The j^{th} agent is characterized by the collection of features $\mathcal{A}_j(t)$ at the t^{th} instant. The characteristic features of the agent include the set of pixels $A_j(t)$ it occupies, its weighted color distribution $h_j(t)$ and the trajectory of the center $c_j(t)$ of the minimum bounding rectangle of $A_j(t)$ for the last τ instants.

$$\mathcal{A}_j(t) = \langle A_j(t), h_j(t), \{c_j(t), \ldots, c_j(t-\tau+1)\}\rangle \tag{6}$$

The pixel set $A_j(t)$ and weighted color distribution $h_j(t)$ are initially learned from the foreground blob extracted at the first appearance of the agent and are updated through out the sequence when (s)he is isolated, unoccluded and well tracked. The color distribution $h_j(t)$ is computed from the b-bin color histogram of the region $A_j(t)$ (in Ω_t) weighted by the Epanechnikov kernel [9] supported over the minimum bounding ellipse of $A_j(t)$ (centered at $c_j(t)$) and is given by,

$$\begin{aligned} h_j[l](t) &= \frac{1}{C_E} \sum_{X \in A_j(t)} \mathbf{K_E}(\|X - c_j(t)\|^2)\delta(l - B_f(X)) \\ C_E &= \sum_{X \in A_j(t)} \mathbf{K_E}(\|X - c_j(t)\|^2) \end{aligned} \tag{8}$$

Where, C_E is the normalizing constant computed from the Epanechnikov kernel $\mathbf{K_E}$ and the function B_f maps the pixel location $X \equiv (x, y)$ to its corresponding color bin derived from $\Omega_t(x, y)$.

The agents in the t^{th} frame are typically localized by their trajectory information and color distribution obtained till the $(t-1)^{th}$ instant. An estimate $c_j^{(0)}(t)$ is obtained by extrapolating from the trajectory $\{c_j(t-1), \ldots, c_j(t-\tau)\}$. The mean-shift iterations [9], initialized at an elliptic region centered at $c_j^{(0)}(t)$ further localize the agent region at $A_j(t) \in \Omega_t$.

IV. Proposed Reasoning Scheme

In this section, we discuss the proposed reasoning scheme for tracking multiple agents in a surveillance scenario. A monocular surveillance system is typically challenged by the problems of occlusion due to unavailability of adequate depth information from a single view. In addition to tracking isolated agents and logging entries or exits, it should also be able to estimate approximate positions of the agents when occluded by another agent or a background object. Occlusions caused by agents occur in cases of crowding, where the system should approximately locate the individual agents while differentiating the occluded ones from the occluding ones. Furthermore, an agent can also get occluded completely by a pillar or a tree thereby disappearing from the set of detected foreground regions. More so, thinner background elements like a pole or tree branches can give rise to multiple foreground regions, which might confuse the system as an appearance of multiple agents. These different cases of occlusions are illustrated in figure 1. The proposed algorithm addresses each of these issues for keeping track of individual agents by processing low level image features guided by higher level intelligent reasoning.

It is worth noting that maintaining an exact track of all the agents is not always possible. As for example, when a person hides behind a tree, the system can only provide an approximate estimate where he has disappeared. Thus, the process of reasoning is performed over two sets, viz. the *active* and the *putative* set of agents. The active set $\mathcal{S}_A(t)$ consists of the agents that are well tracked till the t^{th} instant. On the other hand, the putative set $\mathcal{S}_P(t)$ contains the agents for which the system has lost track. The system typically initializes itself with empty sets and the agents are added or removed accordingly as they enter or leave the field of view. During the process of reasoning, the agents are often swapped between the active and putative sets as the track is lost or restored. We start the process of symbolic reasoning at the t^{th} frame based on the agent sets available from the $(t-1)^{th}$ instant.

The image region occupied by the j^{th} agent $\mathcal{A}_j(t-1)$ ($\mathcal{A}_j(t-1) \in \mathcal{S}_A(t-1), j = 1, \ldots, m_{t-1}$) in the t^{th} frame is predicted from its motion history information and is used for initializing a mean-shift algorithm to localize it further to $A_j(t)$. The process of association is established by computing the overlaps between the set of predicted agent regions and the extracted foreground regions. Equation 9 defines the *fractional overlap measure* $\gamma(\omega_1, \omega_2)$ between two regions ω_1 and ω_2, given by the fraction of ω_1 overlapped with ω_2. It is worth noting that $\gamma(\omega_1, \omega_2)$ is an asymmetric measure and lies in the interval $[0, 1]$.

