2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Multi-Target Visual Tracking with Game Theory-Based Mutual Occlusion Handling*

Xiaolong Zhou, Y.F. Li, *Senior Member, IEEE*, Bingwei He, and Tianxiang Bai

Abstract—**Tracking multiple moving targets in video is still a challenge because of mutual occlusion problem. This paper presents a Gaussian mixture probability hypothesis density-based visual tracking system with game theory-based mutual occlusion handling. First, a two-step occlusion reasoning algorithm is proposed to determine the occlusion region. Then, the spatial constraint-based appearance model with other interacting targets' interferences is modeled. Finally, an *n*-person, non-zero-sum, non-cooperative game is constructed to handle the mutual occlusion problem. The individual measurements within the occlusion region are regarded as the players in the constructed game competing for the maximum utilities by using the certain strategies. The Nash Equilibrium of the game is the optimal estimation of the locations of the players within the occlusion region. Experiments conducted on publicly available videos demonstrate the good performance of the proposed occlusion handling algorithm.**

I. Introduction

Tracking multiple moving targets in video is crucial in intelligent video surveillance system. It is helpful to activity analysis or high-level event understanding. However, the mutual occlusion problem makes it a challenge.

Recently, Gaussian mixture probability hypothesis density (GM-PHD) filter [1-3] to multi-target tracking in video has received considerable attention. Compared with the traditional association-based techniques, the difficulty caused by data association is avoided in the GM-PHD filter. However, the standard GM-PHD filter-based tracking system fails in tracking the individual targets when the mutual occlusion occurs among them. This paper focuses on proposing an effective algorithm to handle this problem.

Extensive methods [4-9] have been presented to solve the mutual occlusion problem. Though, the problem of tracking multiple interacting targets in mutual occlusion is still far from being completely solved and remains a challenge. For example, Xing et al. [5] build a dedicated observation model that maintains three discriminative cues including appearance, size and motion. The target appearance is modeled as the color histogram in HSV color space in discriminative region of the target. The mutual occlusion problem is then handled by a two-way Bayesian inference method. Vezzani et al. [6] generate two different images to represent the target model: the appearance image and a probability mask. The appearance image contains the RGB color of each point of the target and the corresponding probability mask reports their reliability. However, the appearance models proposed above cannot handle the situation when interacting targets have similar color distributions. To remedy this, Papadourakis and Argyros [7] model the target by using an ellipse and a Gaussian mixture model (GMM). The ellipse accounts for the position and spatial distribution of an object and a GMM represents its color distribution. The occlusion handling method proposed is based on both the spatial and the appearance components on a target's model. Similarly, Hu et al. [8] model the human body as a vertical ellipse and use the spatial-color mixture of Gaussian appearance model to model the spatial layout of the colors in a person. The occlusion is deduced using the current states of the interacting targets and handled using the proposed appearance model. However, the aforementioned appearance models do not consider the mutual interferences between the interacting targets, which may affect the tracking precision as mutual occlusion occurs. To remedy this, a robust appearance model that considers both the spatial constraint of the target and the interferences of the other interacting targets is proposed. Follow it, an effective mutual occlusion handling algorithm based on the game theory is proposed.

Game theory is the study of multi-person decision making, which was first proposed by Nash [10]. He stated that in non-cooperative games there exist sets of optimal strategies (so called Nash equilibrium) used by the players in a game such that no player can benefit by unilaterally changing his or her strategy if the strategies of the other players remain unchanged. Game theory has been applied to disciplines ranging from economics to engineering [11]. However, to the best of our knowledge, there are a few applications in visual tracking [4, 12] and fewer in mutual occlusion handling [4] based on the game theory. In this paper, we develop a GM-PHD filter-based system with game theory-based mutual occlusion handling to track multiple moving targets in video, especially to track the interacting targets in mutual occlusion.

The remainder of this paper is organized as follows. Section II presents basic knowledge on the GM-PHD filter. Section III first introduces the occlusion reasoning algorithm, and then describes the game theory-based mutual occlusion handling algorithm. Some experimental results on publicly available videos are discussed in Section IV, and followed by concluding remarks in Section V.

*Research supported by Research Grants Council of Hong Kong (Project No. CityU 118311), City University of Hong Kong (Project No. 7008176) and the grants from the National Natural Science Foundation of China (No. 51175087 and 61273286).

Xiaolong Zhou and Y.F. Li are with the Department of Mechanical and Biomedical Engineering of City University of Hong Kong, Hong Kong, (e-mail: mexlz@hotmail.com, meyfli@cityu.edu.hk,). Bingwei He is with the School of Mechanical Engineering and Automation of Fuzhou University, Fuzhou, China, (e-mail: mebwhe@fzu.edu.cn). Tianxiang Bai is with the Department of Research and Development, ASM Pacific Technology Ltd. Kwai Chung, Hong Kong, (e-mail: tianxiangbai@gmail.com).

