Pattern Recognition ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity

Yeol-Min Seong, HyunWook Park*

Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Article history:
Received 12 July 2011
Received in revised form
13 December 2011
Accepted 1 March 2012

Keywords:
Change perception
Distinguishability
Motion likelihood field
Multiple target tracking
Recognition process of human

ABSTRACT

Object tracking is crucial to surveillance systems, which provide target information including position, size, and velocity. This paper presents a data association process combining two primary components of visual features and spatiotemporal prediction. In addition, the change perception and the visual distinguishability are utilized to adaptively combine the two primary components. The proposed spatiotemporal prediction is performed on several consecutive frames in order to cover the irregular motion of targets. The prediction is then filtered with a change perception mask to determine whether the candidate observations have motion or not. In addition, the level of contribution of a visual feature is adjusted by the proposed distinguishability to maintain a reward-penalty balance. The proposed method is applied to various video sequences having small targets and abrupt motions, and the experimental results show consistent tracking performance.

1. Introduction

Target tracking is an important element in computer vision, such as robot navigation, automatic video indexing, and visual surveillance for security purposes. Since a tracking system should be able to track multiple targets and retrieve the correct identity of each target in various situations, most tracking algorithms have utilized two primary components, the object's visual features and the spatiotemporal prediction [1–3].

In the last few decades, many tracking methods have been proposed to solve the tracking problems based on the particle filter [1,3–7]. These methods basically perform a probabilistic propagation process using a stochastic motion model to obtain an estimation of the posterior distribution without restrictive linear Gaussian assumptions regarding the transition and likelihood models. Color and segmentation cues [8] or pixel-wise appearance models [5] were applied to the particle filter. Markov Chain Monte Carlo (MCMC) method was used to reduce the computational cost of high-dimensional state space in multi-object tracking [9]. Variants of the MCMC-based particle filter have been also proposed to solve specific problems such as multiple interacting targets [7], sequences with abrupt motion [10], small object tracking [11], and multiple motion models [12].

When these algorithms are applied to multiple targets, the computational cost is exponentially increased, because a large number of particles are required to estimate the posterior density of the object's state well. In addition, the particles cannot be accurately predicted if the object's motion is complex [12]. To reduce the heavy computation amount for multiple targets, several algorithms initially performed object detection using background subtraction [13–15] or supervised learning [16], and then performed data association to assign consistent identities. In [17], motion detection, background/foreground modeling, and shadow removal techniques were introduced for human motion detection and tracking. Recently, learning based tracking methods were proposed to improve the tracking performance [18–20]. The Tracking-Learning-Detection (TLD) method [18,19] improved the performance of a binary classifier by using the structured unlabeled data. The positive and negative (P–N) constraints of the TLD method provided a feedback about the performance of the classifier which was iteratively improved in a bootstrapping fashion.

The main goal of the proposed method is to develop a robust algorithm to track small objects, especially having low visual distinguishability and irregular motion. In the present work, a new likelihood model for data association is proposed based on the object recognition process of the human brain. The data association process follows the theory of [21] and the two primary components are used in a simultaneous manner. In addition to the two primary components, the change perception and the distinguishability are also considered for a robust and stable data association of indistinguishable objects having irregular motion.

The remainder of this paper is organized as follows. In Section 2, four components of the likelihood model for data association are introduced, which are based on the human object recognition

* Corresponding author. Tel.: +82 42 350 3466; fax: +82 42 350 8361.
E-mail address: hwpark@kaist.ac.kr (H. Park).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.03.005

process. The overall process of the proposed multiple target tracking is described in Section 3. Section 4 presents experimental results, and we conclude the paper in Section 5.

2. Four components of the proposed likelihood model for data association

Most tracking algorithms based on stochastic inference initially predict a few possible candidates using a transition model, and then determine the corresponding identities by comparing the visual features using a likelihood model [5]. Since they sequentially perform the initial prediction of candidates and identification of the object, investigating the visual similarity can be meaningless if the true target is not included in the candidates.