$$\gamma(\omega_1, \omega_2) \stackrel{def}{=} \frac{|\omega_1 \cap \omega_2|}{|\omega_1|}; \tag{9}$$

We define the matrix $\Theta_{AF}(t)$ whose $(j,i)^{th}$ element ($j = 1, \ldots, m_{t-1}, i = 1, \ldots, n_t$) is set to 1 if the *localization confidence* of the j^{th} agent in $F_i(t)$, given by $\gamma(A_j(t), F_i(t))$ exceeds a certain threshold η_A and to 0 otherwise.

$$\Theta_{AF}[j,i](t) = \begin{cases} 1; & \gamma(A_j(t), F_i(t)) \geq \eta_A \\ 0; & \text{Otherwise} \end{cases} \tag{10}$$

Similarly, we can also define the matrix $\Psi_{FA}(t)$ whose $(i,j)^{th}$ element ($i = 1, \ldots, n_t; j = 1, \ldots, m_{t-1}$) is set to 1 if the *attribution confidence* of $F_i(t)$ to the j^{th} agent, given by $\gamma(F_i(t), A_j(t))$ exceeds a certain threshold η_F and to 0 otherwise.

$$\Psi_{FA}[i,j](t) = \begin{cases} 1; & \gamma(F_i(t), A_j(t)) \geq \eta_F \\ 0; & \text{Otherwise} \end{cases} \tag{11}$$

The number of foreground regions per agent ($\Theta_A[j](t)$) and the number of agents per foreground region ($\Theta_F[i](t)$) can be computed by summing up along the respective columns and rows of the matrix $\Theta_{AF}(t)$. Similar quantities ($\Psi_A[j](t)$ and $\Psi_F[i](t)$) can also be derived from $\Psi_{FA}(t)$. Hence, a number of measures can be defined from the thresholded localization and attribution confidence matrices and are given by,

$$\begin{aligned} \Theta_A[j](t) &= \sum_{i=1}^{n_t} \Theta_{AF}[j,i](t) \\ \Theta_F[i](t) &= \sum_{j=1}^{m_{t-1}} \Theta_{AF}[j,i](t) \\ \Psi_A[j](t) &= \sum_{i=1}^{n_t} \Psi_{FA}[i,j](t) \\ \Psi_F[i](t) &= \sum_{j=1}^{m_{t-1}} \Psi_{FA}[i,j](t) \end{aligned} \tag{12}$$

The proposed reasoning scheme is performed by extensively using the quantities introduced in equation 12 for handling the situations introduced in section I. The following subsections address each of these issues and defines the predicates for identifying the same.

A. Isolated, Unoccluded and Well Tracked Agents

The j^{th} agent in the active set $\mathcal{S}_A(t-1)$ is completely visible, if it is isolated from others and not occluded by any background objects. If this agent is properly tracked, then the localization confidence should be significantly high. Thus, the Boolean predicate $\text{ISOUNOCCTRK}_j(t)$ signifying the *isolated, unoccluded and well tracked* state of the j^{th} can be expressed as,

$$\text{ISOUNOCCTRK}_j(t) = \exists i[\Theta_{AF}[j,i](t) = 1] \wedge [\Theta_F[i](t) = 1] \tag{13}$$

Once the isolated and unoccluded status of the agent is ensured, its color distribution, pixel set and current position are updated so as to achieve best tracking performances in the subsequent frames.

B. Losing Track of Agents

The system might lose track of an agent due to two main reasons. Firstly, it may fail to associate the predicted agent region with any of the detected foreground regions, if the agent is occluded by larger background objects, like a tree or a pillar. Secondly, the mean-shift algorithm might fail to reach the detected foreground region corresponding to the agent due to inadequate motion information. In this case, the j^{th} agent under consideration will have no significant overlap with any of the detected foreground regions. Thus, the boolean predicate $\text{LOSTTRACK}_j(t)$ signifying the *track loss* of the j^{th} agent can be expressed as,

$$\text{LOSTTRACK}_j(t) = [\Theta_A[j](t) = 0] \wedge [\Psi_A[j](t) = 0] \tag{14}$$

The j^{th} agent is transferred from the active to the putative set as the system loses the track of the same and none of its current position or color distribution data are updated.

