Game-theoretical occlusion handling for multi-target visual tracking

Xiaolong Zhou [a], Y.F. Li [a,*], Bingwei He [b]

[a] Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Kowloon, Hong Kong
[b] School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China

ABSTRACT

Multi-target visual tracking is a challenge because of data association and mutual occlusion in the interacting targets. This paper presents a Gaussian mixture probability hypothesis density based multi-target visual tracking system with game-theoretical occlusion handling. Firstly, the spatial constraint based appearance model with other interacting targets' interferences is modeled. Then, a two-step occlusion reasoning algorithm is proposed. Finally, an n-person, non-zero-sum, non-cooperative game is constructed to handle the mutual occlusion problem. The individual targets within the occlusion region are regarded as the players in the constructed game to compete for the maximum utilities by using the certain strategies. A Nash Equilibrium of the game is the optimal estimation of the locations of the players within the occlusion region. Experiments on video sequences demonstrate the good performance of the proposed occlusion handling algorithm.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Multi-target visual tracking (MTVT) has emerged as an active research topic in the past two decades due to its widespread applications in many areas, including intelligent surveillance, smart rooms, visual human–computer interfaces, autonomous robotics, augmented reality and video compression. It extends the single target visual tracking [1–5] to a situation where the number of targets is not known and is also varying with time. Moreover, the data association and the mutual occlusion problems make MTVT a challenge.

Recently, the random finite set (RFS) approach [6] to multi-target tracking in radar system has received considerable attention. Compared with the traditional association-based techniques, the difficulty caused by the data association is avoided in the RFS formulation. The probability hypothesis density (PHD) filter [7], which aims to propagate recursively the first order moment or the intensity function associated with the multi-target posterior density, provides a computationally tractable alternative. Gaussian mixture PHD (GM-PHD) filter [8], whose posterior intensity function is estimated by a sum of weighted Gaussian components that can be propagated analytically in time, is a closed form solution to the PHD filter recursion. Although the PHD filter is originated and widely used in radar tracking [7–9], it has been widely explored in visual tracking recently [10–15]. For example,

Maggio et al. [10] use the PHD filter to compensate for missing detections and to remove noises and clutters. Wang et al. [11] propose the data-driven importance function for a particle PHD filter and apply it to track pedestrians in the field of view of a camera. Vo et al. [12] demonstrate that the RFS approach enables a tractable solution to the multi-object estimation problem for image data. Pham et al. [14] incorporate the color feature representation into the GM-PHD filter to track multiple objects. In [15], a modified detection method is combined with the GM-PHD filter and an auction algorithm is proposed to automatically calculate the targets' trajectories.

The PHD filter based tracking system avoids the data association problem thus it is computationally appealing. However, it fails in tracking the individual targets when mutual occlusion occurs in the interacting targets. This paper focuses on proposing an effective algorithm to handle the mutual occlusion problem.

Extensive methods, i.e. multiple camera fusing or stereo vision based methods [16–19], Monte Carlo based probabilistic methods [20,21] and appearance model based deterministic methods [22–32], have been presented to solve the mutual occlusion problem. Though, the problem of tracking multiple interacting targets in occlusion is still far from being completely solved and remains a challenge. Compared with the other two classes of occlusion handling methods, tracking with the appearance model based deterministic methods offers several advantages in applications: generality, flexibility, computational efficiency, and large amount of information [26]. The appearance models proposed in the existing literature for occlusion handling include appearance template based [22,23], kernel density based [24,25] and color histogram based [26–32]. For example, Vezzani et al. [26] use an

* Corresponding author. Tel.: +852 3442 8410; fax: +852 2788 0172.
E-mail addresses: mexlz@hotmail.com (X. Zhou), meyfli@cityu.edu.hk (Y.F. Li),
mebwhe@fzu.edu.cn (B. He).

appearance driven tracking model to overcome large and long-lasting occlusions. They 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. Based on this target model, they classify the invisible regions into dynamic occlusions, scene occlusions and apparent occlusions. Xing et al. [27] build a dedicated observation model that maintains three discriminative cues including the appearance, size and motion. The target appearance is modeled as the color histogram in HSV color space in the discriminative region of the target. The mutual occlusion problem is then handled by a two-way Bayesian inference method. However, the above appearance models cannot handle the situation when interacting targets have similar color distributions. To remedy this, Papadourakis et al. [28] model the target by using an ellipse and a Gaussian mixture model (GMM). The ellipse accounts for the position and the 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. [29] model the human body as a vertical ellipse and use the spatial-color mixture of Gaussian appearance model [33] to model the spatial layout of the colors in a person in an image. The occlusion is deduced by using the current states of the interacting targets. However, they do not consider the mutual interferences among the interacting targets, which may affect the tracking precision as occlusion occurs.

To solve the above problems, this paper proposes a robust appearance model that takes both the spatial constraint of the target and the interferences of other interacting targets into consideration. A game-theoretical occlusion handling algorithm is then proposed to bridge the joint measurements estimation and the Nash Equilibrium of a game. Game theory is the study of multi-person decision making, which was first proposed by Nash [34]. 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 [35]. However, to the best of our knowledge, there are a few applications in visual tracking [25,36,37] and fewer in occlusion handling [25] based on the game theory. In this paper, we develop a GM-PHD based MTVT system with game-theoretical occlusion handling to track multiple moving targets in video, especially to track the interacting targets in mutual occlusion. The flow chart of the proposed system is shown in Fig. 1. The main contributions of this paper include those given here.

(1) The spatial constraint based appearance model with other interacting targets' interferences is modeled and is used for measure of the similarity between the candidate target and the target model. Compared with the conventional color
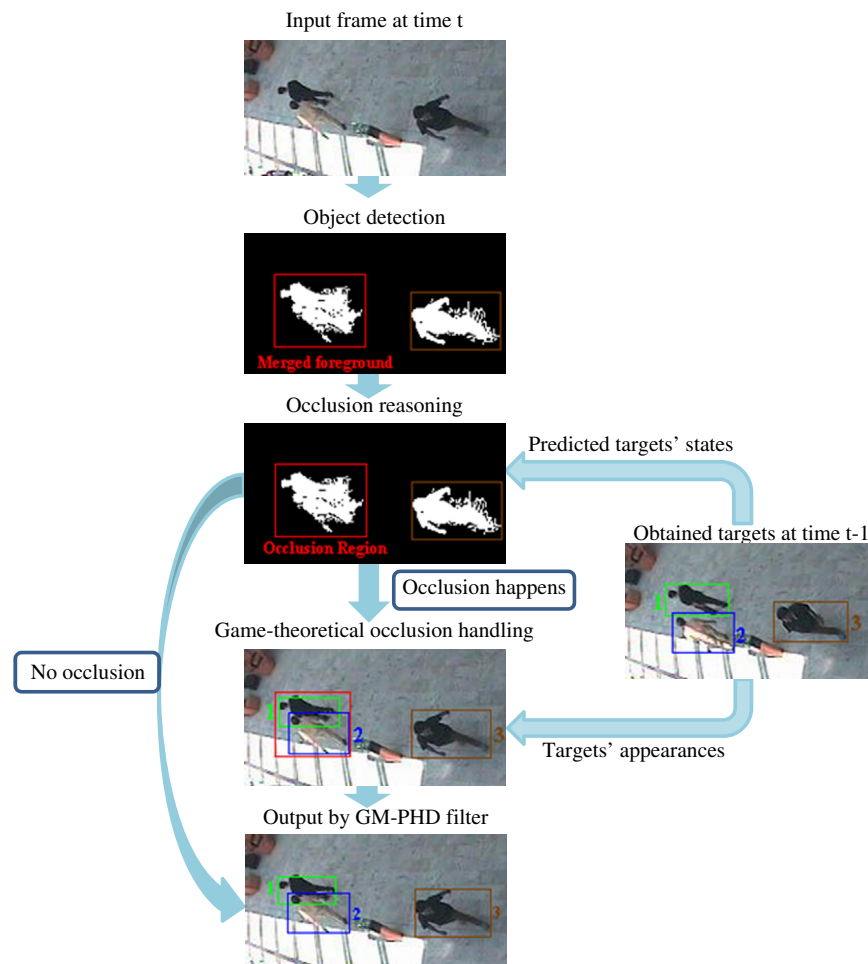

Fig. 1. Flow chart of the proposed MTVT system with game-theoretical occlusion handling.

histogram based appearance model, the proposed model is more robust when the interacting targets in mutual occlusion have similar color distributions.

(2) A two-step occlusion reasoning algorithm that includes occlusion prediction and occlusion determination is proposed. The candidate occlusion region is firstly predicted according to the overlapping between the support regions of the interacting targets. The final occlusion region is then determined by incorporating the detection results (measurements).

(3) A robust game-theoretical occlusion handling algorithm is proposed to obtain the individual measurements within the occlusion region. Different from the other conventional occlusion handling algorithms, an n-person, non-zero-sum, non-cooperative game is constructed to bridge the joint measurements estimation and the Nash Equilibrium. The individual targets are regarded as the players in the game to compete 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.

Supplementary material related to this article can be found online at doi:10.1016/j.patcog.2013.02.013.

The remainder of this paper is organized as follows. Section 2 describes the models of the targets and the measurements, Section 3 presents some backgrounds on the PHD filter and the GM-PHD filter, and Section 4 first introduces the occlusion reasoning algorithm, and then describes the game-theoretical occlusion handling algorithm in detail. Some experimental results on the real video data sets are discussed in Section 5, and followed by concluding remarks in Section 6.

2. Target modeling

2.1. Target state model and measurement model

The target region in an image is approximated with a $w \times h$ rectangle. Let the kinematic state of a target i at time t be $\mathbf{x}_t^i = \{\mathbf{l}_t^i, \mathbf{v}_t^i, \mathbf{s}_t^i\}$, where $\mathbf{l}_t^i = \{l_{x,t}^i, l_{y,t}^i\}$ is the image location of the target, $\mathbf{v}_t^i = \{v_{x,t}^i, v_{y,t}^i\}$ is the 2D velocity of the target and $\mathbf{s}_t^i = \{w_t^i, h_t^i\}$ is the size of the target's bounding box. $i = 1,...,N_t$ 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 $\mathbf{z}_t^j = \{\mathbf{l}_t^j, \mathbf{s}_t^j\}$, $j = 1,...,M_t$ and M_t is the number of measurements at time t. The targets' states set and the measurements set at time t can be denoted as $\mathbf{X}_t = \{\mathbf{x}_t^1, \cdots, \mathbf{x}_t^{N_t}\}$ and $\mathbf{Z}_t = \{\mathbf{z}_t^1, \cdots, \mathbf{z}_t^{M_t}\}$, respectively.