A recent study in the field of cognitive neuroscience demonstrated how an object's identity (ID) and persistence are neurally represented [21], by investigating the influence of spatiotemporal continuity and visual similarity on object recognition using functional magnetic resonance imaging (fMRI) experiments. It was shown that the visual information was integrated into the spatiotemporal information at quite an earlier phase of the recognition process, indicating that it was necessary to modify the previous theories that separate visual information processing and spatiotemporal information processing. Thus, the conventional tracking algorithms involving sequential consideration of the two primary components also can be improved by simultaneously utilizing the primary components.

Fig. 1 shows how four components of the proposed likelihood model for data association are combined. As pointed out in [21], the visual feature provides identities of the targets and the spatiotemporal prediction provides trajectories of the targets. For each observation, the scores of the visual similarity (d_{SS}) and the spatiotemporal prediction (d_{MLF}) are computed and then combined. In addition, the change perception (d_{CP}) and the distinguishability (d_D) are utilized to adaptively combine the spatiotemporal prediction and the visual similarity, respectively.

To accurately predict various motions of targets, the proposed spatiotemporal prediction uses more than one previous frames and the prediction result is filtered by a change perception mask to secure probable candidates for the correct data association. Fig. 2 demonstrates the role of the change perception mask. In Fig. 2(a) and (b) whose objects have regular motion, the object position is well predicted as marked with an ellipse. In Fig. 2(c), however, the tracker misses the true position of the object (marked with a dashed rectangle) due to an abrupt change of motion direction, where the wrong predicted region (marked with an ellipse) is a part of the background. This tracking failure is due to the fact that the small object in the example sequence is hardly distinguishable from clutter in the background. This error can be avoided by applying the proposed change perception, which is described in Section 3.

Visual features such as color, edge, optical flow, and texture are used to identify the target from other objects or clutter. When the visual features become ambiguous, however, the target identification can be difficult, as demonstrated in Fig. 3, where the partly visible trajectories of two objects are marked with dashed arrows. In Fig. 3(a) where two objects are distinguishable, it can be easily estimated as the two arrows cross each other. If two objects are much similar, as shown in Fig. 3(b), it is not easy to estimate whether the arrows of the objects cross each other. The spatiotemporal prediction becomes more important than the visual features for identification, if there are rival candidates, as demonstrated in Fig. 3(b). Thus, a learning process based on the distinguishability is proposed in this paper to adjust the contribution level of the visual features for robust tracking. Details of the distinguishability are described in Section 3.

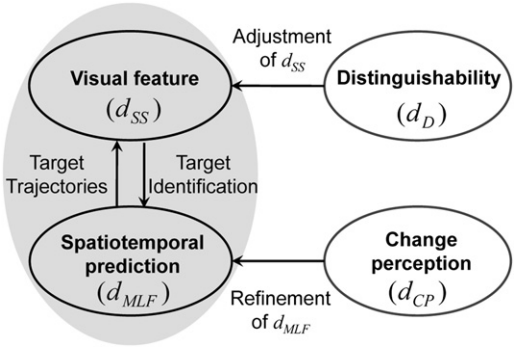


Fig. 1. Four spatial and temporal features of the proposed likelihood model for data association and their relationship.

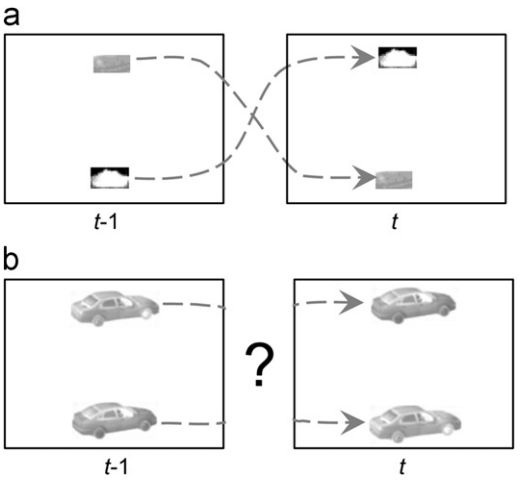


Fig. 3. Examples of ambiguous identification for crossing objects: (a) Clearly distinguishable objects, and (b) indistinguishable objects.