C. Partial Occlusions by Background Objects

In most practical cases, the field of view is not just a collection of objects at infinity. There might be objects like trees, pillars, walls, tables, etc. at finite distances from the camera which can occlude the agents in the scene. Such occlusions account for partial visibility or disappearance of agents in the scene. The system loses track of an agent as it is not detected by background subtraction under complete occlusion and is transfered to the putative set. However, a partially occluded agent is detected as either a single or a collection of disjoint foreground blobs which are the respective distorted or fragmented form of the unoccluded case. Figure 3(a)-(b) illustrates the cases of partial occlusions leading to fragmentation of detected foreground regions. The detected foreground blob(s) corresponding to the partially occluded agent $\mathcal{A}_j(t)$ exhibit overlap(s) with the localized image region $A_j(t)$. Thus, either of the predicates $[\Theta_A[j](t) \geq 1]$ or $[\Psi_A[j](t) \geq 1]$ can be used in this case.

Proposition 1: Partial occlusions are detected with higher confidence by the predicate $[\Psi_A[j](t) \geq 1]$ than $[\Theta_A[j](t) \geq 1]$.

Proof: Consider the case of the j^{th} agent getting fragmented into n disjoint foreground regions $F_i(t)$ $(i = 1, \ldots, n)$. The foreground blobs being a subset of the localized agent region, $|A_j(t)| > \sum_{i=1}^{n} |F_i(t)|$ and $|F_i(t) \cap A_j(t)| = |F_i(t)|$. Thus, the upper bound on the localization confidence $\gamma(A_j(t), F_i(t))$ can be derived as,

$$\gamma(A_j(t), F_i(t)) = \frac{|F_i(t)|}{|A_j(t)|} < \frac{|F_i(t)|}{\sum_{i=1}^{n} |F_i(t)|} = \frac{|F_i(t)|}{n\overline{F_t}} \quad (15)$$

Where, $\overline{F_t}$ is the average size of the foreground regions. Now, from the upper bound result obtained in 15, it can be easily shown that,

$$P(\gamma(A_j(t), F_i(t)) \geq \eta_A) < P(\frac{|F_i(t)|}{n|\overline{F_t}|} \geq \eta_A) \quad (16)$$

The expression $\frac{|F_i(t)|}{n|\overline{F_t}|}$ assumes non-negative values. Hence, using Markov inequality [16], we have,

$$\begin{aligned} P(\frac{|F_i(t)|}{n|\overline{F_t}|} \geq \eta_A) &< \frac{1}{\eta_A}\mathcal{E}(\frac{|F_i(t)|}{n|\overline{F_t}|}) \\ \Rightarrow P(\frac{|F_i(t)|}{n|\overline{F_t}|} \geq \eta_A) &< \frac{1}{n\eta_A} \\ \Rightarrow P(\gamma(A_j(t), F_i(t)) \geq \eta_A) &< \frac{1}{n\eta_A} \\ \Rightarrow P([\Theta_A(j) \geq 1]) &< \frac{1}{n\eta_A} \end{aligned} \quad (17)$$

Where, $\mathcal{E}$ denotes the expected value of a random variable. On the other hand, from the attribution confidence measure, we have,

$$\forall i \gamma(F_i(t), A_j(t)) = 1 \Rightarrow P[\Psi_A[j](t) \geq 1] = 1 \quad (18)$$

Thus, the predicate $[\Psi_A(j) \geq 1]$ is a stronger feature for detecting the partial occlusion of the j^{th} agent. However, to obtain the above as a stronger predicate, we should ensure that $\frac{1}{n\eta_A}$ always remain lesser than 1.0. Now, there can be a minimum of two fragments ($n = 2$) and hence we should always choose $\eta_A > 0.5$. The Boolean predicate $\text{IsoPartOcc}_j(t)$ signifying the situations of *isolation and partial occlusions* can be defined as,

$$\begin{aligned} \text{IsoPartOcc}_j(t) = \forall i [\Psi_{FA}[i,j](t) = 1] \wedge [\Theta_F[i](t) = 1] \\ \wedge [\Psi_A[j](t) \geq 1] \end{aligned} \quad (19)$$

The color distribution of a partially occluded agent being unreliable, only its current position is updated.

D. Crowding

The process of associating agents with the corresponding (detected) foreground regions is very often challenged by the phenomenon of crowding. In this case, multiple agents group together giving rise to a single foreground blob. More so, one agent might very often occlude others while crowding. Such events of merging of agent regions are very common in cases of agents crossing each other, standing together, etc. Figure 3(c)-(d) shows an example of a crowding condition, where a few agents merge to form a single foreground region. In such cases, multiple agents will have significant overlaps with a single foreground region. Thus, the i^{th} foreground region is detected to be congested by multiple agents if either of the predicates $[\Theta_F[i](t) > 1]$ or $[\Psi_F[i](t) > 1]$ is true.