2.2. Target appearance model

Similar to [29], 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's pixels. $k = 1,...,K$ and $(\omega_k^i, \mu_k^i, \Sigma_k^i)$ represent the weight, the mean and the covariance matrix of the kth Gaussian component of the mixture. The measure of the similarity $P_s(p^i, q^i)$ between the candidate p^i and the model q^i is defined as the probability that the p^i's colors are drawn from q^i [30]

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

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 p^i (as shown in Fig. 2). $g_{\mathbf{l}^i} = G_{\mathbf{l}^i}/(R_{\mathbf{l}^i} + G_{\mathbf{l}^i} + B_{\mathbf{l}^i})$, $r_{\mathbf{l}^i} = R_{\mathbf{l}^i}/(R_{\mathbf{l}^i} + G_{\mathbf{l}^i} + B_{\mathbf{l}^i})$ and $I_{\mathbf{l}^i} = (R_{\mathbf{l}^i} + G_{\mathbf{l}^i} + B_{\mathbf{l}^i})/3$. N_i is the number of the foreground pixels in Ω_i.

$$N(x; \mu, \Sigma) = [2\pi|\Sigma|]^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}.$$
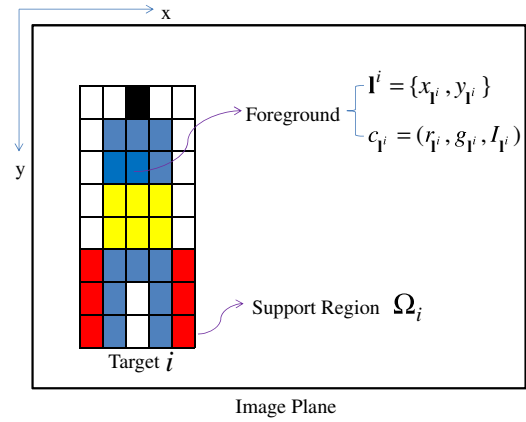

Fig. 2. Schematic diagram of the foreground and the support region of the target i.

However, the above appearance model is robust only when targets have different color distributions. It may fail when the interacting targets have similar appearances. In such case, a Gaussian spatial constraint is applied according to [38]. The measure of the similarity is then improved as

$$P_s(p^i, q^i) = \exp\left\{\frac{1}{N_i} \sum_{\Omega_i} \log\left\{N(\mathbf{l}^i; \mathbf{l}_t^i, \Sigma_t^i) \sum_{k=1}^K \omega_k^i N(c_{\mathbf{l}^i}; \mu_k^i, \Sigma_k^i)\right\}\right\} \quad (2)$$

where $\Sigma_t^i = [(w_t^i/2)^2, 0; 0, (h_t^i/2)^2]$.

Except for the similar appearances, the interferences by other interacting targets p^j within the occlusion region are also needed to be considered when occlusion occurs in the interacting targets. To take this point into account, the measure of the similarity is finally formulated as

$$P_s(p^i, q^i | p^j) = \exp\left\{\frac{1}{N_i}\left[\sum_{\Psi_1} \log(N(\mathbf{l}_t^i) \bullet \sigma_1) + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \log(N(\mathbf{l}_t^i) \bullet \sigma_1)\right]\right\} \quad (3)$$

where $N(\mathbf{l}_t^i) = N(\mathbf{l}^i; \mathbf{l}_t^i, \Sigma_t^i)$, $\sigma_1 = \sum_{k=1}^K \omega_k^i N(c_{\mathbf{l}^i}; \mu_k^i, \Sigma_k^i)$, $\sigma_2 = \sum_{j=1}^{N_o} \sum_{k=1}^K \omega_k^j N(c_{\mathbf{l}^i}; \mu_k^j, \Sigma_k^j)$, $\Psi_1 = \Omega_i - \Omega_i \cap \Omega_j$ and $\Psi_2 = \Omega_i \cap \Omega_j$. N_o is the number of interacting targets within the occlusion region. Note that if there is no occlusion or overlapping between the targets, Eq. (3) degenerates to Eq. (2).

3. Gaussian mixture probability hypothesis density filter

3.1. Probability hypothesis density filter

By definition [7], the PHD $D_t(\mathbf{x}_t)$, is the density whose integral on any region S of the state space is the expected number N_t of target contained in S.

$$N_t = \int_S D_t(\mathbf{x}_t) \, d\mathbf{x}_t \quad (4)$$

where $\mathbf{x}_t$ is the element of $\mathbf{X}_t$. The approximate expected target states are given by the local maxima of the PHD. The PHD filter propagates the intensity functions of multi-target RFS recursively. In general, there are two steps of each cycle of the PHD filter: prediction and update.

(1) *Prediction:* Suppose that we have known the PHD $D_{t-1}(\mathbf{x}_{t-1})$ at time $t-1$, and then the predicted PHD is given by

$$D_{t|t-1}(\mathbf{x}_t) = \gamma_t(\mathbf{x}_t) + \int [p_{sv,t}(\mathbf{x}_{t-1}) f_{t|t-1}(\mathbf{x}_t | \mathbf{x}_{t-1})$$
$$+ p_{sp,t}(\mathbf{x}_t)] D_{t-1}(\mathbf{x}_{t-1}) \, d\mathbf{x}_{t-1} \quad (5)$$

where $f_{t|t-1}(\mathbf{x}_t|\mathbf{x}_{t-1})$ denotes the single-target Markov transition density. $p_{sv,t}(\mathbf{x}_{t-1})$, $p_{sp,t}(\mathbf{x}_t)$ and $\gamma_t(\mathbf{x}_t)$ denote the probabilities of survived targets, spawned targets and new-born targets, respectively.

(2) *Update:* The predicted PHD can be corrected with the measurements $\mathbf{Z}_t$ at time step t to get the updated PHD. We assume the number of clutters is Poisson distributed with the average rate of λ_t, and the probability density of the spatial distribution of clutters is $c_t(\mathbf{z}_t)$. $\mathbf{z}_t$ is the element of $\mathbf{Z}_t$. Let the detection probability of a target with the state $\mathbf{x}_t$ be $p_{d,t}(\mathbf{x}_t)$. Then, the updated PHD is given by

$$D_t(\mathbf{x}_t) = [1-p_{d,t}(\mathbf{x}_t)]D_{t|t-1}(\mathbf{x}_t)$$
$$+ \sum_{\mathbf{z}_t \in \mathbf{Z}_t} \frac{p_{d,t}(\mathbf{x}_t)g_t(\mathbf{z}_t|\mathbf{x}_t)D_{t|t-1}(\mathbf{x}_t)}{\lambda_t c_t(\mathbf{z}_t) + \int p_{d,t}(\mathbf{x}_t)g_t(\mathbf{z}_t|\mathbf{x}_t)D_{t|t-1}(\mathbf{x}_t)d\mathbf{x}_t} \quad (6)$$

Where $g_t(\mathbf{z}_t|\mathbf{x}_t)$ denotes the single-target likelihood.

3.2. Gaussian mixture PHD filter

To implement the GM-PHD filter, several assumptions have to be made first

(1) *Each target follows a linear dynamic model where the process and observation noises are Gaussian*

$$f_{t|t-1}(\mathbf{x}_t|\mathbf{x}_{t-1}) = N(\mathbf{x}_t; \mathbf{F}_t\mathbf{x}_{t-1}, \mathbf{Q}_t) \quad (7)$$

$$g_t(\mathbf{z}_t|\mathbf{x}_t) = N(\mathbf{z}_t; \mathbf{H}_t\mathbf{x}_t, \mathbf{R}_t) \quad (8)$$

Where $N(\cdot; \mathbf{m}, \mathbf{P})$ denotes a Gaussian component with the mean $\mathbf{m}$ and the covariance $\mathbf{P}$. $\mathbf{F}_t$ 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 noise and the measurement noise, respectively.

(2) *The survival and detection probabilities are independent of the target state.* That is $p_{sv,t}(\mathbf{x}_{t-1}) = p_{sv}$ and $p_{d,t}(\mathbf{x}_t) = p_d$.

(3) *The birth intensity $\gamma_t(\mathbf{x}_t)$, can be represented by the Gaussian mixtures.* That is

$$\gamma_t(\mathbf{x}_t) = \sum_{i=1}^{J_{\gamma,t}} \omega_{\gamma,t}^{(i)} N(\mathbf{x}_t; \mathbf{m}_{\gamma,t}^{(i)}, \mathbf{P}_{\gamma,t}^{(i)}) \quad (9)$$

where $J_{\gamma,t}$, $\omega_{\gamma,t}^{(i)}$, and $\mathbf{P}_{\gamma,t}^{(i)}$ are the given parameters of the Gaussian mixture density of the newborn target state.

According to [8], the GM-PHD filter is implemented by following steps.

(1) *Prediction:* Suppose the prior density $D_{t-1}(\mathbf{x}_{t-1})$ has the form

$$D_{t-1}(\mathbf{x}_{t-1}) = \sum_{i=1}^{J_{t-1}} \omega_{t-1}^{(i)} N(\mathbf{x}_{t-1}; \mathbf{m}_{t-1}^{(i)}, \mathbf{P}_{t-1}^{(i)}) \quad (10)$$

Then the predicted intensity $D_{t|t-1}(\mathbf{x}_t)$ is given by

$$D_{t|t-1}(\mathbf{x}_t) = \gamma_t(\mathbf{x}_t) + p_{sv}\sum_{i=1}^{J_{t-1}} \omega_{t-1}^{(i)} N(\mathbf{x}_t; \mathbf{m}_{sv,t|t-1}^{(i)}, \mathbf{P}_{sv,t|t-1}^{(i)}) \quad (11)$$

where $\mathbf{m}_{sv,t|t-1}^{(i)} = \mathbf{F}_{t-1}\mathbf{m}_{t-1}^{(i)}$, $\mathbf{P}_{sv,t|t-1}^{(i)} = \mathbf{Q}_t + \mathbf{F}_{t-1}\mathbf{P}_{t-1}^{(i)}\mathbf{F}_{t-1}^T$. The spawned targets in the prediction of the PHD filter (Eq. (5)) usually come from the needs of military applications in radar tracking, e.g., an airplane sends a missile [11]. However, there is no such situation in our visual tracking scenario. For simplicity, we assume that all targets consist of the survival targets and the newborn targets.