II. GM-PHD FILTER

The kinematic state of a target i at time t is denoted as $\mathbf{x}_t^i=\{\mathbf{l}_t^i,\mathbf{v}_t^i,\mathbf{s}_t^i\}$. $\mathbf{l}_t^i=\{l_{x,t}^i,l_{y,t}^i\}$, $\mathbf{v}_t^i=\{v_{x,t}^i,v_{y,t}^i\}$ and $\mathbf{s}_t^i=\{w_t^i,h_t^i\}$ are the location, velocity and bounding box size of the target, respectively. $i=1,...,N_t$ and N_t is the number of targets at time t. Similarly, the measurement model of a target j at time t is denoted as $\mathbf{z}_t^j=\{\mathbf{l}_{z,t}^j,\mathbf{s}_{z,t}^j\}$. $j=1,...,N_{m,t}$ and $N_{m,t}$ is the number of measurements at time t. The target states set and the measurements set are denoted as $\mathbf{X}_t=\{\mathbf{x}_t^1,\ldots,\mathbf{x}_t^{N_t}\}$ and $\mathbf{Z}_t=\{\mathbf{z}_t^1,\ldots,\mathbf{z}_t^{N_{m,t}}\}$, respectively. The measurements are obtained by object detection. Any object detection method can be incorporated into our system. Because the contribution of this paper mainly focuses on handling the mutual occlusion problem, a simple background subtraction algorithm proposed in [13] is used to obtain the measurements.

This paper develops a visual tracking system based on the GM-PHD filter. In our visual tracking scenario, we assume that all targets consist of survival targets and newborn targets. According to [14], the GM-PHD filter for visual tracking is implemented by:

Prediction: Suppose the prior density $D_{t-1}(\mathbf{x}_{t-1})$ has the form: $D_{t-1}(\mathbf{x}_{t-1})=\sum_{i=1}^{J_{t-1}}\omega_{t-1}^{(i)}\mathcal{N}(\mathbf{x}_{t-1};\mathbf{m}_{t-1}^{(i)},\mathbf{P}_{t-1}^{(i)})$, then the predicted intensity $D_{t|t-1}(\mathbf{x}_t)$ is given by:

$$D_{t|t-1}(\mathbf{x}_t)=\gamma_t(\mathbf{x}_t)+p_{sv}\sum_{i=1}^{J_{t-1}}\omega_{t-1}^{(i)}\mathcal{N}(\mathbf{x}_t;\mathbf{m}_{sv,t|t-1}^{(i)},\mathbf{P}_{sv,t|t-1}^{(i)}) \quad (1)$$

where $\mathcal{N}(\cdot;\mathbf{m},\mathbf{P})$ denotes a Gaussian component with the mean $\mathbf{m}$ and the covariance $\mathbf{P}$. J_{t-1} and $\omega_{t-1}^{(i)}$ are the number and weight of the Gaussian component, respectively. $\mathbf{x}_t$ is the element of $\mathbf{X}_t$. $\gamma_t(\mathbf{x}_t)$ is the birth intensity of the newborn targets, which is determined by using the method proposed in [13]. p_{sv} is the survival probability of the survival targets.

Update: $D_{t|t-1}(\mathbf{x}_t)$ can be expressed as a Gaussian mixture: $D_{t|t-1}(\mathbf{x}_t)=\sum_{i=1}^{J_{t|t-1}}\omega_{t|t-1}^{(i)}\mathcal{N}(\mathbf{x}_t;\mathbf{m}_{t|t-1}^{(i)},\mathbf{P}_{t|t-1}^{(i)})$, then the posterior intensity $D_t(\mathbf{x}_t)$ is given by:

$$D_t(\mathbf{x}_t)=(1-p_d)D_{t|t-1}(\mathbf{x}_t)+\sum_{\mathbf{z}_t\in\mathbf{Z}_t}D_{g,t}(\mathbf{x}_t;\mathbf{z}_t) \quad (2)$$

$$D_{g,t}(\mathbf{x}_t;\mathbf{z}_t)=\sum_{i=1}^{J_{t|t-1}}\omega_{g,t}^{(i)}(\mathbf{z}_t)\mathcal{N}(\mathbf{x}_t;\mathbf{m}_{g,t}^{(i)}(\mathbf{z}_t),\mathbf{P}_{g,t}^{(i)}(\mathbf{z}_t)) \quad (3)$$

$$\omega_{g,t}^{(i)}(r_t,\mathbf{z}_t)=\frac{p_d\omega_{t|t-1}^{(i)}\mathcal{N}(\mathbf{z}_t;\mathbf{m}_{h,t}^{(i)},\mathbf{P}_{h,t}^{(i)})}{\lambda_t c_t(\mathbf{z}_t)+p_d\sum_{i=1}^{J_{t|t-1}}\omega_{t|t-1}^{(i)}\mathcal{N}(\mathbf{z}_t;\mathbf{m}_{h,t}^{(i)},\mathbf{P}_{h,t}^{(i)})} \quad (4)$$

where p_d is the detection probability and $\mathbf{z}_t$ is the element of $\mathbf{Z}_t$. λ_t is the average rate of the Poisson distributed clutters. $c_t(\mathbf{z}_t)$ is the probability density of the spatial distribution of clutters. $\mathbf{m}_{sv,t|t-1}^{(i)}=\mathbf{F}_t\mathbf{m}_{t-1}^{(i)}$, $\mathbf{P}_{sv,t|t-1}^{(i)}=\mathbf{Q}_t+\mathbf{F}_t\mathbf{P}_{t-1}^{(i)}\mathbf{F}_t^T$, $\mathbf{m}_{g,t}^{(i)}(\mathbf{z}_t)=\mathbf{m}_{t|t-1}^{(i)}+K(\mathbf{z}_t-\mathbf{H}_t\mathbf{m}_{t|t-1}^{(i)})$, $K=\mathbf{P}_{t|t-1}^{(i)}\mathbf{H}_t^T\bullet(\mathbf{H}_t\mathbf{P}_{t|t-1}^{(i)}\mathbf{H}_t^T+\mathbf{R}_t)^{-1}$,

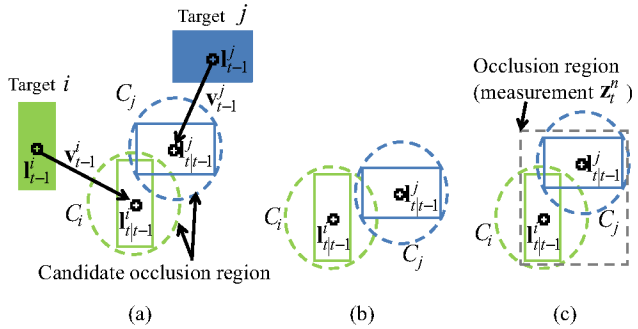

Fig. 1. Illustration of occlusion prediction and determination. (a) Occlusion prediction. (b) No occlusion. (c) Occlusion occurs and occlusion region is determined.