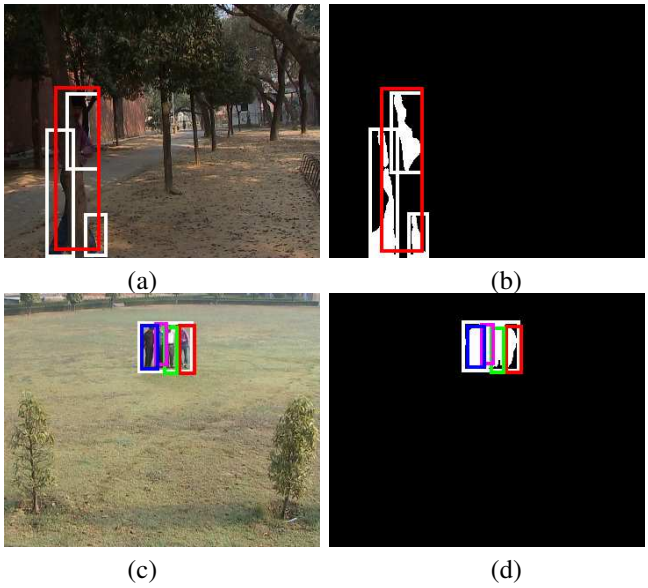

Fig. 3. Detecting partial occlusions and crowding (a) An agent is partially visible, being occluded by a tree; (b) Multiple foreground regions are detected as the result of background subtraction of (a); (c) A few agents stand together to form a group; (d) The detected foreground region corresponding to the agents in (c)

Proposition 2: Crowding is more efficiently detected by the predicate $[\Theta_F[i](t) > 1]$ than $[\Psi_F[i](t) > 1]$.

Proof: Consider the case of m agents crowding together at the foreground region $F_i(t)$. The agent regions being subsets of the foreground blob, $|F_i(t)| \leq \sum_{j=1}^{m} |A_j(t)|$ and $|F_i(t) \cap A_j(t)| = |A_j(t)|$. Thus, the lower bound on the attribution confidence measure can be derived as,

$$\gamma(F_i(t), A_j(t)) = \frac{|A_j(t)|}{|F_i(t)|} \geq \frac{|A_j(t)|}{\sum_{j=1}^{m} |A_j(t)|} = \frac{|A_j(t)|}{m\overline{A_t}} \quad (20)$$

Where, $\overline{A_t}$ is the average size of the agents in the crowd. This lower bound is achieved in the case of all agents grouping together, each being fully visible. Now, by using Markov inequality for this case, we have,

$$\begin{aligned} P(\gamma(F_i(t), A_j(t)) > \eta_F) &= P(\frac{|A_j(t)|}{m\overline{A_t}} \geq \eta_F) \\ P[\Psi_F[i](t) > 1] &\leq \frac{1}{m\eta_F} \end{aligned} \quad (21)$$

On the other hand, from the localization confidence measure, we have,

$$\forall j \gamma(A_j(t), F_i(t)) = 1 \Rightarrow P([\Theta_F[i](t) > 1]) = 1 \quad (22)$$

Thus, the condition $[\Theta_F[i](t) > 1]$ is a stronger feature for detecting crowding at the i^{th} foreground region. By similar reasoning, as in the case of partial occlusions, we can argue that we should always choose $\eta_F > 0.5$ to ensure the above proposition. The Boolean predicate $\text{CROWD}_i(t)$ signifying the case of *crowding* at the i^{th} foreground blob is given by,

$$\text{CROWD}_i(t) = [\Theta_F[i](t) > 1] \quad (23)$$

The individual agents in a crowd are never localized accurately with exact contour descriptions and thus the color distributions obtained from their current positions are not worth updating. Hence, only the current position is updated to keep continuous track through motion information.