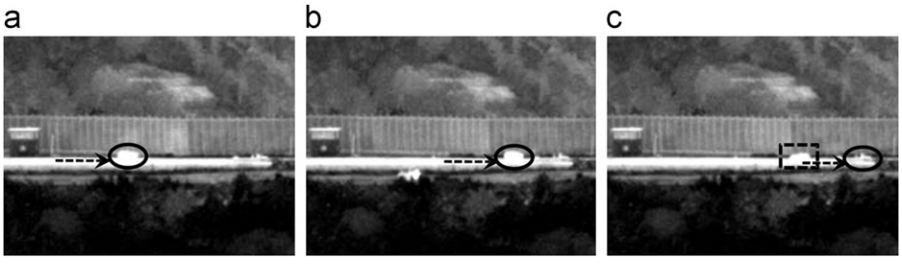


Fig. 2. Role of the change perception for object tracking. The dashed arrows denote the predicted motion of the target and the ellipses are the predicted positions of the target: (a) Time $t-2$, (b) time $t-1$, and (c) tracking fail at time t, where dashed rectangle is the real target.

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

3. The proposed multiple target tracking

The overall flowchart of the proposed multiple target tracking algorithm is presented in Fig. 4. After target IDs of the initial frame are assigned, the proposed tracking method is applied to the following frames. For every frame, a change perception mask is generated to find the appeared and disappeared regions, as described in Section 3.3. If new observations that do not correspond to motion likelihood fields (MLFs) of the previous frame (described in section 3.2) appear in the current frame, its size is evaluated to eliminate clutters. After the clutter elimination, the observation is considered as a new target and a new ID is assigned to the target. For the existing targets that are tracked from the previous frames, the corresponding IDs are assigned to each observation in the current frame using the proposed likelihood model as described in Section 3.5. After assigning IDs to all targets, the tracking process is repeated for the next frame.

To perform multiple target tracking, object representation of the ith object in the tth frame is defined as follows,

$$\mathbf{O}_t^i = \{\mathbf{x}_t^i, \dot{\mathbf{x}}_t^i, \ddot{\mathbf{x}}_t^i, \boldsymbol{\sigma}_t^{i2}, \mathbf{s}_t^i\}, \tag{1}$$

where $\mathbf{x}_t^i$ denotes the object center vector, and $\mathbf{s}_t^i$ is the size vector of the ith object in the x, y coordinates. $\dot{\mathbf{x}}_t^i$ and $\ddot{\mathbf{x}}_t^i$ are the velocity and the acceleration vectors of the object, respectively, and $\boldsymbol{\sigma}_t^{i2}$ is the variance vector of the velocity. Using this representation, a detailed explanation of the proposed tracking algorithm for multiple targets is described in the following sections.

3.1. Visual feature of similarity score

A color histogram is an important visual feature, which is invariant to rotation and scale changes and robust to partial occlusions [22]. However, it cannot reflect the spatial shape of objects. Use of the edge energy can reflect the object shape and increase robustness to illumination change [22]. The visual feature of the proposed method is a similarity score, which utilizes both the color histogram and the edge energy. The similarity score, d_{SS}, between the ith object in the tth frame and jth object in the $t-1$th frame is defined as follows,

$$d_{SS}(\mathbf{O}_t^i, \mathbf{O}_{t-1}^j) = s_h \rho[h(\mathbf{O}_t^i), h(\mathbf{O}_{t-1}^j)] + s_e \rho[e(\mathbf{O}_t^i), e(\mathbf{O}_{t-1}^j)], \tag{2}$$

$$\rho[h(\mathbf{O}_t^i), h(\mathbf{O}_{t-1}^j)] = \sum_{u=1}^{m} \sqrt{h(\mathbf{O}_t^i)_u h(\mathbf{O}_{t-1}^j)_u}, \tag{3}$$

$$\rho[e(\mathbf{O}_t^i), e(\mathbf{O}_{t-1}^j)] = \sum_{v=1}^{N} \sqrt{e(\mathbf{O}_t^i)_v e(\mathbf{O}_{t-1}^j)_v}, \tag{4}$$

where $h(\mathbf{O}_t^i)$ denotes the normalized histogram of the ith object in the tth frame, which is quantized into m-bins [6]. $e(\mathbf{O}_t^i)$ denotes the normalized edge energy of the object, which is quantized into N-bins. In (2), s_h and s_e are the proportion constant of the histogram and the edge energy, respectively. $h(\mathbf{O}_t^i)_u$ and $e(\mathbf{O}_t^i)_v$ are the histogram values of uth and vth bins, respectively.