(2) *Update:* The predicted intensity $D_{t|t-1}(\mathbf{x}_t)$ can be expressed as a Gaussian mixture of the form

$$D_{t|t-1}(\mathbf{x}_t) = \sum_{i=1}^{J_{t|t-1}} \omega_{t|t-1}^{(i)} N(\mathbf{x}_t; \mathbf{m}_{t|t-1}^{(i)}, \mathbf{P}_{t|t-1}^{(i)}) \quad (12)$$

then the posterior intensity $D_t(\mathbf{x}_t)$ is given by

$$D_t(\mathbf{x}_t) = (1-p_d)D_{t|t-1}(\mathbf{x}_t) + \sum_{\mathbf{z}_t \in \mathbf{Z}_t} D_{g,t}(\mathbf{x}_t; \mathbf{z}_t) \quad (13)$$

$$D_{g,t}(\mathbf{x}_t; \mathbf{z}_t) = \sum_{i=1}^{J_{t|t-1}} \omega_{g,t}^{(i)}(\mathbf{z}_t) N(\mathbf{x}_t; \mathbf{m}_{g,t}^{(i)}(\mathbf{z}_t), \mathbf{P}_{g,t}^{(i)}(\mathbf{z}_t)) \quad (14)$$

$$\omega_{g,t}^{(i)}(r_t, \mathbf{z}_t) = \frac{p_d \omega_{t|t-1}^{(i)} N(\mathbf{z}_t; \mathbf{m}_{h,t}^{(i)}, \mathbf{P}_{h,t}^{(i)})}{\lambda_t c_t(\mathbf{z}_t) + p_d \sum_{i=1}^{J_{t|t-1}} \omega_{t|t-1}^{(i)} N(\mathbf{z}_t; \mathbf{m}_{h,t}^{(i)}, \mathbf{P}_{h,t}^{(i)})} \quad (15)$$

where $\mathbf{m}_{g,t}^{(i)}(\mathbf{z}_t) = \mathbf{m}_{t|t-1}^{(i)} + K(\mathbf{z}_t - \mathbf{H}_t\mathbf{m}_{t|t-1}^{(i)})$, $\mathbf{P}_{g,t}^{(i)}(\mathbf{z}_t) = (\mathbf{I} - K\mathbf{H}_t)\mathbf{P}_{t|t-1}^{(i)}$, $\mathbf{m}_{h,t}^{(i)} = \mathbf{H}_t\mathbf{m}_{t|t-1}^{(i)}$, $\mathbf{P}_{h,t}^{(i)} = \mathbf{R}_t + \mathbf{H}_t\mathbf{P}_{t|t-1}^{(i)}\mathbf{H}_t^T$, $K = \mathbf{P}_{t|t-1}^{(i)}\mathbf{H}_t^T(\mathbf{H}_t\mathbf{P}_{t|t-1}^{(i)}\mathbf{H}_t^T + \mathbf{R}_t)^{-1}$.

(3) *Pruning, merging and target states extraction:* From the above two steps, it can be shown the number of components of the predicted and the posterior intensities increases with time, which can be a problem in implementation. Therefore, some pruning and merging algorithms [8] should be applied to prune those components that are irrelevant with the target intensity and to merge components that share the same intensity peak into one component. Given the posterior intensity $D_t(\mathbf{x}_t)$ at time t, it can be represented as $D_t(\mathbf{x}_t) = \{\omega_t^{(i)}, \mathbf{m}_t^{(i)}, \mathbf{P}_t^{(i)}\}_{i=1}^{J_t}$. Given a pruning threshold τ_p and a merging threshold τ_m, the details of the pruning and merging algorithms are shown in Fig. 3.

After the pruning and merging, the peaks of the intensity $\hat{D}_t(\mathbf{x}_t)$ are points of the highest local concentration of the expected number N_t of targets. The estimate of the multi-target state is the set of N_t ordered of the mean with the largest weights.

4. Game-theoretical occlusion handling

4.1. Occlusion reasoning

A simple two-step occlusion reasoning algorithm that includes occlusion prediction and occlusion determination is proposed. The occlusion region is deduced according to the predicted targets' states and the detection results (measurements).

(1) Occlusion prediction
As shown in Fig. 4(a), C_i (or C_j) is a circle centered at $\mathbf{l}_{t|t-1}^i$ (or $\mathbf{l}_{t|t-1}^j$) with the 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 the scale 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 determined only as $C_i \cap C_j \neq \emptyset$ ($i \neq j$). That is

$$\|\mathbf{l}_{t|t-1}^i - \mathbf{l}_{t|t-1}^j\| < \|\mathbf{s}_{t|t-1}^i\| + \|\mathbf{s}_{t|t-1}^j\| \quad (16)$$

Otherwise, no occlusion occurs.

(2) Occlusion determination
There are two possible situations, no occlusion and occlusion, exist in the candidate occlusion region (shown as Fig. 4(b) and (c)). According to the fact that the overlapping between the occlusion targets always increases gradually, the first detected merged foreground of the occlusion region is always larger than the corresponding single target. To further determine the occlusion region, the detection results

1. Input $D_t(\mathbf{x}_t) = \left\{\omega_t^{(i)}, \mathbf{m}_t^{(i)}, \mathbf{P}_t^{(i)}\right\}_{i=1}^{J_t}$.

2. Pruning: $I = \left\{i : 1 \leq i \leq J_{t,p}, \omega_t^{(i)} > \tau_p\right\}$.

3. Merging: Set $l = 0$, repeat: $l = l+1$, $j = \arg\max_{i \in I} \omega_t^{(i)}$,

$$L = \left\{i \in I \left| \left(\mathbf{m}_t^{(i)} - \mathbf{m}_t^{(j)}\right)^T \left(\mathbf{P}_t^{(i)}\right)^{-1} \left(\mathbf{m}_t^{(i)} - \mathbf{m}_t^{(j)}\right) \leq \tau_m \right.\right\},$$

$$\hat{\omega}_t^{(l)} = \sum_{i \in L} \omega_t^{(i)}, \quad \hat{\mathbf{m}}_t^{(l)} = \frac{1}{\hat{\omega}_t^{(l)}} \sum_{i \in L} \omega_t^{(i)} \mathbf{m}_t^{(i)}, \quad \hat{\mathbf{P}}_t^{(l)} = \frac{1}{\hat{\omega}_t^{(l)}} \sum_{i \in L} \omega_t^{(i)} \left(\mathbf{P}_t^{(i)} + \left(\hat{\mathbf{m}}_t^{(l)} - \mathbf{m}_t^{(i)}\right)\left(\hat{\mathbf{m}}_t^{(l)} - \mathbf{m}_t^{(i)}\right)^T\right),$$

$I = I \setminus L$, until $I = \varnothing$.

4. Output $\hat{D}_t(\mathbf{x}_t) = \left\{\hat{\omega}_t^{(i)}, \hat{\mathbf{m}}_t^{(i)}, \hat{\mathbf{P}}_t^{(i)}\right\}_{i=1}^{l}$.

Fig. 3. Procedure of the pruning and merging algorithm.
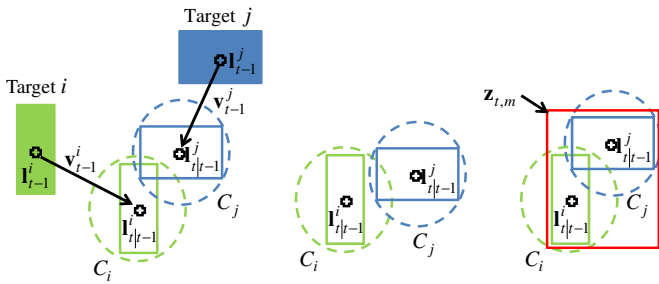

Fig. 4. Illustration of occlusion prediction and determination. (a) Occlusion prediction, (b) no occlusion and (c) occlusion occurs and occlusion region is determined as $\mathbf{z}_{t,m}$.

(measurements) obtained at the current time step is incorporated. If a measurement $\mathbf{z}_{t,m} = \{\mathbf{l}_{t,m}, \mathbf{s}_{t,m}\}$ within the candidate occlusion region satisfies the following condition, the measurement is regarded as an occlusion region.

$$\|\mathbf{s}_{t,m}\| > \varepsilon \cdot \max(\|\mathbf{s}_{t|t-1}^i\|, \|\mathbf{s}_{t|t-1}^j\|) \tag{17}$$

where ε is a scale factor. The size of the detected target may be slightly changed between consecutive frames due to the changes of the target's pose or of the depth of view. On the contrary, compared with the size of the target before mutual occlusion, the size of the target after mutual occlusion is significantly changed because it is merged with other targets. Consequently, we set $\varepsilon = 1.2$ in our experiments to determine the occlusion region more correctly.

4.2. Occlusion handling

As occlusion occurs, the measurement $\mathbf{z}_{t,m}$ that contains the merged foreground $\mathbf{F}_{t,m}$ is determined by occlusion reasoning. It is assumed that the target's size keeps constant during the occlusion. Once the occlusion region is determined, the identities (IDs) and the total number N_o of the targets in the occlusion region are determined. Then the appearance models of the corresponding
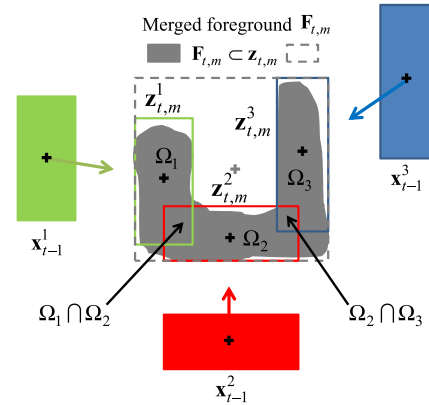

Fig. 5. Schematic diagram of occlusion handling. Three targets compete for the merged foreground to reach a Nash Equilibrium.

targets obtained at time $t-1$ are constructed. As shown in Fig. 5, given the merged foreground $\mathbf{F}_{t,m}$ and the predicted targets' states $\{\mathbf{x}_{t|t-1}^i\}_{i=1}^{N_o}$ ($N_o = 3$ in Fig. 5), the goal is to obtain the optimal individual measurements $\{\mathbf{z}_{t,m}^i\}_{i=1}^{N_o}$ within the occlusion region. In other words, the optimal solution $(\mathbf{z}_{t,m}^{1*}, \ldots, \mathbf{z}_{t,m}^{N_o*})$ is to maximize the similarity probability between the joint measurements $(\mathbf{z}_{t,m}^1, \ldots, \mathbf{z}_{t,m}^{N_o})$ and the foreground $\mathbf{F}_{t,m}$.

$$(\mathbf{z}_{t,m}^{1*}, \ldots, \mathbf{z}_{t,m}^{N_o*}) = \arg\max_{\{\mathbf{z}_{t,m}^i\}_{i=1}^N} P(\mathbf{z}_{t,m}^1, \ldots, \mathbf{z}_{t,m}^{N_o} | \mathbf{F}_{t,m}) \tag{18}$$

According to the Bayesian theory, we have

$$\arg\max_{\{\mathbf{z}_{t,m}^i\}_{i=1}^N} P(\mathbf{z}_{t,m}^1, \ldots, \mathbf{z}_{t,m}^{N_o} | \mathbf{F}_{t,m}) = \arg\max_{\{\mathbf{z}_{t,m}^i\}_{i=1}^N} P(\mathbf{z}_{t,m}^i, \mathbf{z}_{t,m}^{-i} | \mathbf{F}_{t,m})$$

$$= \arg\max_{\{\mathbf{z}_{t,m}^i\}_{i=1}^N} P(\mathbf{z}_{t,m}^i | \mathbf{F}_{t,m}, \mathbf{z}_{t,m}^{-i}) P(\mathbf{z}_{t,m}^{-i} | \mathbf{F}_{t,m}) \tag{19}$$

where $\mathbf{z}_{t,m}^{-i} = \{\mathbf{z}_{t,m}^j\}_{j=1,j\neq i}^{N_o}$. To obtain the optimal solution of Eq. (19), a game-theoretical occlusion handling algorithm is proposed to bridge the joint measurements estimation and the Nash Equilibrium of a game.