E. Identifying New Regions

The occurrence of a new region is either caused by the entry of an agent or the re-appearance of one from the putative set. Thus, a new region $F_i(t)$ does not have any prior association (here overlap) with the agents in the active set $\mathcal{S}_A(t-1)$. Now, in case of the entry of an agent, a certain number of frames are required for him to appear completely and hence it is not a wise choice to learn his features from partial information available from the initial frames. Thus, an agent detected as the new region $F_i(t)$ is only added to the active set if its fractional overlap with an inner region ${}_i\Omega_t \subset \Omega_t$ (typically chosen by leaving a few border pixels of the image from each side) exceeds a certain threshold η_E. Hence, combining these conditions, the boolean predicate $\text{NEWREGION}_i(t)$ signifying the identification of a *new region* can be expressed as,

$$\text{NEWREGION}_i(t) = [\Theta_F[i](t) = 0] \wedge [\Psi_F[i](t) = 0] \wedge [\gamma(F_i(t),_i \Omega_t) > \eta_E] \quad (24)$$

Once it is ensured that, $F_i(t)$ is a new region, we match its features against the agents in the putative set $\mathcal{S}_P(t-1)$. If a match is found, the track of the corresponding agent is restored by re-initializing its features computed from $F_i(t)$ and is transfered to the active set. However, if no match is found, occurrence of $F_i(t)$ is accounted to the appearance of a new agent $\mathcal{A}_{m_t}$ $(m_t = m_{t-1} + 1)$ and is added to the active set.

F. Detecting Agent's Exit

The process of detecting an agent's exit from the field of view is of prime importance as the system needs to identify the agents to be removed from the active set. The exit of the j^{th} agent from the boundary of the image is detected when the fractional overlap between the region $A_j(t)$ and the inner region ${}_i\Omega_t$ falls below $1 - \eta_E$. Thus, the boolean predicate $\text{EXIT}_j(t)$ signifying the *exit* of the j^{th} agent is given by,

$$\text{EXIT}_j(t) = [\gamma(A_j(t),_i \Omega_t) < 1 - \eta_E] \quad (25)$$

Once an agent is detected to exit from the field of view, it is removed from the active set.

G. Splitting of Agents

A surveillance system often encounters the case of splitting, where two or more agents might enter the scene in a group and separate later within the field of view. In such a case, they will be initially learned as a single agent. A split occurs when the relative velocity between the separating agents is considerably high. It is worth noting that a split may eventually get detected as a fragmentation for a few frames, depending on the relative velocity between the separating agents. Afterward, the tracker converges on one of the agents and lose the others, which eventually emerge as new regions and are added as new agents. In some cases, it might also lose track of both, which are then learned as two new agents. Figure 4 shows the sequence of the event of split from initial stages of fragmentation detection to identification of a new region.

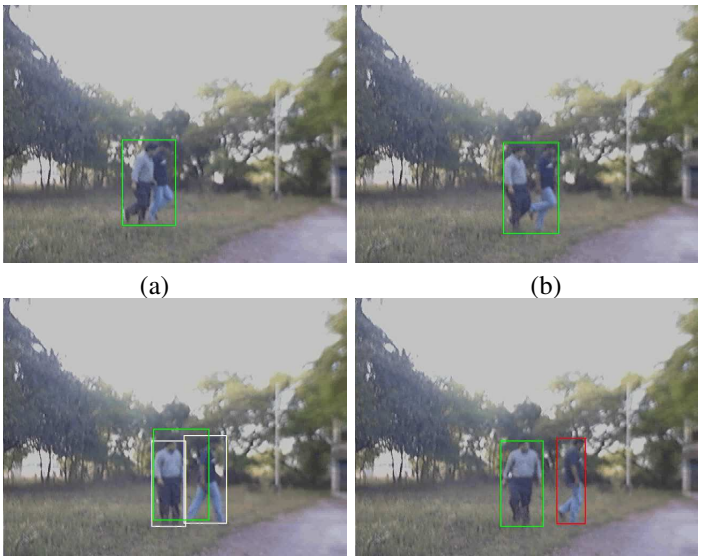

Fig. 4. *Splitting* of agents (a) Two agents entering the scene together (frame 360); (b) The agents tend to split (frame 370); (c) The system detects fragmentation even when the agents have separated (frame 372); (d) The new region is identified as the system loses track of one agent and is added as a new agent marked by red rectangle (frame 374).