$\mathbf{P}_{g,t}^{(i)}(\mathbf{z}_t)=(\mathbf{I}-K\mathbf{H}_t)\mathbf{P}_{t|t-1}^{(i)}$, $\mathbf{m}_{h,t}^{(i)}=\mathbf{H}_t\mathbf{m}_{t|t-1}^{(i)}$, $\mathbf{P}_{h,t}^{(i)}=\mathbf{R}_t+\mathbf{H}_t\mathbf{P}_{t|t-1}^{(i)}\mathbf{H}_t^T$. $\mathbf{F}_t$ and $\mathbf{H}_t$ are the transition and the measurement matrices, respectively. $\mathbf{Q}_t$ and $\mathbf{R}_t$ are the covariance matrices of the process noises and the measurement noises, respectively.

It can be shown the number of components of the predicted and posterior intensities increases with time. This can be a problem in implementation. Therefore, we use the pruning and merging algorithms proposed in [14] to prune those components that are irrelevant to the target intensity and to merge those components that share the same intensity peak into one component. The peaks of the intensity are points of the highest local concentration of the expected number N_t of targets. The estimate of the multi-target states is the set of N_t ordered of the mean with the largest weights.

The standard GM-PHD filter-based visual tracking system fails in tracking individual targets when mutual occlusion occurs. To remedy this, an effective algorithm based on the game theory is proposed.

III. GAME THEORY-BASED MUTUAL OCCLUSION HANDLING

A. Occlusion Reasoning

A simple occlusion reasoning algorithm that includes occlusion prediction and determination is proposed.

1) **Occlusion prediction**: In Fig. 1(a), C_i (or C_j) is a circle at center $\mathbf{l}_{t|t-1}^i$ (or $\mathbf{l}_{t|t-1}^j$) with radius $\|\mathbf{s}_{t|t-1}^i\|$ (or $\|\mathbf{s}_{t|t-1}^j\|$). $\mathbf{l}_{t|t-1}^i$ and $\mathbf{s}_{t|t-1}^i$ are the location and size of the predicted target state $\mathbf{x}_{t|t-1}^i$, respectively. $\|\bullet\|$ is the Euclidean norm (hereinafter the same). The candidate occlusion region is predicted only when $C_i\bigcap C_j\neq\varnothing$ ($i\neq j$). That is,

$$\|\mathbf{l}_{t|t-1}^i-\mathbf{l}_{t|t-1}^j\|<\|\mathbf{s}_{t|t-1}^i\|+\|\mathbf{s}_{t|t-1}^j\| \quad (5)$$

Otherwise, no occlusion occurs.

2) **Occlusion determination**: Two situations are possible in the candidate occlusion region: no occlusion and occlusion (shown in Fig. 1(b) and 1(c), respectively). According to the fact that overlapping between the occlusion targets always increases gradually, the size of the first detected merged measurement is always larger than the size of the

corresponding single target. Therefore, if a measurement $\mathbf{z}_t^n = \{\mathbf{l}_{z,t}^n, \mathbf{s}_{z,t}^n\}$ ($n = 1, \ldots, N_{m,t}$) within the candidate occlusion region satisfies Eq. (6), this measurement is regarded as an occlusion region.

$$\left\|\mathbf{s}_{z,t}^n\right\| > \varepsilon \cdot \max\left\{\left\|\mathbf{s}_{t|t-1}^i\right\|, \left\|\mathbf{s}_{t|t-1}^j\right\|\right\} \tag{6}$$

where ε is a scale factor. The size of the detected target may slightly be changed between consecutive frames because of the changes in the target's pose or because of the depth of view. Compared with the size of the target before occlusion, the size of the target after occlusion is largely changed because it is merged with other targets. Consequently, ε is set to 1.2 to determine the occlusion region correctly.

B. Game Theory-Based Mutual Occlusion Handling

As mutual occlusion occurs, the measurement $\mathbf{z}_t^n$ is determined by the occlusion reasoning. Once it is determined, the identities and number N_o^n of the targets involved in this occlusion region are determined. To track them, we firstly model the appearances of the corresponding targets obtained at time $t-1$. We then propose an effective game theory-based algorithm to determine the locations of the targets in the occlusion region.