4.3. N-person, non-zero-sum, non-cooperative game constructing

In game theory [34], a non-cooperative game is one in which players make decisions independently. As mutual occlusion occurs, the individual measurements involved in the occlusion region are competing to maximize the similarity probability between the measurements and the foreground independently. 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. Since the size of a target is assumed constant, the estimation of the measurements $(\mathbf{z}_{t,m}^1, \cdots, \mathbf{z}_{t,m}^{N_o})$ is simplified as the estimation of the locations $(\mathbf{l}_t^1, \cdots, \mathbf{l}_t^{N_o})$ of the measurements. The player, strategy of the player and the corresponding utility are defined as follows:

Player: The individual measurement $\mathbf{z}_{t,m}^i$ corresponding to the target $i \in \{1, \cdots, N_o\}$ in the occlusion region.
Strategy: The motion of the player, i.e. the location $\mathbf{l}_t^i$ of the player in the image. $\mathbf{l}_t^i = \{l_{x,t}^i, l_{y,t}^i\} \in R^2$.
Utility function: $U^i(\mathbf{l}_t^i, \mathbf{l}_t^{-i}) = P(\mathbf{z}_{t,m}^i | \mathbf{F}_{t,m}, \mathbf{z}_{t,m}^{-i}) P(\mathbf{z}_{t,m}^{-i} | \mathbf{F}_{t,m})$, where $\mathbf{l}_t^{-i} = \{\mathbf{l}_t^j\}_{j=1, j \neq i}^{N_o}$.

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

Definition 1. (Barron [35]): The *best-response* of a player i to the profile of strategies $\mathbf{l}_t^{-i}$ is the strategy for that player such that

$$U^i(\mathbf{l}_t^i, \mathbf{l}_t^{-i}) \geq U^i(\mathbf{l}_t^{i\prime}, \mathbf{l}_t^{-i}), \forall \mathbf{l}_t^{i\prime} \in R^2 \tag{20}$$

Hence, a Nash Equilibrium of the game is a strategy profile for which every player's strategy is a *best-response* to the other players' strategies.

Definition 2. (Barron [35]): $(\mathbf{l}_t^{1*}, \cdots, \mathbf{l}_t^{N_o*})$ is a Nash Equilibrium for the game with utility $\{U^i(\mathbf{l}_t^i, \mathbf{l}_t^{-i})\}_{i=1, \cdots, N_o}$, if every player's strategy is a *best-response* to the other players' strategies

$$U^i(\mathbf{l}_t^{i*}, \mathbf{l}_t^{-i*}) \geq U^i(\mathbf{l}_t^i, \mathbf{l}_t^{-i*}), \text{ for every player } i \tag{21}$$

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

$$\max U^i(\mathbf{l}_t^i, \mathbf{l}_t^{-i*}) = \max P(\mathbf{z}_{t,m}^i | \mathbf{F}_{t,m}, \mathbf{z}_{t,m}^{-i*}) P(\mathbf{z}_{t,m}^{-i*} | \mathbf{F}_{t,m}) \propto \max P(\mathbf{z}_{t,m}^i | \mathbf{F}_{t,m}, \mathbf{z}_{t,m}^{-i*}) \tag{22}$$

where $P(\mathbf{z}_{t,m}^i | \mathbf{F}_{t,m}, \mathbf{z}_{t,m}^{-i*})$ is the similar probability between the measurement $\mathbf{z}_{t,m}^i$ and the merged foreground $\mathbf{F}_{t,m}$ with the interferences of other fixed measurements $\mathbf{z}_{t,m}^{-i*}$.

$$\max P(\mathbf{z}_{t,m}^i | \mathbf{F}_{t,m}, \mathbf{z}_{t,m}^{-i*}) \propto \max P_s(p^i, q^i | p^j) \propto \max f(\mathbf{l}_t^i) \tag{23}$$

where

$$f(\mathbf{l}_t^i) = \sum_{\Psi_1} \log\left(N(\mathbf{l}_t^i) \cdot \sigma_1\right) + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \log\left(N(\mathbf{l}_t^i) \cdot \sigma_1\right) \tag{24}$$

Then we set the derivative of $f(\mathbf{l}_t^i)$ with respect to $\mathbf{l}_t^i$ to zero

$$\frac{\partial f(\mathbf{l}_t^i)}{\partial \mathbf{l}_t^i} = \sum_{\Psi_1} \frac{1}{N(\mathbf{l}_t^i)} \cdot \frac{\partial N(\mathbf{l}_t^i)}{\partial \mathbf{l}_t^i} + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \cdot \frac{1}{N(\mathbf{l}_t^i)} \cdot \frac{\partial N(\mathbf{l}_t^i)}{\partial \mathbf{l}_t^i}$$
$$= \sum_{\Psi_1} (\Sigma_t^i)^{-1} \cdot (\mathbf{l}^i - \mathbf{l}_t^i) + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \cdot (\Sigma_t^i)^{-1} \cdot (\mathbf{l}^i - \mathbf{l}_t^i) = 0 \tag{25}$$

$$\sum_{\Psi_1} \mathbf{l}^i - N_i' \cdot \mathbf{l}_t^i + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \cdot \mathbf{l}^i - \mathbf{l}_t^i \cdot \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} = 0 \tag{26}$$

where N_i' is the number of foreground pixels in the support region Ψ_1. $\mathbf{l}_t^i$ can be calculated by Eq. (26) and is regarded as the *best-response* $\mathbf{l}_t^{i*}$ of the player i

$$\mathbf{l}_t^{i*} = \left(\sum_{\Psi_1} \mathbf{l}^i + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2} \cdot \mathbf{l}^i\right) \bigg/ \left(N_i' + \sum_{\Psi_2} \frac{\sigma_1}{\sigma_2}\right) \tag{27}$$

The location of the player i is initialized by the corresponding predicted target's location $\mathbf{l}_{t|t-1}^i$, $i \in \{1, \cdots, N_o\}$. Given the initialized locations $\mathbf{l}_t^{-i*}$, the *best-response* $\mathbf{l}_t^{i*}$ of the player i can be calculated by Eq. (27). $\mathbf{l}_t^{i*}$ can be 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. (28). $\Delta\mathbf{l}$ is the difference of the *best-response* sets between the consecutive iteration cycles.

$$\Delta\mathbf{l} < T_{NE} \tag{28}$$

where $\Delta\mathbf{l} = \left| \{\mathbf{l}_t^{1*}, \cdots, \mathbf{l}_t^{N_o*}\}_{\text{iteration } j} - \{\mathbf{l}_t^{1*}, \cdots, \mathbf{l}_t^{N_o*}\}_{\text{iteration } j-1} \right|$ ($j=1,2, \ldots$). $\{\mathbf{l}_t^{1*}, \cdots, \mathbf{l}_t^{N_o*}\}_{\text{iteration } 0}$ is the initialized locations set. T_{NE} is the given threshold. The smaller the 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.

4.4. Algorithm summary

The procedure of the proposed multi-target visual tracking system and the corresponding occlusion reasoning and game-theoretical occlusion handling algorithms are shown in Figs. 6–8.

5. Experiments and discussions

The proposed visual tracking system is tested on the real-world scenarios. In particular, the contribution of the proposed game-theoretical occlusion handling algorithm is assessed.

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 and $\sigma_v = 3$ 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 set as the detection probability $p_d = 0.99$, the survival probability $p_{sv} = 0.95$, the average distributed rate of clutters $\lambda_t = 0.01$, the spatial distribution of clutters $c_t(\mathbf{z}_t) = (\text{image area})^{-1}$, the pruning threshold $\tau_p = 0.1$ and the merging threshold $\tau_m = 5$.

Any object detection method can be incorporated into our tracking system. Because of the contribution of this paper mainly focuses on effectively handling the mutual occlusion problem, a simple background subtraction algorithm based on RGB color for the object detection is utilized. It is assumed that the static background image has already known. First, each pixel in the background image is modeled as R, G and B channels. Then the difference between the current image and the background image for each channel is calculated, and the pixel is labeled as the foreground if the difference of one channel is larger than the threshold τ ($\tau = 20$ in our experiments). Finally, the morphological operator is employed to eliminate the isolated noises and the 8-connected components labeling algorithm is used to connect the detected foreground pixels to a set of regions. The regions are enclosed by a set of rectangles.

The proposed tracking system is evaluated on the video data sets coming from 'ViSOR' [39], 'CAVIAR' [40] and 'PETS2006' [41]. All the videos are captured from the single static camera. The

1. Input the frame F_t at time t and the targets' states $\mathbf{X}_{t-1} = \{\mathbf{x}_{t-1}^i\}_{i=1}^{N_{t-1}}$ at time $t-1$.

2. Predict the targets' states $\mathbf{X}_{t|t-1}$ by the GM-PHD filter.

    Obtain the measurement $\mathbf{Z}_t$ at time t by the background subtraction method.

3. Call the "**Occlusion reasoning algorithm**" in Fig. 7 to determine the occlusion regions.

    If occlusion occurs, go to step 4. Otherwise, go to step 5.

4. Construct the appearance models $\{q^i\}_{i=1}^{N_o}$ of those targets involved in the occlusion region.

    Call the "**Game-theoretical occlusion handling algorithm**" in Fig. 8 to obtain the locations of the

    individual measurements.

5. Update the targets' states by the GM-PHD filter.

    Output the targets' states $\mathbf{X}_t = \{\mathbf{x}_t^i\}_{i=1}^{N_t}$ and targets' IDs.

Fig. 6. Procedure of the proposed multi-target visual tracking system.

1. Input the predicted targets' states $\mathbf{X}_{t|t-1}$ and the measurements $\mathbf{Z}_t$.

2. **Occlusion prediction**:

    Generate a circle C_i of predicted target i with center $\mathbf{l}_{t|t-1}^i$ (or $\mathbf{l}_{t|t-1}^j$) and radius $\left\|\mathbf{s}_{t|t-1}^i\right\|$ (or $\left\|\mathbf{s}_{t|t-1}^j\right\|$).