V. RESULTS

The proposed methodology is tested offline on sets of image sequences obtained from outdoor surveillance settings. Here, we have arranged for a few persons playing a chasing game (firstly, in an open field and secondly, in an area with a number of trees), thereby tracing out complicated trajectories along with substantial occlusions among the agents themselves and with the background objects (trees). Here, although the tracking fails at several instants due to the complexity of motion and/or severe occlusions, intelligent reasoning provides robustness to the system by automatically resuming the track of the agents. The view under surveillance is assumed to be free from any intrusion for the first few $(T = 100)$ frames as the system initializes. Whenever an intrusion occurs, the system detects a change in the background and pixels belonging to

the intruder's image region are extracted as foreground regions. Initially, for the first few ($\tau = 4$) 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 set. Intelligent reasoning processes this set to detect the cases of several event primitives. The results of multi-agent tracking by the proposed approach under both dynamic and static occlusions are shown in figures 5 and 6 respectively.

The tracking performance of the j^{th} agent at the t^{th} instant is evaluated by the fraction of the ground-truth region of the same ($G_j(t)$) overlapped with the region $A_j(t)$, localized by the proposed algorithm and is thus given by the quantity $\gamma(G_j(t), A_j(t))$. Hence, if there are $m_g(t)$ number of agents present in the ground-truth marked images at the t^{th}, instant, then the overall performance $\mathcal{P}$ for a video of T frames is given by,

$$\mathcal{P} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m_g(t)}\sum_{j=1}^{m_g(t)}\gamma(G_j(t), A_j(t)) \qquad (26)$$

The above measure of overall performance $\mathcal{P}$ signifies the average fraction of the actual agent regions (or ground-truth regions) localized by the tracking algorithm in a certain video sequence. The overall performance varies, as the thresholds η_A and η_F are changed. It is evident from equations 10 and 11 that, as the thresholds η_A and η_F are increased, the detection rates of correspondences between predicted agent regions and foreground blobs are reduced. On the other hand, from sections IV-C and IV-D we learn that these thresholds should always be maintained above 0.5. To achieve optimal performances, we have chosen $\eta_A = \eta_F = 0.6$ and $\eta_E = 0.9$ and an overall tracking performance of 82.54% was observed. The proposed intelligent reasoning scheme for monocular visual surveillance is implemented on a standard 1.6 GHz Pentium-4 PC. The current implementation operates on images of resolution 320x240 at 7.5 FPS.

VI. Conclusion

In this paper, we have proposed an algorithm for multiple agent tracking while disambiguating occlusions through intelligent reasoning with a comprehensive set of surveillance event primitives. The system was found to track multiple agents satisfactorily in several complex situations. An agent is primarily detected as a change mask by the process of background subtraction. The surveillance system maintains a set of different agents, where the color distribution and motion history form the signature of each. Our algorithm processes a dynamic set of agent signatures in an intelligent manner to identify a variety of event primitives such as isolation, track loss, partial or complete occlusions, crowding, splitting and entry/exit. The proposed scheme is not restricted by any prior agent shape/motion models or ground plane assumptions and thus performs satisfactorily in relatively unconstrained environments.

This work reports a significant component of our ongoing project on semantic analysis of surveillance videos. Further enhancements to the multi-agent tracking algorithm aims at incorporating shape descriptors as agent signature and the use of Kalman filters along with a more sophisticated version of mean-shift tracking. The current algorithm is able to identify the reliability of feature updates along with various event primitives like crowding, partial occlusions, splitting and entry/exit. The future goals include parsing of temporal sequences of detected event primitives for the generation of scene activities thereby leading to a truly smart surveillance system.

References

[1] R. Collins, A. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 745–746, Aug 2000.

[2] C. R. Wren, A. Azarbayejani, T. Darell, and A. Penaland, "Pfinder: Real-time tracking of the human body," *Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780–785, July 1997.

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

[4] C. Needham and R. Boyle, "Tracking multiple sports players through occlusion, congestion and scale," in *Proceedings of the 12th British Machine Vision Conference*, 2001, pp. 93–102.

[5] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[6] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environments," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, July 2004, pp. 406–413.

[7] S. McKenna, S. Jabri, Z. Duric, and A. Rosenfeld, "Tracking groups of people," *Computer Vision and Image Understanding*, vol. 80, 2000.

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

[9] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Computer Vision and Pattern Recognition*, vol. 2, 2000, pp. 142–149.

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

[11] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, June 1999, pp. 246–252.

[12] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 90, pp. 1151–1163, July 2002.

[13] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh;, "Background modeling and subtraction of dynamic scenes," in *Proceedings of Ninth IEEE International Conference on Computer Vision*, vol. 2, October 2003, pp. 1305–1312.

[14] N. Paragios and V. Ramesh, "A mrf based approach for real-time subway monitoring," in *CVPR 2001: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE Computer Society, 2001, pp. 1034–1040.