3.2. Spatiotemporal prediction

To make more robust prediction for irregular motions as well as regular motions, the proposed method chooses as many candidate positions as possible based on the combination of directions and magnitudes of the previous motions. A graphical example of the spatiotemporal prediction process is given in Fig. 5, where a black square with $\mathbf{x}_t^i$ denotes the ith object's position at time t. At time t, n-previous positions and velocities are used to predict the Motion Likelihood Field (MLF) as follows: first, the candidate unit vectors, which are marked with arrows at $\mathbf{x}_{t-1}^i$, are calculated using the previous motion directions in order to obtain the candidate motion directions of the MLF. These unit vectors show various directions when the previous motion of the object is irregular. The directions are then quantized with a quantization step of 15°. Then, the candidate positions at tth frame, which are marked with small circles, are determined by using the motion velocities of the previous n-frames. The (p, q)th candidate velocity vector, $\dot{\mathbf{x}}_{t-1,p,q}^i$, is defined as follows,

$$\dot{\mathbf{x}}_{t-1,p,q}^i = |\dot{\mathbf{x}}_p^i| \cdot \mathbf{u}_{t-1}^{i,q}, \quad (p = t-n, \cdots, t, q = 1, \cdots, n_q), \tag{5}$$

where $|\dot{\mathbf{x}}_{t-1}^i| = ((|\mathbf{x}_{t-1}^i - \mathbf{x}_{t-2}^i|)/T)$ denotes the magnitude of the velocity vector at $t-1$, and $\mathbf{u}_{t-1}^{i,q}$ is the qth quantized unit vector. n_q denotes the number of the quantized unit vectors, and the interval of the time step is $T=1$ in this paper. Since $\mathbf{x}_t^i$ is not defined when $p=t$, $|\dot{\mathbf{x}}_t^i|$ is estimated using the acceleration vector

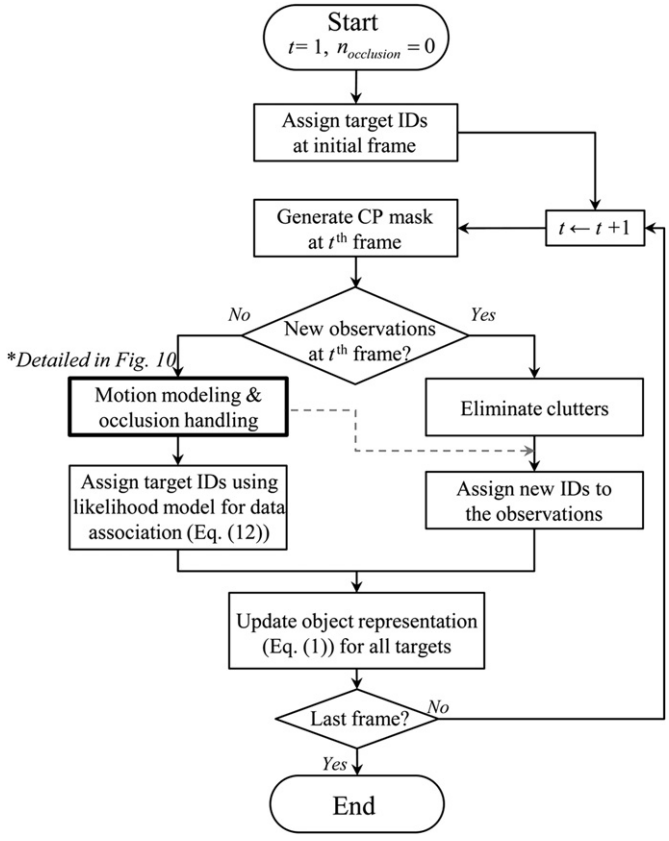


Fig. 4. Overall flowchart of the proposed multiple target tracking algorithm.