    If $C_i \bigcap C_j \neq \varnothing$ ($i \neq j$), C_i and C_j are regarded as the candidate occlusion regions, go to step 3.

    Else, no occlusion occurs, return.

3. **Occlusion determination**:

    Within the candidate occlusion regions, if a measurement $\mathbf{z}_{t,m} = \{\mathbf{l}_{t,m}, \mathbf{s}_{t,m}\} \in \mathbf{Z}_t$ satisfies the Equation

    (17), occlusion occurs. Then output the occlusion region (measurement) $\mathbf{z}_{t,m}$.

    Else, no occlusion occurs, return.

Fig. 7. Occlusion reasoning algorithm.

following experiments are conducted to validate the effectiveness of the proposed occlusion handling algorithm during the occlusion period. The challenging issues involved in the occlusion period for each data set are listed in Table 1. Moreover, the comparisons between the GM-PHD filter based *conventional tracking system without occlusion handling* (CTS-OH) and the *proposed tracking system with game-theoretical occlusion handling* (PTS+OH) are presented to show the performances in handling the listed challenging issues.

5.1. Qualitative analysis

Figs. 9–14 show the detection results (the first row in Figs. 9–14), the tracking results of the CTS-OH (the second row in Figs. 9–14) and of the PTS+OH (the third row in Figs. 9–14).

As the targets come near, occlusion occurs and the targets are detected as a merged foreground. Without occlusion handling, the system may lose the targets or track the merged foreground as the target. Conventionally, simple color appearance model based occlusion handling algorithm can deal well with the targets with different color distributions. However, it fails when the interacting targets have similar appearances. On the contrary, the proposed appearance model considers both the similar appearances and the interferences of other interacting targets. Moreover, the game-theoretical occlusion handling algorithm provides the optimal solutions of multiple targets' locations within the occlusion region. Therefore, by incorporating the proposed appearance model into the game-theoretical occlusion handling algorithm, our tracking system can perform robustly as mutual occlusion occurs in the interacting targets.

1. Input the measurement $\mathbf{z}_{t,m}$, the foreground $\mathbf{F}_{t,m}$, the targets' appearance models $\{q^i\}_{i=1}^{N_o}$, and the predicted locations $\{\mathbf{l}^i_{t|t-1}\}_{i=1}^{N_o}$ of the targets within the occlusion region. The location $\mathbf{l}^i_{t|t-1}$ is the initial location of the player i.

2. $j=1$, iteration cycle j

   While ($\Delta \mathbf{l} < T_{NE}$)

   {For $i=1:N_o$, given $\mathbf{l}_t^{-i*}$

   Calculate $\Psi_1 = \Omega_i - \Omega_i \bigcap \Omega_j$ and $\Psi_2 = \Omega_i \bigcap \Omega_j$.

   Calculate the number N_i' of foreground pixels in Ψ_1.

   Calculate $\sum_{\Psi_1} \mathbf{l}^i$, $\sum_{\Psi_2} \frac{\sigma_1}{\sigma_2}\square^i$ and $\sum_{\Psi_2} \frac{\sigma_1}{\sigma_2}$.

   Calculate $\mathbf{l}_t^{i*}$ by Equation (27).

   End}

   $j = j+1$;

   Calculate $\Delta \mathbf{l} = \left| \{\mathbf{l}_t^{1*}, \cdots, \mathbf{l}_t^{N_o*}\}_{iteration\ j} - \{\mathbf{l}_t^{1*}, \cdots, \mathbf{l}_t^{N_o*}\}_{iteration\ j-1} \right|$.

3. Output the individual measurement $\mathbf{z}_{t,m}^i = \{\mathbf{l}_t^{i*}, \mathbf{s}_{t-1}^i\}$, $i=1,\cdots,N_o$.

Fig. 8. Game-theoretical occlusion handling algorithm.

Table 1
Challenging issues involved in the occlusion period.

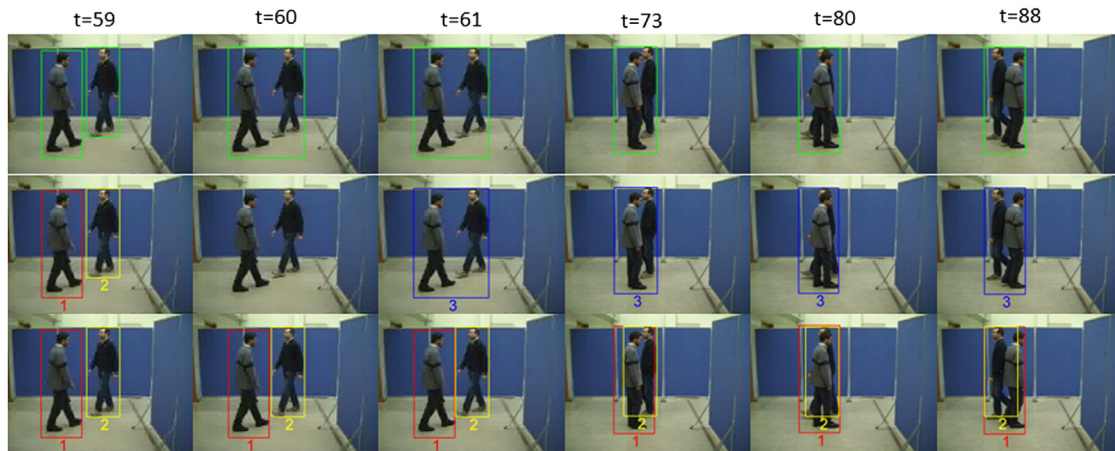| Data Sets	Challenging issues involved in the occlusion period
ViSOR #1	• Two interacting targets with different appearances • Partial and total occlusion • Targets merge once and split once
ViSOR #2	• Two interacting targets with similar appearances • Partial and total occlusion • Targets merge once and split once
ViSOR #3	• Three interacting targets with different appearances in rapid moving • Partial occlusion • Targets merge five times and split five times
CAVIAR	• Three interacting targets, in which two targets have similar appearances • Partial occlusion • Targets merge twice and split twice
PETS2006 #1	• Three interacting targets with similar appearances • Partial and total occlusion • Targets merge twice and split once
PETS2006 #2	• Four interacting targets, in which three targets have similar appearances • Partial occlusion • Targets merge three times and split zero times

Fig. 9. Tracking results comparison for 'ViSOR #1'. The first row: detection results; second row: tracking results of the CTS-OH; third row: tracking results of the PTS+OH. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Fig. 10. Tracking results comparison for 'ViSOR #2'. The first row: detection results; second row: tracking results of the CTS-OH; third row: tracking results of the PTS+OH. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)
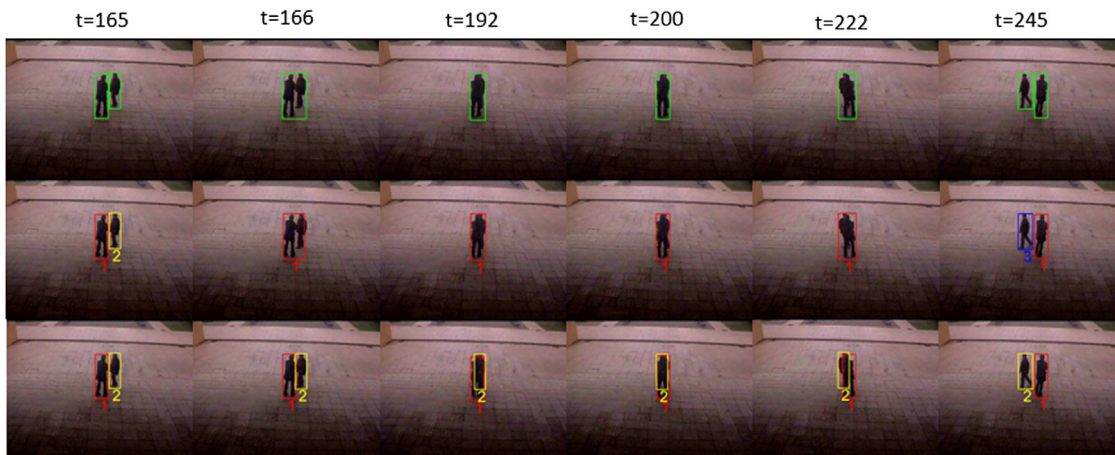