[15] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. Buhmann, "Topology free hidden markov models: Application to background modeling," in *Proceedings of Eighth IEEE International Conference on Computer Vision*, vol. 1, 2001, pp. 294–301.

[16] B. Bhat, *Modern Probability Theory*. 2nd Edition, Halsted Press (John Wiley and Sons), 1981.

[17] 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*, August 2001, pp. 334–339.

[18] A. Branca, G. Attolico, and A. Distante, "Cast shadow removing in foreground segmentation," in *International Conference on Pattern Recognition*, vol. 1, August 2002, pp. 214–217.

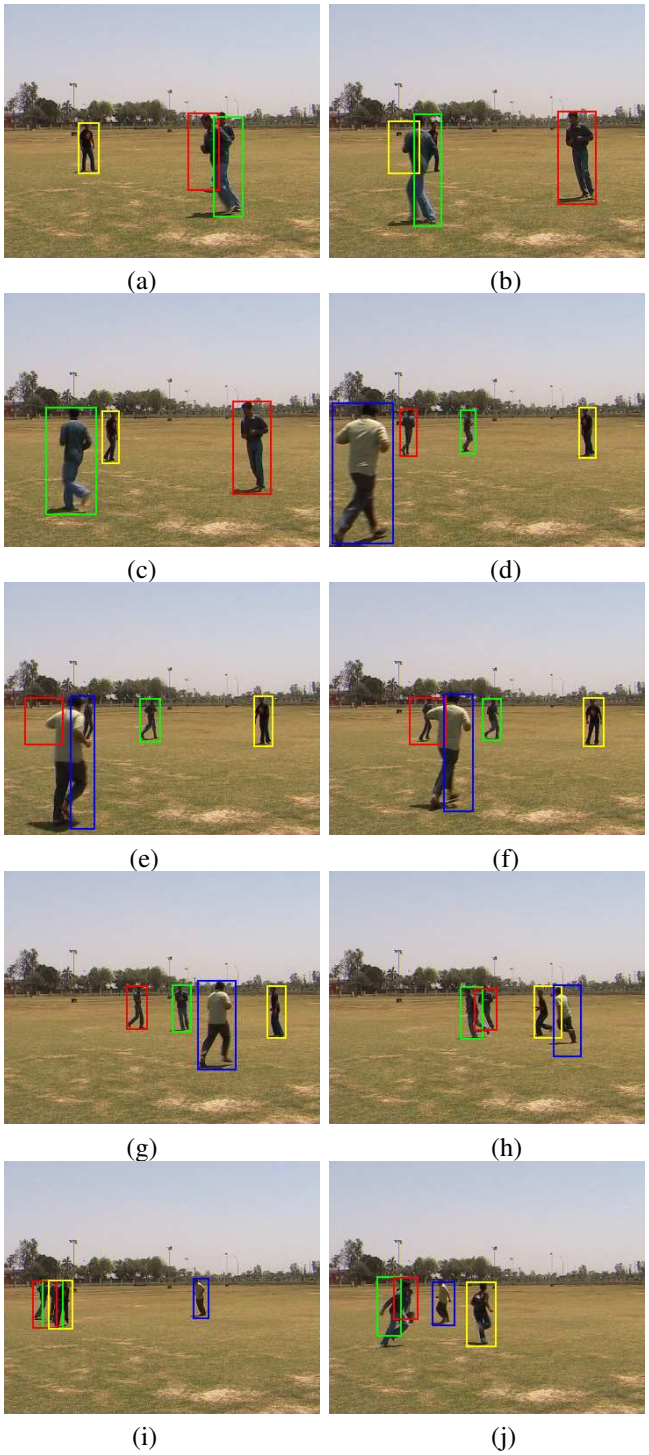

Fig. 5. Results of tracking four persons under dynamic occlusions, while they play a chasing game in an open field. Agent number 1, 2, 3 and 4 are marked with green, red, yellow and blue respectively. (a) Three agents in a circular motion, the first partially occluding the second (frame 1172); (b) First agent partially occluding the third and the system loses the track of the later (frame 1200); (c) The track of the third agent is automatically restored (frame 1205); (d) Fourth agent enters the scene (frame 1379); (e) The fourth agent partially occludes the second and the system loses the track of the later (frame 1384); (f) The track of the second agent is automatically resumed (frame 1395); (g) Four agents in a circular motion and are isolated from each other (frame 1420); (h) Fourth agent chases the others and all of them are properly tracked (frame 1457); (i) The agents are well tracked as the first three are in a group while partially occluding each other and the fourth one is isolated (frame 1528); (j) The system maintains the track as the chasing game continues with dynamic occlusions among agents (frame 1622).