1) Target appearance modeling

The appearance of a target i is modeled as a GMM $q^i = q^i(\omega_k^i, \mu_k^i, \Sigma_k^i)$, representing the color distribution of target pixels. $k = 1, \ldots, K$ and $(\omega_k^i, \mu_k^i, \Sigma_k^i)$ represents the weight, mean and covariance matrix of the kth Gaussian component of the mixture. K is the number of Gaussian components. The measure of the similarity $P_s(p^i, q^i)$ between the candidate p^i (for a target i after occlusion) and the model q^i (for a target i before occlusion) is defined as the probability that p^i's colors are drawn from q^i [9]:

$$P_s(p^i, q^i) = \exp\left\{\frac{1}{N_i}\sum_{\Omega_i}\log\left\{\sum_{k=1}^{K}\omega_k^i N(c_{\mathbf{l}^i}; \mu_k^i, \Sigma_k^i)\right\}\right\} \tag{7}$$

where $c_{\mathbf{l}^i} = (r_{\mathbf{l}^i}, g_{\mathbf{l}^i}, I_{\mathbf{l}^i})$ is the color of the pixel located in $\mathbf{l}^i$ within the support region Ω_i of target i. N_i is the number of foreground pixels in Ω_i. $g_{\mathbf{l}^i} = G_{\mathbf{l}^i}/(R_{\mathbf{l}^i} + G_{\mathbf{l}^i} + B_{\mathbf{l}^i})$, $r_{\mathbf{l}^i} = R_{\mathbf{l}^i}/(R_{\mathbf{l}^i} + G_{\mathbf{l}^i} + B_{\mathbf{l}^i})$ and $I_{\mathbf{l}^i} = (R_{\mathbf{l}^i} + G_{\mathbf{l}^i} + B_{\mathbf{l}^i})/3$.

However, the aforementioned appearance model may fail when targets have similar color distributions. For such case, a Gaussian spatial constraint is applied according to [15]. Aside from similar color distributions, interferences by other interacting targets p^j within the occlusion region are also need to be considered. To take the above points into account, the measure of the similarity is then improved as:

$$P_s(p^i, q^i \,|\, p^j) = \exp\left\{\frac{1}{N_o^n}\left[\sum_{\Psi_1}\log\left(N(\mathbf{l}_t^i)\cdot\sigma_1\right) + \sum_{\Psi_2}\frac{\sigma_1}{\sigma_2}\log\left(N(\mathbf{l}_t^i)\cdot\sigma_1\right)\right]\right\} \tag{8}$$

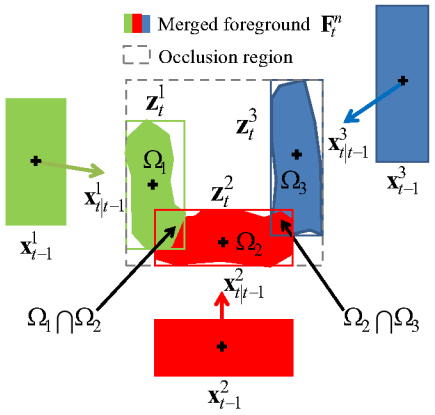

Fig. 2. A schematic diagram of mutual occlusion handling.

where $\sigma_1 = \sum_{k=1}^{K}\omega_k^j N(c_{\mathbf{l}^i}; \mu_k^j, \Sigma_k^i)$, $\sigma_2 = \sum_{j=1}^{N_o^n}\sum_{k=1}^{K}\omega_k^j N(c_{\mathbf{l}^i}; \mu_k^j, \Sigma_k^j)$, $N(\mathbf{l}_t^i) = N(\mathbf{l}^i; \mathbf{l}_t^i, \Sigma_t^i)$, $\Psi_1 = \Omega_i - \Omega_i \cap \Omega_j$, $\Psi_2 = \Omega_i \cap \Omega_j$ and $\Sigma_t^i = [(w_t^i/2)^2, 0; 0, (h_t^i/2)^2]$.

In Fig. 2, given the merged foreground $\mathbf{F}_t^n$ and the predicted target states $\{\mathbf{x}_{t|t-1}^i\}_{i=1}^{N_o^n}$ ($N_o^n = 3$ in Fig. 2), the goal is to obtain the optimal individual measurement of each target within the occlusion region. In other words, the optimal solution $(\mathbf{z}_t^{1*}, \ldots, \mathbf{z}_t^{N_o^n *})$ is obtained by maximizing the similarity probability between the $(\mathbf{z}_t^1, \ldots, \mathbf{z}_t^{N_o^n})$ and the $\mathbf{F}_t^n$.

$$\begin{aligned}(\mathbf{z}_t^{1*}, \ldots, \mathbf{z}_t^{N_o^n *}) &= \arg\max_{\{\mathbf{z}_t^i\}_{i=1}^{N_o^n}} P(\mathbf{z}_t^1, \ldots, \mathbf{z}_t^{N_o^n} \,|\, \mathbf{F}_t^n) = \arg\max_{\{\mathbf{z}_t^i\}_{i=1}^{N_o^n}} P(\mathbf{z}_t^i, \mathbf{z}_t^{-i} \,|\, \mathbf{F}_t^n) \\ &= \arg\max_{\{\mathbf{z}_t^i\}_{i=1}^{N_o^n}} P(\mathbf{z}_t^i \,|\, \mathbf{F}_t^n, \mathbf{z}_t^{-i}) P(\mathbf{z}_t^{-i} \,|\, \mathbf{F}_t^n)\end{aligned} \tag{9}$$

where $\mathbf{z}_t^{-i} = \{\mathbf{z}_t^j\}_{j=1, j\neq i}^{N_o^n}$. To obtain the optimal solution of Eq. (9), a game theory-based mutual occlusion handling algorithm is proposed to bridge the joint measurements estimation and the Nash equilibrium of a game.