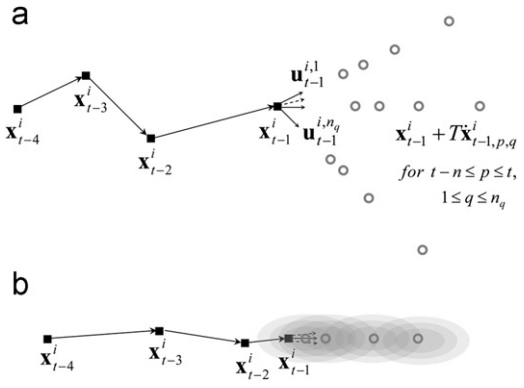


Fig. 5. Spatiotemporal prediction process: (a) Motion with various directions, and (b) relatively constant motion.

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

as follows,

$$|\dot{\mathbf{x}}_t^i| = |\dot{\mathbf{x}}_{t-1}^i| + |\ddot{\mathbf{x}}_{t-1}^i| \cdot T, \tag{6}$$

where $|\ddot{\mathbf{x}}_{t-1}^i| = ((|\dot{\mathbf{x}}_{t-1}^i| - |\dot{\mathbf{x}}_{t-2}^i|)/T)$ is the magnitude of the acceleration vector at $t-1$.

The MLF as the spatiotemporal prediction, is given by the estimated candidate positions and the standard deviation of the velocity, $\boldsymbol{\sigma}_t^i$, which is calculated separately for horizontal and the vertical directions. The MLF of the jth object at $t-1$, having the shape of mixture of Gaussians, is defined using the definition of bivariate kernel density estimation [23] as follows,

$$D_{t-1}^j(\mathbf{x};\mathbf{H}) = C \sum_{p=t-n}^{t} \sum_{q=1}^{n_q} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_{t-1,p,q}^j) \tag{7}$$

where $\mathbf{x}=(x,y)^T$ and $\mathbf{x}_{t-1,p,q}^j = \mathbf{x}_{t-1}^j + T \cdot \dot{\mathbf{x}}_{t-1,p,q}^j$, which is the (p, q)th candidate position vector. In Eq. (7), $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$ where

$$\mathbf{H} = \begin{bmatrix} h_{t-1,x}^2 & 0 \\ 0 & h_{t-1,y}^2 \end{bmatrix}$$

is the bandwidth matrix and the kernel function $K(\cdot)$ is the Gaussian kernel in this paper. It is well known that the optimal bandwidth is estimated by minimizing Mean Integrated Squared Error (MISE) [23]. The optimal bandwidth in the horizontal direction is approximated by $h_{t-1,x} = 1.06\sigma_{t-1,x}(n_q(n+1))^{-1/5}$, where $\sigma_{t-1,x}$ is the standard deviation of the velocity in the x direction and $n_q(n+1)$ is the number of candidate positions. The bandwidth in the vertical direction can be similarly defined and C is a normalizing constant that sets the range of $D_{t-1}^j(\mathbf{x};\mathbf{H})$ to [0,1]. The MLF score of the ith observation at t with respect to the jth object at $t-1$ is defined as follows,

$$d_{MLF}(\mathbf{O}_t^i, \mathbf{O}_{t-1}^j) = D_{t-1}^j(\mathbf{x}_t^i;\mathbf{H}). \tag{8}$$

In Fig. 5(b), the Gaussian shape of each MLF is illustrated with an ellipse, where a MLF with a higher score represents that the position of the observation at time t is close to the candidate positions. During the MLF estimation, the number of candidate positions is adaptively determined as $n_q(n+1)$, where n_q decreases as the motion becomes more regular and n is the number of previous frames. For example, the motion direction of the object in Fig. 5(b) can be simplified with the quantization process because the motion is relatively regular, and the computation amount can be reduced as a result.

Fig. 6 shows the simulation results of the MLFs according to various motions, where the arrows show the pathway of an object marked with a rectangle. In case of a regular motion, both methods have similar candidates. However, the proposed MLF has several candidates in the case of irregular motions so that it can reduce the prediction error.