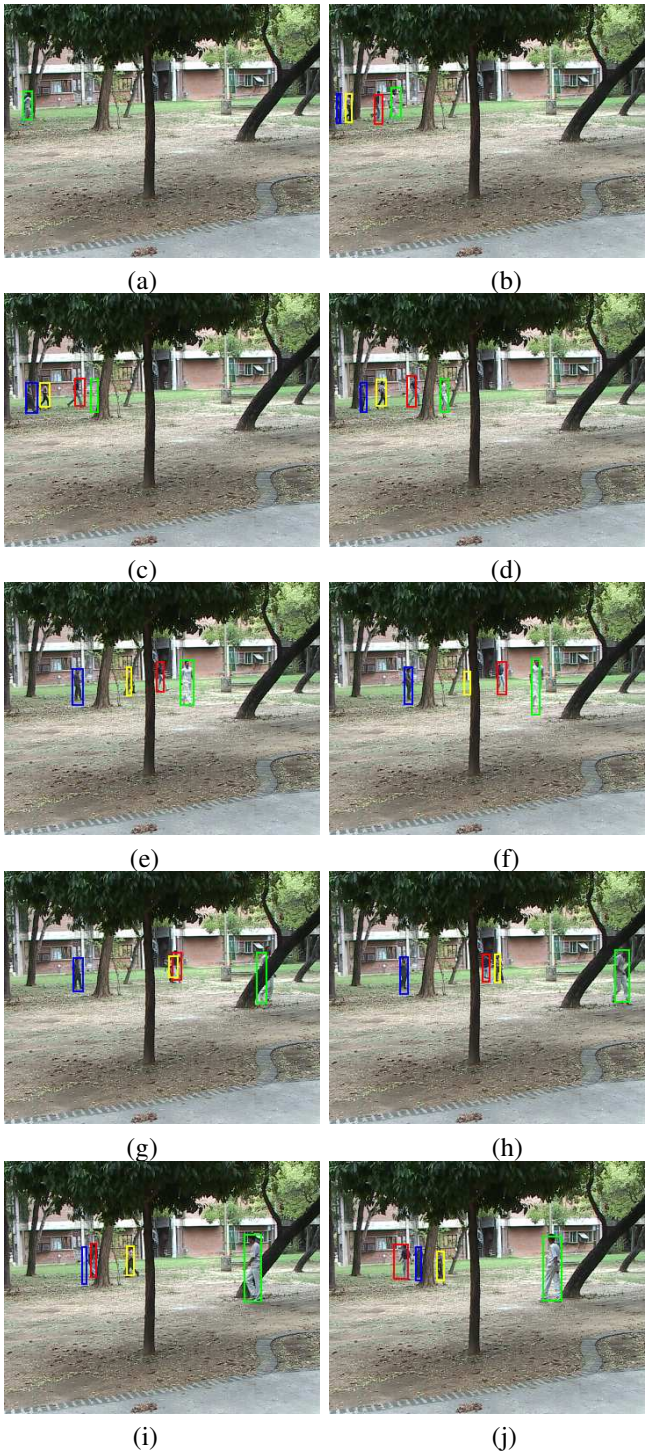

Fig. 6. Results of tracking four persons under static and dynamic occlusions. Agent #1, #2, #3 and #4 are marked with green, red, yellow and blue respectively. (a) Two agents enter the scene together (frame 897); (b) Agents #1 and #2 split and are distinguished while #3 and #4 enter (frame 917); (c) Agent #1 disappears behind a tree and the system loses his track (frame 934); (d) Agent #1 reappears and the track is restored automatically (frame 943); (e) Agent #3 is partially occluded by a tree and is tracked with some success (frame 1000); (f) The system loses track of agent #3 as he fully disappears behind a tree (frame 1019); (g) Agent #2 partially occludes #3 and both are well tracked, while #1 is partially occluded by a tree trunk and is tracked with some success (frame 1053); (h) Agent #4 reappears and his track is resumed automatically, the other three agents being isolated from each other (frame 1071); (i) Agent #2 partially occludes #4 and the track of the later is lost (frame 1130); (j) The track of agent #4 is automatically restored as he comes out of occlusion - all the agents are isolated and the system continues maintaining the track of each (frame 1148).