2) Game theory-based mutual occlusion handling

In game theory [10], a non-cooperative game is one in which players make decisions independently. As mutual occlusion occurs, the individual measurements involved in the occlusion region compete to independently maximize the similarity probability between the measurements and the foreground. Therefore, it is reasonable to construct a non-cooperative game to bridge the joint measurements estimation and the Nash equilibrium. In this paper, we construct an *n*-person, non-zero-sum, non-cooperative game and assume that the target's size keeps the constant during the occlusion. The estimation of measurements $(\mathbf{z}_t^1, \ldots, \mathbf{z}_t^{N_o^n})$ is simplified as the estimation of the locations $(\mathbf{l}_{z,t}^1, \ldots, \mathbf{l}_{z,t}^{N_o^n})$ of the measurements. The player, strategy and utility are defined as follows: 1) **Player:** The individual measurement $\mathbf{z}_t^i$ originating from the target i within the occlusion region; 2) **Strategy:** Motion of the player, i.e. the location

$\mathbf{l}_{z,t}^{i} = \{l_{x,t}^{i}, l_{y,t}^{i}\} \in \mathbb{R}^2$ of the player. 3) **Utility:** $U^i(\mathbf{l}_{z,t}^{i}, \mathbf{l}_{z,t}^{-i}) = P(\mathbf{z}_t^i \,|\, \mathbf{F}_t^n, \mathbf{z}_t^{-i}) P(\mathbf{z}_t^{-i} \,|\, \mathbf{F}_t^n)$, $\mathbf{l}_{z,t}^{-i} = \{\mathbf{l}_{z,t}^{j}\}_{j=1, j\neq i}^{N_o^n}$.

To find a Nash equilibrium of the game, the *best-response* should be defined first.

Definition 1 [11]: The *best-response* of a player i to the strategies $\mathbf{l}_{z,t}^{-i}$ is the strategy for that player such that:

$$U^i(\mathbf{l}_{z,t}^{i}, \mathbf{l}_{z,t}^{-i}) \geq U^i(\mathbf{l}_{z,t}^{i'}, \mathbf{l}_{z,t}^{-i}), \ \forall \mathbf{l}_{z,t}^{i'} \in \mathbb{R}^2 \tag{10}$$

Definition 2 [11]: $(\mathbf{l}_{z,t}^{1*}, \ldots, \mathbf{l}_{z,t}^{N_o^n *})$ is a Nash equilibrium for the game with utility $\{U^i(\mathbf{l}_{z,t}^{i}, \mathbf{l}_{z,t}^{-i})\}_{i=1,\ldots,N_o^n}$, if every player's strategy is a *best-response* to the other players' strategies:

$$U^i(\mathbf{l}_{z,t}^{i*}, \mathbf{l}_{z,t}^{-i*}) \geq U^i(\mathbf{l}_{z,t}^{i}, \mathbf{l}_{z,t}^{-i*}), \text{ for every player } i \tag{11}$$

Given $\mathbf{l}_{z,t}^{-i*}$, the goal is to determine the *best-response* of the player i. That is:

$$\max U^i(\mathbf{l}_{z,t}^{i}, \mathbf{l}_{z,t}^{-i*}) \propto \max P(\mathbf{z}_t^i \,|\, \mathbf{F}_t^n, \mathbf{z}_t^{-i*}) \tag{12}$$

Maximizing the $P(\mathbf{z}_t^i \,|\, \mathbf{F}_t^n, \mathbf{z}_t^{-i*})$ is equal to maximizing the measure of similarity $P_s(p^i, q^i \,|\, p^i)$. To obtain the $\mathbf{l}_{z,t}^{i*}$, we set the derivative of $P_s(p^i, q^i \,|\, p^i)$ with respect to $\mathbf{l}_{z,t}^{i}$ to zero.

$$\sum_{\Psi_1} \mathbf{l}^i - N_i' \bullet \mathbf{l}_{z,t}^{i} + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \bullet \mathbf{l}^i - \mathbf{l}_{z,t}^{i} \bullet \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} = 0 \tag{13}$$

where N_i' is the number of foreground pixels in the support region Ψ_1. $\mathbf{l}_{z,t}^{i}$ is calculated by Eq. (13) and regarded as the *best-response* $\mathbf{l}_{z,t}^{i*}$ of the player i:

$$\mathbf{l}_{z,t}^{i*} = \left(\sum_{\Psi_1} \mathbf{l}^i + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \bullet \mathbf{l}^i \right) \Big/ \left(N_i' + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \right) \tag{14}$$

The location $\mathbf{l}_{z,t}^{i}$ of the player i is initialized by the corresponding predicted target's location $\mathbf{l}_{t|t-1}^{i}$. Given the initialized locations $\mathbf{l}_{z,t}^{-i*}$, the *best-response* $\mathbf{l}_{z,t}^{i*}$ of the player i can be calculated by Eq. (14). $\mathbf{l}_{z,t}^{i*}$ is iteratively updated until the process reaches an equilibrium. The equilibrium is obtained when the maximum component of the difference vector $\Delta\mathbf{l}$ satisfies Eq. (15). $\Delta\mathbf{l}$ is the difference of the *best-response* sets between consecutive iteration cycles.

$$\max(\Delta\mathbf{l}) < T_{NE} \tag{15}$$

where $\Delta\mathbf{l} = \left| \{\mathbf{l}_{z,t}^{1*}, \ldots, \mathbf{l}_{z,t}^{N_o^n *}\}_{iteration\ j} - \{\mathbf{l}_{z,t}^{1*}, \ldots, \mathbf{l}_{z,t}^{N_o^n *}\}_{iteration\ j-1} \right|$. $\{\mathbf{l}_{z,t}^{1*}, \ldots, \mathbf{l}_{z,t}^{N_o^n *}\}_{iteration\ 0}$ is the initialized locations set. T_{NE} is the given threshold. The smaller T_{NE} is, the more iteration time needed, while the more precise results obtained. In our experiments, we set T_{NE} =1 pixel to achieve a trade-off between the efficiency and the precision. When the iteration terminates at iteration cycle j, the final *best response* set $\{\mathbf{l}_{z,t}^{1*}, \ldots, \mathbf{l}_{z,t}^{N_o^n *}\}_{iteration\ j}$ is determined as the Nash equilibrium of the game. This Nash equilibrium is regarded as the optimal locations of the measurements. The measurements $(\mathbf{z}_t^{1*}, \ldots, \mathbf{z}_t^{N_o^n *})$ with the $\{\mathbf{l}_{z,t}^{1*}, \ldots, \mathbf{l}_{z,t}^{N_o^n *}\}_{iteration\ j}$ are then incorporated into the GM-PHD filter to update the states of the targets within the occlusion region.