Fig. 11. Tracking results comparison for 'ViSOR #3'. The first row: detection results; second row: tracking results of the CTS-OH; third row: tracking results of the PTS+OH. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)
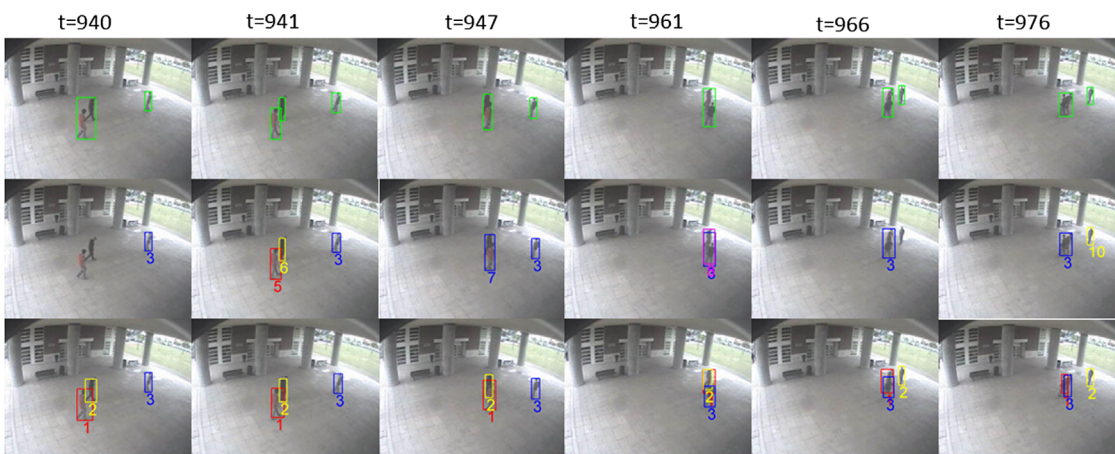
In Fig. 9 ('ViSOR #1'), the occlusion occurs from $t=60$ to $t=97$. When this occurs, the CTS-OH fails in tracking the two targets at $t=60$, and it then tracks the merged detection as a single newborn target at $t=61$. As described in Section 3.2, the estimate of the multi-target states is the set of targets ordered of the mean with the largest weights of the Gaussian components. Because of the large differences between the merged detection and the two predicted targets' states (targets 1 and 2) at $t=60$, the
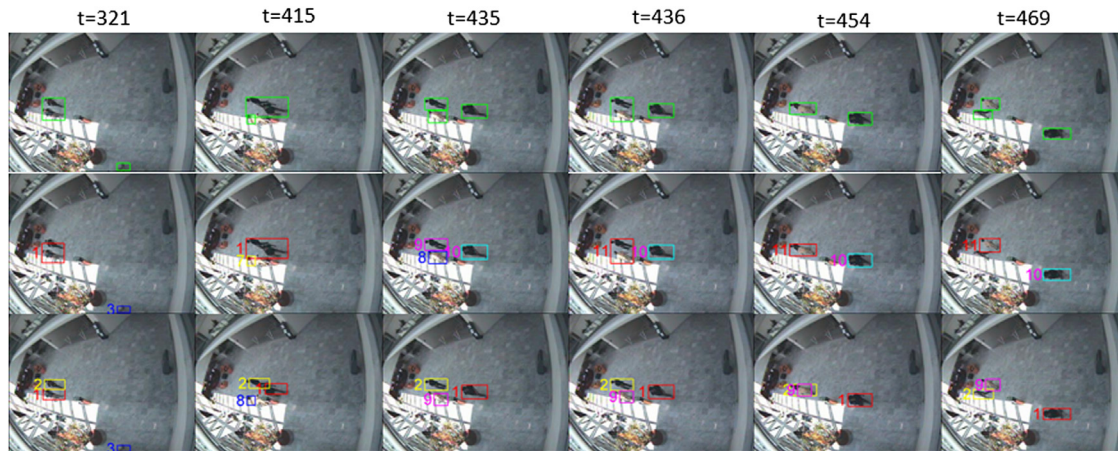
Fig. 12. Tracking results comparison for 'CAVIAR'. The first row: detection results; second row: tracking results of the CTS-OH; and third row: tracking results of the PTS+OH. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Fig. 13. Tracking results comparison for 'PETS2006 #1'. The first row: detection results; second row: tracking results of the CTS-OH; third row: tracking results of the PTS+OH. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)
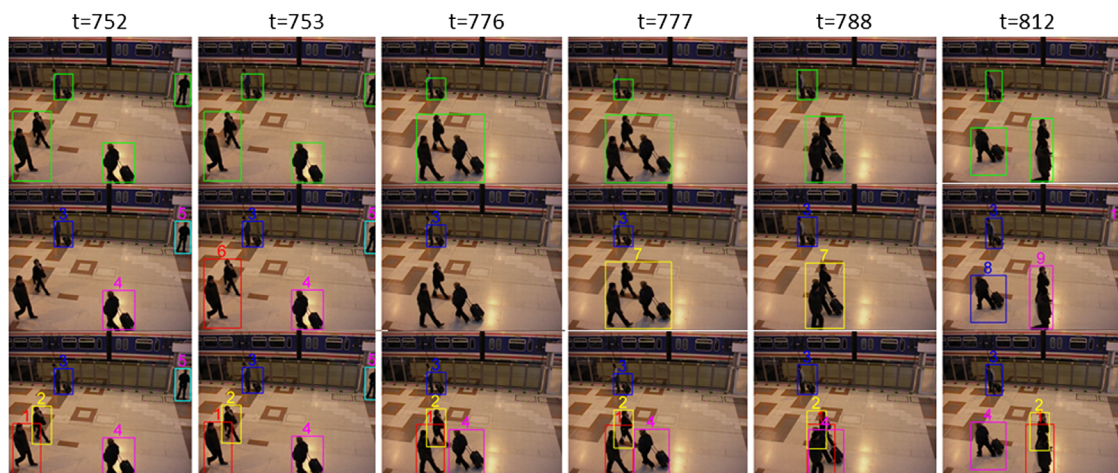
corresponding weights of the components are smaller than the given threshold. Consequently, the CTS-OH fails in tracking them as the occlusion occurs. However, the PTS+OH can succeed in tracking the targets during the whole occlusion period, even when target 2 is totally occluded by target 1 (shown as $t=80$ in Fig. 9).

In Fig. 10 ('ViSOR #2'), the occlusion occurs from $t=166$ to $t=241$. When this occurs, the difference between target 1 and the merged detection is smaller than the difference between target 2 and the merged detection. So the CTS-OH fails in tracking target 2 while tracks the merged detection as the other target 1 from $t=166$. Then, target 2 is re-tracked as a newborn target 3 from $t=245$. However, the PTS+OH can succeed in tracking the targets during the whole occlusion period, even though the targets have similar appearances.

In Fig. 11 ('ViSOR #3'), the occlusion among the three targets is complicated. Targets 1 and 2 merge first at $t=940$ and then they split and merge several times until they merge with target 3 at $t=961$. From $t=966$, target 2 splits out while targets 1 and 3 remain merge. Similarly, the CTS-OH loses the targets or tracks the merged detection as a target during the occlusion. On the contrary, the PTS+OH performs robustly in frequently merging and splitting among the targets. However, when the interacting targets with similar

appearances suddenly change their motion directions (shown as $t=976$ in Fig. 11), the PTS+OH fails in tracking them correctly. This is because the strategy of the player (target) is the motion of the player (target). Suddenly changing of the motion direction results in approaching the wrong Nash Equilibrium. This makes the PTS+OH fail to track the correct targets.

In Fig. 12 ('CAVIAR'), target 1 merges with target 2 from $t=321$ to $t=427$ and they split at $t=428$. Then target 2 merges with target 9 from $t=436$ to $t=469$. The CTS-OH loses the targets or tracks the merged detection as a target as they merge and re-tracks them as a newborn target as they split, while the PTS+OH succeeds in tracking all the interacting targets. As shown in Fig. 12, target 3 needs to walk through a bright region before he meets the other targets. Because of the similar appearances between target 3 and the bright region (background), the detection is so weak that no target is detected at some time steps. It results in losing target 3 and then re-tracking target 3 as a newborn target (shown as target 8 in the second row and target 9 in the third row in Fig. 12). Moreover, because the size of the target is assumed constant during the occlusion, the tracking precision is affected once the detection obtained before the occlusion is weak (shown as target 9 in the third row in Fig. 12 from $t=436$ to $t=469$).
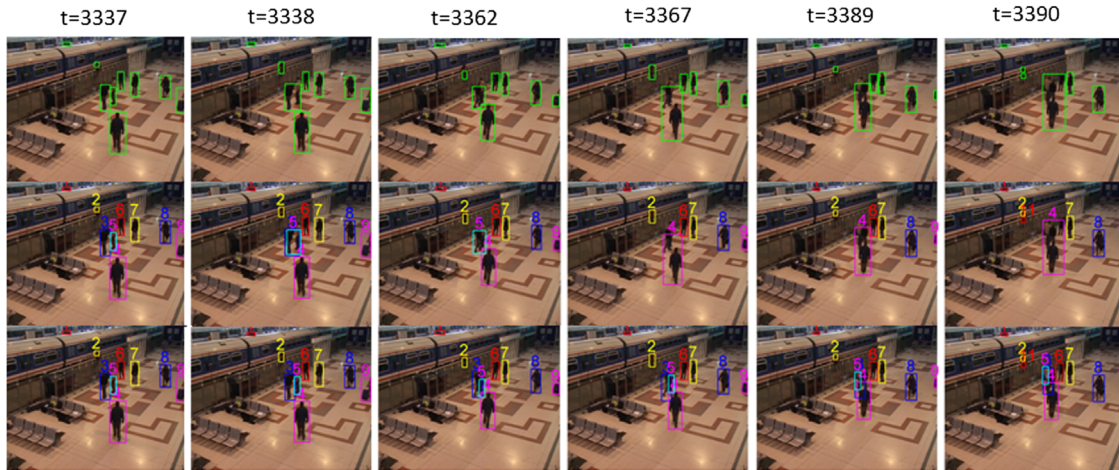
Fig. 14. Tracking results comparison for 'PETS2006 #2'. The first row: detection results; second row: tracking results of the CTS-OH; third row: tracking results of the PTS+OH. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

In Fig. 13 ('PETS2006 #1'), the first two targets come near and merge from $t=752$ to $t=775$. Then they merge with the third target from $t=776$ to $t=803$. Finally the third target splits out while the other two targets remain merge from $t=804$. As stated in Table 1, the main challenge is that the partial and the total occlusions are occurred in the above three interacting targets who have similar appearances. The CTS-OH cannot handle this challenging issue while the PTS+OH can successfully track all the occlusion targets. Moreover, as shown $t=812$ in Fig. 13, both of the tracking systems mis-track the new detected noise as a newborn target (target 10 in the second row and target 6 in the third row in Fig. 13). This could be solved by incorporating a robust birth intensity estimation algorithm for the GM-PHD filter, which will be explored in our future work.

In Fig. 14 ('PETS2006 #2'), there are totally 9 targets involved and most of them move closely, which makes the occlusion reasoning a challenge. Moreover, the complexity of the occlusion is increased gradually and three occlusion targets have similar appearances, which are also challenging issues for occlusion handling. For example, the occlusion first occurs in targets 3 and 5 from $t=3338$ to $t=3366$, then occurs in targets 3, 4 and 5 from $t=3367$ to $t=3389$, finally occurs in targets 3, 4, 5 and 6 from $t=3390$ to $t=3400$. Though, the PTS+OH can correctly determine the occlusion region and successfully track all the occlusion targets, while the CTS-OH loses some of the occlusion targets or tracks the merge detection as a target.

5.2. Quantitative analysis

To show the tracking precision of the targets' positions in occlusion, comparisons between the ground truth and the obtained targets' trajectories are given in Fig. 15 while the average position errors (APE) calculated for these targets are given in Table 2. The APE of a target i in x-direction and y-direction are defined as

$$\text{APE}(x) = \left(\sum_{t=1}^{N_{f,i}} |l_{x,t}^i - l_{x,g,t}^i| \right) / N_{f,i} \tag{29}$$

$$\text{APE}(y) = \left(\sum_{t=1}^{N_{f,i}} |l_{y,t}^i - l_{y,g,t}^i| \right) / N_{f,i} \tag{30}$$

where $\{l_{x,t}^i, l_{y,t}^i\}$ and $\{l_{x,g,t}^i, l_{y,g,t}^i\}$ are the tracked location and the ground truth location of a target i, respectively, for time t. $N_{f,i}$ is the total number of frames for the tracked target i.

From Table 2, the percentage of APE over the average target size is about 3.9–19% for x-direction and about 0.9–20.4% for y-direction, while the percentage of APE over the image size is about 0.2–2.4% for x-direction and about 0.2–3.7% for y-direction. The main reasons of the large APE (shown as the bold letters in Table 2) are the weak detections caused by the significant shades (target 1 in 'ViSOR #1' and targets 1 and 4 in 'PETS2006 #1') and the illumination variation or by similar appearances between the targets and the background (target 3 in 'CAVIAR' and target 5 in 'PETS2006 #2'), and ID switch caused by the sudden changing of motion directions (targets 1 and 3 in 'ViSOR #3'). Without regard to the negative effect of above factors, the percentage of APE over the average target size is improved to 3.9–12.7% for x-direction and to 0.9–12.8% for y-direction.