IV. Experiments and Discussions

In the experiments, the state transition model is a constant velocity model with: $\mathbf{F}_t = [\mathbf{I}_2, T\mathbf{I}_2, \mathbf{0}_2; \ \mathbf{0}_2, \mathbf{I}_2, \mathbf{0}_2; \mathbf{0}_2, \mathbf{0}_2, \mathbf{I}_2]$, $\mathbf{Q}_t = \sigma_v^2 [T^4\mathbf{I}_2/4, T^3\mathbf{I}_2/2, \mathbf{0}_2; T^3\mathbf{I}_2/2, \ T^2\mathbf{I}_2, \mathbf{0}_2; \mathbf{0}_2, \mathbf{0}_2, T^2\mathbf{I}_2]$, where $\mathbf{0}_n$ and $\mathbf{I}_n$ are the $n \times n$ zero and identity matrices. $T = 1$ frame, is the interval between two consecutive time steps. $\sigma_v = 3$ is the standard deviation of the state noise. The measurements follow the measurement likelihood with: $\mathbf{H}_t = [\mathbf{I}_2, \mathbf{0}_2, \mathbf{0}_2; \mathbf{0}_2, \mathbf{0}_2, \mathbf{I}_2]$, $\mathbf{R}_t = \sigma_w^2 \mathbf{I}_4$, where $\sigma_w = 2$ is the standard deviation of the measurement noise. The values of the parameters used in the GM-PHD filter are: p_d=0.99, p_{sv}=0.95, λ_t=0.01 and $c_t(\mathbf{z}_t)$=(image area)$^{-1}$.

The proposed system is evaluated on the following publicly available videos: 415 frames from the 'ViSOR'[1], 951 frames from the 'PETS2006'[2], 922 frames from the 'CAVIAR'[3], and 795 frames from the 'PETS2009'[4].

A. Qualitative Analysis

We compare our *improved tracking system (ITS)* with the *standard GM-PHD filter-based tracking system (STS).*

In Fig. 3 (for 'ViSOR'), two targets with similar color distributions are involved. The numbers involved in the figure are the targets' identities (IDs), which are managed using the algorithm proposed in [14]. Without occlusion handling, the *STS* fails in tracking the target 1 while tracks the merged measurement as the other target 2 from t=166. As targets split, target 1 is re-tracked as a newborn target 3 from t=245. On the contrary, the *ITS* can succeed in tracking them during the whole occlusion period even when they are in total occlusion (shown as t=191 in Fig. 3).

In Fig. 4 (for 'PETS2006'), targets frequently merge and split. The *STS* loses the targets (shown as t=508 and t=777 in Fig. 4) as mutual occlusion occurs (targets merge), and then re-tracks the merged measurement as a newborn target (shown as t=778 in Fig. 4). Similarly, as they split, they are tracked as the newborn targets. On the contrary, the *ITS* performs robustly no matter that targets merge or split. In particular, the *ITS* can handle the situation when three targets with similar color distributions merge together (shown as t=777 and t=778 in Fig. 4).

Similarly in Fig. 5 (for 'CAVIAR') and Fig. 6 (for 'PETS2009'), the *STS* fails in tracking the targets as mutual occlusion occurs, while the *ITS* can successfully track the

[1]Available: http://imagelab.ing.unimore.it/visor/video_categories.asp
[2]Available: http://www.cvg.rdg.ac.uk/PETS2006/data.html
[3]Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.
[4]Available: http://www.cvg.rdg.ac.uk/PETS2009/a.html

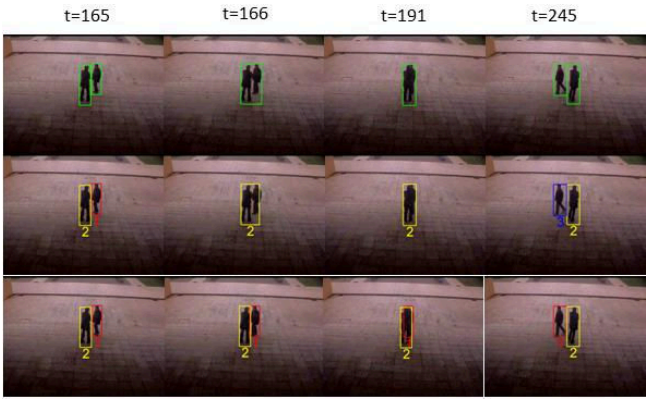

Fig. 3. Tracking results of the 'ViSOR'. First row: detection results. Second row: tracking with the *STS*. Third row: tracking with the *ITS*.