The CLEAR MOT metrics [42] is also used to evaluate the occlusion tracking performance. This returns a precision score MOTP (intersection over union of bounding boxes) and an accuracy score MOTA (composed of false positive rate (FPR), miss rate (MR) and mismatch rate (MMR)).

$$\text{MOTP} = \frac{\sum_{i,t}[S(gb_t^i \cap tb_t^i)/S(gb_t^i \cup tb_t^i)]}{\sum_t c_t} \tag{31}$$

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{32}$$

where $S(\bullet)$ is a function to compute the area. gb_t^i and tb_t^i are the ground truth box and the associated tracked box of a target i, respectively, for time t. c_t is the number of matched targets found for time t. m_t, fp_t, mme_t, and g_t are the number of misses, of false positives, of mismatches, and of ground truth, respectively, for time t.

To show the robustness of the proposed occlusion handling algorithm, the performance is only evaluated in the occlusion period. The results in Table 3 show that the PTS+OH outperforms the CTS-OH both in MOTP and MOTA. For 'ViSOR #3' because of the challenging situation of suddenly changing the motion directions of targets 1 and 3 (as shown in Fig. 11), the IDs are switched between them and thus results in non-zero MMR in the PTS+OH. For 'CAVIAR', because of the weak detection of target 3, the PTS+OH loses the target and re-tracks it as a newborn target in
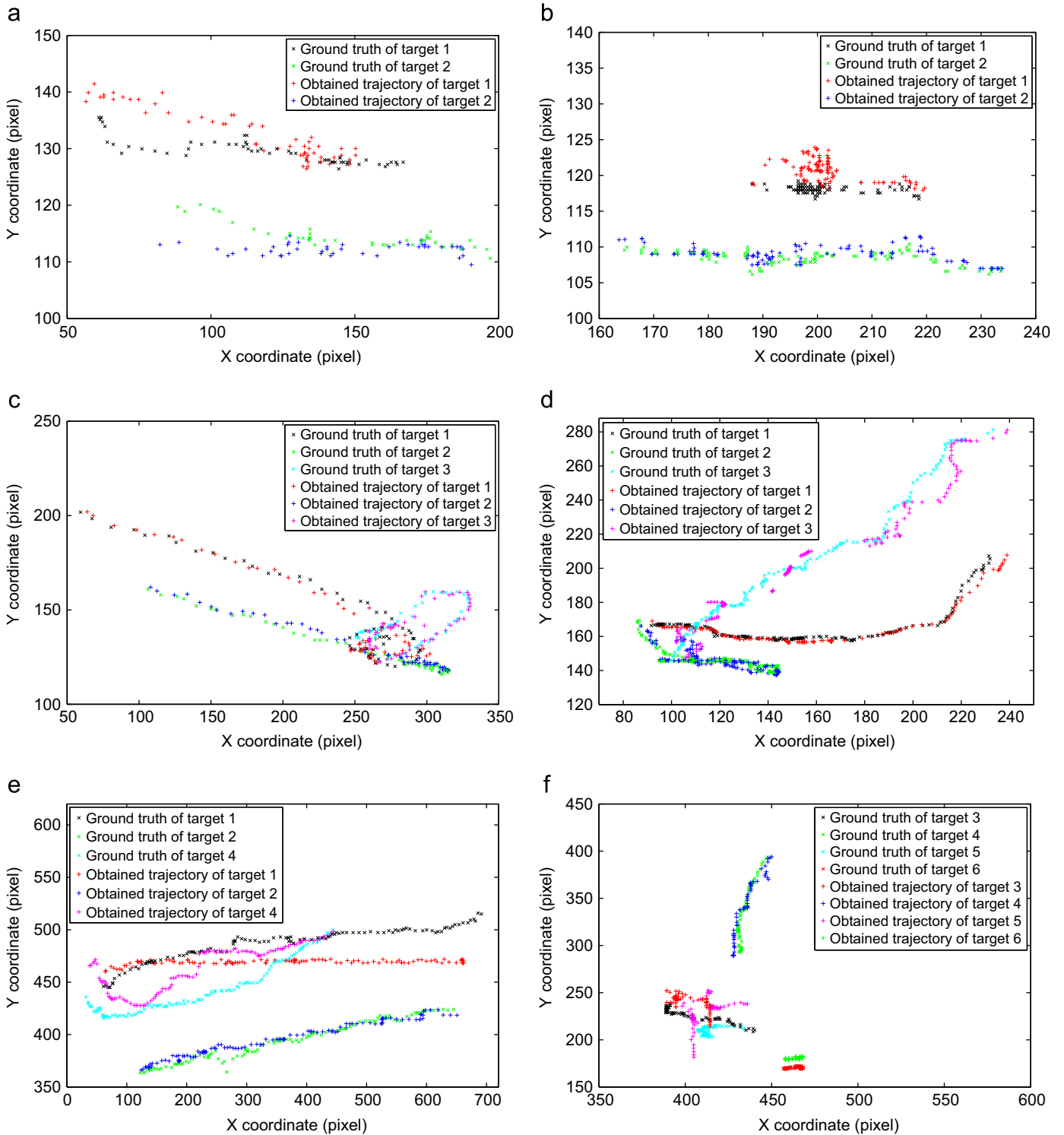
Fig. 15. Comparisons between the ground truth and the obtained trajectories of targets in occlusion. (a) For data set 'ViSOR #1', (b) for data set 'ViSOR #2,' (c) for data set 'ViSOR #3', (d) for data set 'CAVIAR', (e) for data set 'PETS2006 #1' and (f) for data set 'PETS2006 #2'. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

some time steps (shown as in Fig. 12) which results in non-zero MR and MMR.

We also compare our occlusion handling method with the state-of-the-art occlusion handling methods including the kernel density based method reported by Yang et al. [25], the color appearance model based method reported by Vezzani et al. [26], and the spatial-color appearance model based method reported

by Hu et al. [29]. All the above methods can track the targets in mutual occlusion. The main difference is the tracking precision. The results in Table 4 show the MOTP scores of the above methods for the data sets listed in Table 1. Compared with the kernel density based method [25], our method achieves the better MOTP scores for four data sets including 'ViSOR #2', 'ViSOR #3', 'CAVIAR' and 'PETS2006 #1' while gets the lower MOTP scores for

the data sets 'ViSOR #1' and 'PETS2006 #2'. The reason for the lower MOTP scores is that we only use a simple background subtraction method for object detection. The tracking precision score can be decreased once the targets are over-detected (shown as target 1 in Fig. 9) or weak-detected (shown as targets 2 and 5 in Fig. 14). Compared with the color appearance model based method [26], our method outperforms for all data sets. In particular, our method significantly improves the tracking precision when the targets have similar appearances (shown as the MOTP scores of the data sets 'ViSOR #2', 'CAVIAR', 'PETS2006 #1' and 'PETS2006 #2' in Table 4). Compared with the spatial-color appearance model based method [29], our method can achieve the better scores when the targets are in severe occlusion (shown as the MOTP scores of the data sets 'ViSOR #2', 'ViSOR #3', 'CAVIAR' and 'PETS2006 #1' in Table 4). This is because we consider the interferences by other interacting targets in mutual occlusion in the proposed appearance model. The lower MOTP scores for the data sets 'ViSOR #1' and 'PETS2006 #2' are caused by the weak object detection method.

5.3. Discussions

Although all the above experiments have validated the PTS+OH in handling the challenging issues listed in Table 1, some other issues need to be discussed furthermore.

(1) Application restriction: The GM-PHD filter has a closed form solution under linear Gaussian assumptions on the target dynamics and births (see Section 3.2). However, the linear Gaussian multi-target model is not general enough to accommodate maneuvering targets, once these targets follow jump Markov system models. To apply system for some more general cases, the sequential Monte Carlo implementation of the PHD filter [43] can be incorporated to accommodate any Markovian target dynamics.

(2) Noises elimination: Some noises exist in the detection results due to the environmental uncertainty (shown as $t=812$ in Fig. 13). In this paper, we use the generic GM-PHD filter to track multiple moving targets. It has the ability to filter those noises close to the survival targets. However, as noises are far away from the survival targets, they may be tracked as the newborn targets. To remedy this, a robust birth intensity estimation algorithm should be explored to accurately determine the intensity of the newborn targets.

(3) Sudden motion changing: The proposed game-theoretical occlusion handling algorithm selects the motion of a target as the strategy. This is reasonable and effective as the interacting targets move routinely. However, as the interacting targets with similar appearances move rapidly meanwhile change their motion directions drastically (shown as $t=976$ of 'ViSOR #3' in Fig. 11), the players (targets) involved in the constructed game may reach a wrong Nash Equilibrium. Consequently, the proposed algorithm fails in tracking these targets correctly for this challenging situation.

(4) Processing speed: The proposed tracking system is implemented in Matlab, with an Inter Core2Duo 2.20 GHz and 2 GB of memory. Without any code optimization, the average runtimes per frame during occlusion tracking for the above data sets are about 0.25–0.80 fps. More than 95% of the runtimes are consumed in searching the Nash Equilibrium

Table 2
Average position errors (APE) of targets in occlusion (unit: pixel).

Data Sets	Average target size	Image size	Target's ID	[a]Or/ Od	APE (x)	APE (y)
ViSOR #1	70×190	360×270	1	Or	**7.4**	**3.3**
			2	Od	4.9	1.8
ViSOR #2	30×80	384×288	1	Or	1.4	2.7
			2	Od	1.4	0.7
ViSOR #3	20×55	384×288	1	Or+Od	**3.8**	**3.1**
			2	Od	1.2	1.6
			3	Or+Od	**2.7**	**2.1**
CAVIAR	48×27	384×288	1	Or	2.1	1.0
			2	Od	3.8	1.6
			3	Or	**9.1**	**5.5**
PETS2006 #1	100×170	720×576	1	Or	**10.4**	**19.9**
			2	Od	12.2	3.2
			4	Or+Od	**5.0**	**20.1**
PETS2006 #2	36×112	720×576	3	Or+Od	4.6	14.4
			4	Or	2.5	3.1
			5	Od	**5.3**	**21.3**
			6	Od	1.4	10.2

[a] Or represents an occluding target; Od represents an occluded target; Or+Od represents either an occluding target or an occluded target in different time steps.

Table 4
Occlusion tracking performance comparison with the state-of-the-art methods.

Data sets	Our method	Kernel density based method [25]	Color appearance model based method [26]	Spatial-color appearance model based method [29]
	MOTP (%)	MOTP (%)	MOTP (%)	MOTP (%)
ViSOR #1	81.68	82.56	81.53	82.24
ViSOR #2	90.83	87.92	81.88	85.51
ViSOR #3	88.64	83.57	87.96	88.43
CAVIAR	81.87	79.14	74.59	79.68
PETS2006 #1	83.94	79.85	72.61	75.96
PETS2006 #2	79.09	81.26	75.87	80.42

Table 3
Occlusion tracking performance comparison with the CTS-OH.