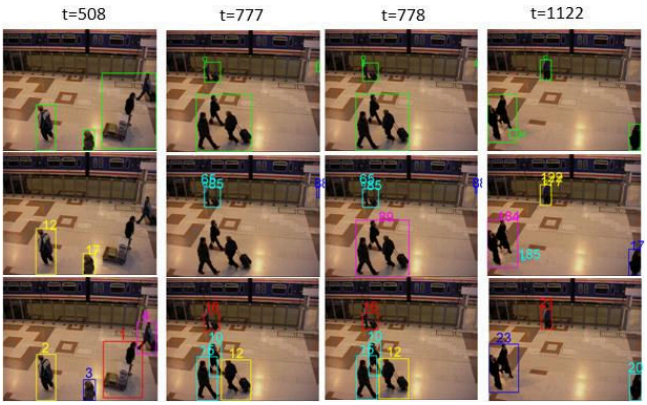

Fig. 4. Tracking results of the 'PETS2006'. First row: detection results. Second row: tracking with the *STS*. Third row: tracking with the *ITS*.

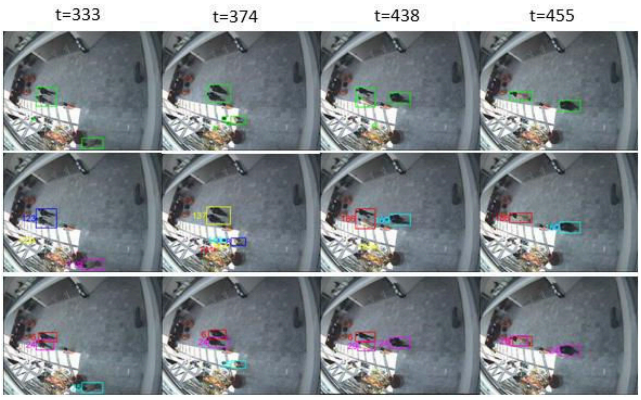

Fig. 5. Tracking results of the 'CAVIAR'. First row: detection results. Second row: tracking with the *STS*. Third row: tracking with the *ITS*.

Fig. 6. Tracking results of the 'PETS2009'. First row: detection results. Second row: tracking with the *STS*. Third row: tracking with the *ITS*.

TABLE I
TRACKING PERFORMANCE COMPARISON BETWEEN THE *ITS* AND THE *STS*

Data set	System	MOTP (%)	MOTA (%)	FPR (%)	MR (%)	MMR (%)
ViSOR	*ITS*	85.46	99.36	0.13	0.38	0.13
	STS	67.92	89.53	0.13	10.08	0.26
PETS2006	*ITS*	62.92	86.16	6.43	7.12	0.29
	STS	42.86	34.4	49.21	14.65	1.74
CAVIAR	*ITS*	80.64	78.65	19.56	0.96	0.83
	STS	65.78	33.93	52.73	10.13	3.21
PETS2009	*ITS*	58.47	87.21	0.11	11.45	1.23
	STS	49.76	46.17	0.23	19.94	6.12

targets in occlusion. Particularly, as several occlusions simultaneously occur in different targets groups (shown as t=723 and t=728 in Fig. 6), the *ITS* still can robustly track the targets in each occlusion region.

B. Quantitative Analysis

The CLEAR MOT metrics [16] are used to evaluate the occlusion tracking performance. The metrics return a multi-object tracking precision (MOTP) score and a multi-object tracking accuracy (MOTA) score. The MOTA is composed of the miss rate (MR), the false positive rate (FPR), and the mismatch rate (MMR). We compare the *ITS* with the *STS* and the state-of-the-art tracking systems according to the CLEAR MOT metrics.

TABLE II
TRACKING PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART TRACKING SYSTEMS ON THE DATA SET 'PETS2006'

System	Our *ITS*	Joo and Chellappa [17]	Torabi and Bilodeau [18]	Zuikifley and Moran [19]
MOTP (%)	62.92	49.8	56.87	58.16
MOTA (%)	86.16	92.21	96.56	98.75

Comparison with the *STS*: As mutual occlusion occurs, the *STS* may lose the targets or track the merged measurement as one target. This results in a large MR (shown as in Table I). On the contrary, the *ITS* can robustly handle the mutual occlusion problem. The results in Table I show that the *ITS* outperforms the *STS* both in MOTP and MOTA.

Comparison with the state-of-the-art tracking systems: We also compare the *ITS* with the state-of-the-art results reported in [17-19] for the data set 'PETS2006' (shown as in Table II), and in [20-22] for the data set 'PETS2009' (shown as in Table III). The results in Table II show that the *ITS* achieves a better MOTP score while gets a lower MOTA score. The results in Table III show that the *ITS* outperforms the results reported by Breitenstein et al. [21] and Yang et al. [22] both in precision and accuracy. When compared with the results reported by Andriyenko et al. [20], the *ITS* achieves a better MOTP score while gets a lower MOTA score. The

TABLE III
TRACKING PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART TRACKING SYSTEMS ON THE DATA SET 'PETS2009'

System	Our *ITS*	Andriyenko et al. [20]	Breitenstein et al. [21]	Yang et al. [22]
MOTP (%)	58.47	56.4	56.3	53.8
MOTA (%)	87.21	89.3	79.7	75.9

reason for a lower MOTA score is that we only use a simple background subtraction method for object detection. This may generate a large amount of noises due to the variable environment, and finally many false positives. This could be improved by using a more robust object detection method.

C. Discussions

Although aforementioned experiments have validated the effectiveness of the proposed occlusion handling algorithm, some other issues need to be discussed furthermore.

1) **Tracking newborn group targets**: As targets firstly enter into the scene in group (e.g. target 23 at t=1122 in Fig. 4 and target 29 at t=507 in Fig. 6), the occlusion handling algorithm cannot be invoked. In such cases, the targets are tracked as one newborn group target. To solve this problem, some more effective object detection methods should be incorporated to accurately detect the targets as they firstly appear in the scene.