Data sets	Occlusion lasting time (frames)	Tracking performance of the CTS-OH				Tracking performance of the PTS+OH			
		MOTP (%)	MOTA (%)	MR (%)	MMR (%)	MOTP (%)	MOTA (%)	MR (%)	MMR (%)
ViSOR #1	38	55.21	49.56	48.72	1.72	81.68	100	0	0
ViSOR #2	76	65.63	50.65	49.35	0	90.83	100	0	0
ViSOR #3	50	66.68	55.33	36.00	8.00	88.64	96.00	0	**4.00**
CAVIAR	150	62.18	69.60	28.63	1.57	81.87	98.23	**0.59**	**0.98**
PETS2006 #1	108	66.64	56.74	30.43	1.96	83.94	89.78	0	0
PETS2006 #2	68	75.48	85.88	12.93	0	79.09	98.81	0	0

of the proposed game-theoretical occlusion handling algorithm because it is a pixel-wise iteration process. To remedy this, employing more efficient appearance model will be helpful.

6. Conclusion

We have developed a GM-PHD filter based multi-target visual tracking system with the game-theoretical occlusion handling algorithm. The proposed target 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 interacting targets had similar appearances in occlusion. We also proposed an effective two-step occlusion reasoning algorithm to correctly determine the occlusion region, in which both the support regions and the detection results were incorporated. Then we proposed a robust game-theoretical occlusion handling algorithm to deal with the mutual occlusion problem in multiple interacting targets. We constructed an n-person, non-zero-sum, non-cooperative game and regarded the individual targets in the occlusion region as the players in the game to compete for the maximum utilities by using the certain strategies. We selected the Nash Equilibrium of the game as the optimal estimation of the locations of the players within the occlusion region. Experiments on video sequences showed that the proposed tracking system achieved promising results in handling mutual occlusions in multiple interacting targets.

In future work we plan to investigate more efficient appearance model for occlusion handling. Furthermore, we would like to investigate more effective strategies in a game to cope with the sudden motion changing problem.

Conflict of interest

None declared.

References

[1] M. Kim, Correlation-based incremental visual tracking, Pattern Recognition 45 (3) (2012) 1050–1060.
[2] D.A. Ross, J. Lim, R.S. Lin, M.–H. Yang, Incremental learning for robust visual tracking, International Journal of Computer Vision 77 (2008) 125–141.
[3] B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, in: IEEE Conference on Computer Vision and Pattern Recognition. June 2009, pp. 983–990.
[4] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5) (2003) 564–577.
[5] T. Bai, Y.F. Li, Robust visual tracking with structured sparse representation appearance model, Pattern Recognition 45 (6) (2012) 2390–2404.
[6] I.R. Goodman, R. Mahler, H.T. Nguyen, Mathematics of Data Fusion, Kluwer Academic Press, Norwell, MA, 1997.
[7] R. Mahler, Multitarget bayes filtering via first-order multitarget moments, IEEE Transactions on Aerospace and Electronic Systems 39 (4) (2003) 1152–1178.
[8] B.-N. Vo, W.K. Ma, The Gaussian mixture probability hypothesis density filter, IEEE Transactions on Signal Processing 54 (11) (2006) 4091–4104.
[9] B. Ristic, D. Clark, B.-N. Vo, Improved SMC implementation of the PHD filter, in: 13th International Conference on Information Fusion. July 2010, pp. 1–8.
[10] I.E. Maggio, M. Taj, A. Cavallaro, Efficient multi-target visual tracking using random finite sets, IEEE Transactions on Circuits and Systems for Video Technology 18 (8) (2008) 1016–1027.

[11] Y.D. Wang, J.K. Wu, A.A. Kassim, W.M. Huang, Data-driven probability hypothesis density filter for visual tracking, IEEE Transactions on Circuits and Systems for Video Technology 18 (8) (2008) 1085–1095.
[12] B.-N. Vo, B.-T. Vo, N.-T. Pham, D. Suter, Joint detection and estimation of multiple objects from image observations, IEEE Transactions on Signal Processing 58 (10) (2010) 5129–5141.
[13] E. Maggio, A. Cavallaro, Learning scene context for multiple object tracking, IEEE Transactions on Image Processing 18 (8) (2009) 1873–1884.
[14] N.T. Pham, W.M. Huang, S.H. Ong, Tracking multiple objects using probability hypothesis density filter and color measurements, in: IEEE International Conference on Multimedia and Expo. July 2007, pp. 1511–1514.
[15] J.J. Wu, S.Q. Hu, Y. Wang, Probability-hypothesis-density filter for multitarget visual tracking with trajectory recognition, Optical Engineering 49 (12) (2010) 12970-11-12970-19.
[16] S.L. Dockstader, A.M. Tekalp, Multiple camera tracking of interacting and occluded human motion, Proceedings of IEEE 89 (10) (2001) 1441–1455.
[17] Z. Wu, N.I. Hristov, T.L. Hedrick, T.H. Kunz, M. Betke, Tracking a large number of objects from multiple views, in: IEEE 12th International Conference on Computer Vision. October 2009, pp. 1546–4553.
[18] S.M. Khan, M. Shah, Tracking multiple occluding people by localizing on multiple scene planes, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (3) (2009) 505–519.
[19] L. Cai, L. He, Y. Xu, Y. Zhao, X. Yang, Multi-object detection and tracking by stereo vision, Pattern Recognition 43 (12) (2010) 4028–4041.
[20] T. Zhao, R. Nevatia, B. Wu, Segmentation and tracking of multiple humans in crowded environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (7) (2008) 1198–1211.
[21] Z. Khan, T. Balch, F. Dellaert, MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 1960–1972.
[22] Y. Wu, T. Yu, G. Hua, Tracking appearances with occlusions, in: IEEE Conference on Computer Vision and Pattern Recognition. June 2003, pp. I-789–I-795.
[23] W.M. Hu, X. Li, W.H. Luo, X.Q. Zhang, Single and multiple object tracking using log-Euclidean Riemannian subspace and block-division appearance model, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (12) (2012) 2420–2440.
[24] G.R. Li, W. Qu, Q.M. Huang, A multiple targets appearance tracker based on object interaction models, IEEE Transactions on Circuits and Systems for Video Technology 22 (3) (2012) 450–464.
[25] M. Yang, T. Yu, Y. Wu, Game-theoretic multiple target tracking, in: IEEE 11th International Conference on Computer Vision. October 2007, pp. 1–8.
[26] R. Vezzani, C. Grana, R. Cucchiara, Probabilistic people tracking with appearance models and occlusion classification: The ad-hoc system, Pattern Recognition Letters 32 (2011) 867–877.
[27] J.L. Xing, H.Z. Ai, L.W. Liu, S.H. Lao, Multiple player tracking in sports video: A dual-mode two-way Bayesian inference approach with progressive observation modeling, IEEE Transactions on Image Processing 20 (6) (2011) 1652–1667.
[28] V. Papadourakis, A. Argyros, Multiple objects tracking in the presence of long-term occlusions, Computer Vision and Image Understanding 114 (2010) 835–846.
[29] W.M. Hu, X. Zhou, M. Hu, S. Manbank, Occlusion reasoning for tracking multiple people, IEEE Transactions on Circuits and Systems for Video Technology 19 (1) (2009) 114–121.
[30] A. Senior, A. Hampapur, Y.L. Tian, L. Brown, S. Pankanti, R. Bolle, Appearance models for occlusion handling, Image Vision Computing 24 (11) (2006) 1233–1243.
[31] C.H. Kuo, C. Huang, R. Nevatia, Multi-target tracking by on-line learned discriminative appearance models, in: IEEE Conference on Computer Vision and Pattern Recognition, June 2010, pp. 685–692.
[32] Y.M. Seong, H.W. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition 45 (9) (2012) 3451–3462.
[33] H.Z. Wang, D. Suter, K. Schindler, C. Shen, Adaptive object tracking based on an effective appearance filter, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (9) (2007) 1661–1667.
[34] J. Nash, Two-person cooperative games, Econometrica 21 (1) (1953) 128–140.
[35] E.N. Barron, Game Theory: an Introduction, Wiley, 2008.
[36] D.B. Gu, A game theory approach to target tracking in sensor networks, IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 41 (1) (2011) 2–13.
[37] C. Soto, B. Song, A.K. Roy-Chowdhury, Distributed multi-target tracking in a self-configuring camera network, in: IEEE Conference on Computer Vision and Pattern Recognition. June 2009, pp. 1486–1493.
[38] X.Q. Zhang, W.M. Hu, G. Luo, S. Manbank, Kernel-yesian framework for object tracking, in: 8th Asian Conference on Computer Vision. November 2007, pp. 821–831.
[39] Available from: ⟨http://imagelab.ing.unimore.it/visor/video_categories.asp⟩.
[40] Available from: ⟨http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/⟩.
[41] Available from: ⟨http://www.cvg.rdg.ac.uk/PETS2006/data.html⟩.
[42] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: The CLEAE MOT Metrics, EURASIP Journal of Image and Video Processing 2008 (2008) 1–10.
[43] B.–N. Vo, S. Singh, A. Doucet, Sequential Monte Carlo methods for multitarget filtering with random finite sets, IEEE Transactions on Aerospace and Electronic Systems 41 (4) (2005) 1224–1245.

Xiaolong Zhou received the B.S. and M.S. degrees in mechanical engineering from Fuzhou University, Fuzhou, China, in 2007 and 2010, respectively. Currently, he is a Ph.D. student in the Department of Mechanical and Biomedical Engineering at City University of Hong Kong, Hong Kong. His research interests include visual tracking, 3D reconstruction and robot vision.

Y.F. Li received the B.S. and M.S. degrees in electrical engineering from Harbin Institute of Technology China. He obtained the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford in 1993. From 1993 to 1995 he was a postdoctoral research staff in the Department of Computer Science at the University of Wales, Aberystwyth, UK. He joined City University of Hong Kong in 1995. His research interests include robot sensing, sensor guided manipulation, robot vision, 3D vision and visual tracking. He has served as an Associate Editor of IEEE Transactions on Automation Science and Engineering (T-ASE) and is currently serving as Associate Editor of IEEE Robotics and Automation Magazine (RAM).

Bingwei He received his B.S. degree in mechanical engineering from the Yanshan University, Hebei, China, in 1992, and his M.S. and Ph.D. degrees in mechanical engineering from the Xi'an JiaoTong University, Shanxi, China, in 1999 and 2003, respectively. He is currently a professor at the School of Mechanical Engineering at Fuzhou University. His research interests include robot vision, reverse engineering, and rapid prototyping.