2) **Processing speed**: The proposed tracking system is implemented in Matlab using a computer with Inter Core 2 Duo 2.20 GHz and 2 GB of memory. Without any code optimization the average runtimes for the above four data sets are about 0.4~1.2 frames per second. More than 95% of the runtimes are consumed in searching the Nash equilibrium of the game, because it is a pixel-wise iteration process. To remedy this, employing a more efficient appearance model will be helpful and will be explored in our future works.

V. CONCLUSION

We have developed a GM-PHD filter-based multi-target visual tracking system with the game theory-based mutual occlusion handling algorithm. We proposed a simple occlusion reasoning algorithm to correctly determine the occlusion region. We proposed a robust game theory-based mutual occlusion handling algorithm based on the proposed target appearance model to deal with the mutual occlusion problem. The proposed appearance model improved the conventional color histogram-based appearance model with the spatial constraint and other interacting targets' interferences, which was more robust as the targets in occlusion had similar appearances. We constructed an *n*-person, non-zero-sum, non-cooperative game and selected the Nash equilibrium of the game as the optimal estimation of the locations of the players within the occlusion region. Experiments conducted on publicly available videos showed that the proposed tracking system achieved promising results in handling mutual occlusions.

REFERENCES

[1] Y.D. Wang, J.K. Wu, W.M. Huang, and A.A. Kassim, "Gaussian mixture probability hypothesis density for visual people tracking," in *10th International Conference on Information Fusion*, 2007, pp. 1-6.

[2] E. Pollard, A. Plyer, B. Pannetier, F. Champagnat, and G. L. Besnerais, "GM-PHD filters for multi-object tracking in uncalibrated aerial videos," in *12th International Conference on Information Fusion*, 2009, pp. 1171-1178.

[3] J.J. Wu and S.Q. Hu, and Y. Wang, "Probability-hypothesis-density filter for multitarget visual tracking with trajectory recognition," *Optical Engineering*, vol. 49, no. 12, pp.12970-11-12970-19, December 2010.

[4] M. Yang, T. Yu and Y. Wu, "Game theory-based multiple target tracking", in *Proc. IEEE 11th Int. Conf. Computer Vision*, Rio de Janeiro, 2007, pp. 1-8.

[5] J. Xing, H. Ai, L. Liu and S. Lao, "Multiple player tracking in sports video: A dual-mode two-way Bayesian inference approach with progressive observation modeling", *IEEE Trans. Image Processing*, vol. 20, no. 6, pp. 1652-1667, June 2011.

[6] R. Vezzani, C. Grana and R. Cucchiara, "Probabilistic people tracking with appearance models and occlusion classification: The AD-HOC system", *Pattern Recognition Letters*, vol. 32, no. 6, pp. 867-877, April 2011.

[7] V. Papadourakis and A. Argyros, "Multiple objects tracking in the presence of long-term occlusions", *Computer Vision and Image Understanding*, vol. 114, no. 7, pp. 835-846, July 2010.

[8] W. Hu, X. Zhou, M. Hu and S. Manbank, "Occlusion reasoning for tracking multiple people", *IEEE Trans. Circuits and Systems for Video Technology*, vol.19, no. 1, pp.114-121, Jan. 2009.

[9] A. Senior, A. Hampapur, Y. L. Tian, L. Brown, S. Pankanti and R. Bolle, "Appearance models for occlusion handling", *Image Vision Computing*, vol. 24, no. 11, pp. 1233-1243, Nov. 2006.

[10] J. Nash, "Two-person cooperative games", *Econometrica*, vol. 21, no. 1, pp. 128-140, January 1953.

[11] E.N. Barron, *Game theory: an introduction*, Wiley, 2008.

[12] D. Gu, "A game theory approach to target tracking in sensor networks", *IEEE Trans. Systems, Man, and Cybernetics – Part B: Cybernetics*, vol. 41, no. 1, pp. 2-13, Feb. 2011.

[13] X. Zhou, Y.F. Li, B. He, T. Bai and Y. Tang, "Birth intensity online estimation in GM-PHD filter for multi-target visual tracking", in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura, 2012, pp. 3893-3898.

[14] B.-N. Vo and W.K. Ma, "The Gaussian mixture probability hypothesis density filter", *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4091-4104, Nov. 2006.

[15] X. Zhang, W. Hu, G. Luo and S. Manbank, "Kernel-bayesian framework for object tracking", in *Proc. 8th Asian Conf. Computer Vision*, Tokyo, 2007, pp. 821-831.

[16] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT Metrics", *EURASIP J. of Image and Video Processing*, vol. 2008, pp. 1-10, Feb. 2008.

[17] S. W. Joo and R. Chellappa, "A multiple-hypothesis approach for multiobject visual tracking," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2849-2854, November 2007.

[18] A. Torabi and G. A. Bilodeau, "A multiple hypothesis tracking method with fragmentation handling," In *Canadian conference on computer and robot vision,* 2009, pp. 8-15.

[19] M. A. Zulkifley and B. Moran, "Robust hierarchical multiple hypothesis tracker for multiple-object tracking," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12319-12331, November 2012.

[20] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1926-1933.

[21] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820-1833, September 2011.

[22] J. Yang, Z. Shi, P. Vela, and J. Teizer, "Probabilistic Multiple People Tracking through Complex Situations," in *IEEE Workshop Performance Evaluation of Tracking and Surveillance*, 2009, pp.79-86.