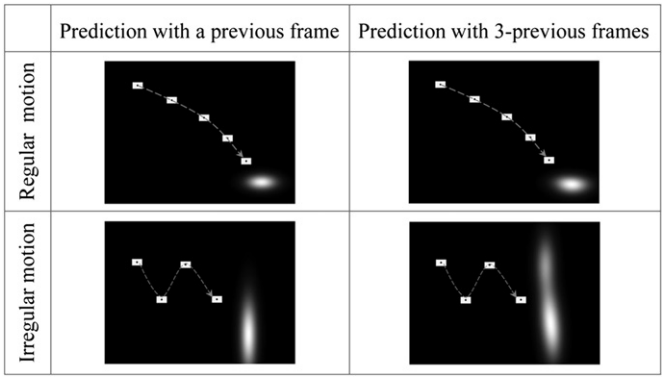


Fig. 6. Motion likelihood fields (MLFs) for regular and irregular motions.

3.3. Change perception

While the proposed MLF can predict abruptly moving targets, it also has the risk of data association with clutter. Although most clutter is different from the targets in terms of d_{SS} and d_{MLF}, some clutter cannot be easily distinguished, as shown in Fig. 2. In the proposed method, a change perception (CP) is defined in order to remove clutters.

A binary mask for the change region, called CP mask, can be defined by thresholding and morphological operations [24] as follows,

$$d_{CP} = M(|f_t - f_{t-1}|, \gamma) \bullet B, \tag{9}$$

where f_t is the tth frame, and γ is a threshold value determined by Otsu's method [25]. $M(|f_t - f_{t-1}|, \gamma)$ denotes the binarization of $|f_t - f_{t-1}|$with the threshold of γ. In the paper, we assume that there are no illumination changes during video capture and a global motion of the camera is compensated before this process, so that the foreground change region can be detected by Eq. (9). Although false negative regions might be still present in d_{CP}, they can be corrected by further data association steps such as evaluation of the visual similarity and MLF score. The morphological closing operation, denoted by $\bullet$, plays a role of filling holes in $M(|f_t - f_{t-1}|, \gamma)$. We initially use a 7×7 window for the structuring element B, and then increase the window size with an increment of two pixels in both horizontal and vertical directions until it becomes the object's size that is assigned in the first frame. The iteration is terminated either when the window size reaches the object's size or when no further change occurs in the CP mask. d_{CP} includes two regions of the appeared and disappeared regions in the current frame (f_t) compared to the previous frame (f_{t-1}). The disappeared region is defined by the intersection region between d_{CP} and the object region in the previous frame. The appeared region is the intersection region between d_{CP} excluding the disappeared region and the object's MLF.

On the basis of the investigation of d_{CP} and MLF, motion types of objects can be classified into four cases, as shown in Fig. 7. When the object has normal motion without occlusions or abrupt stops, it is classified as 'Case 1', which has both appeared and disappeared regions. In 'Case 2', the object is occluded by the background that has only disappeared region. If there is only appeared region, it is classified as 'Case 3', where the object appears from occlusion. If the object abruptly stops or has no motion, it is classified as 'Case 4', where both appeared and disappeared regions are not defined. As shown in Fig. 7, the MLFs (marked with gray ellipses) of 'Case 1' and 'Case 3' are well predicted, whereas the MLFs of 'Case 2' and 'Case 4' are filtered out due to the CP mask. The motion analysis and the occlusion handling are further described in Section 3.5.

The MLF of the irregular motion of Fig. 6 results in a maximum MLF value of max_d_{MLF} in Fig. 8(a), which may mislead the target tracking. To obtain more reasonable candidates, the CP mask of Fig. 8(b) is applied to filter out the false regions. Consequently, the maximum MLF is defined at a more reasonable position (max_$d_{CP}d_{MLF}$), as shown in Fig. 8(c).

3.4. Distinguishability

The ratio of similarity (*RoS*) is defined as the ratio of the second-largest similarity score to the largest similarity score. If the *RoS* is close to 1, there are one or more rival candidates for a

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

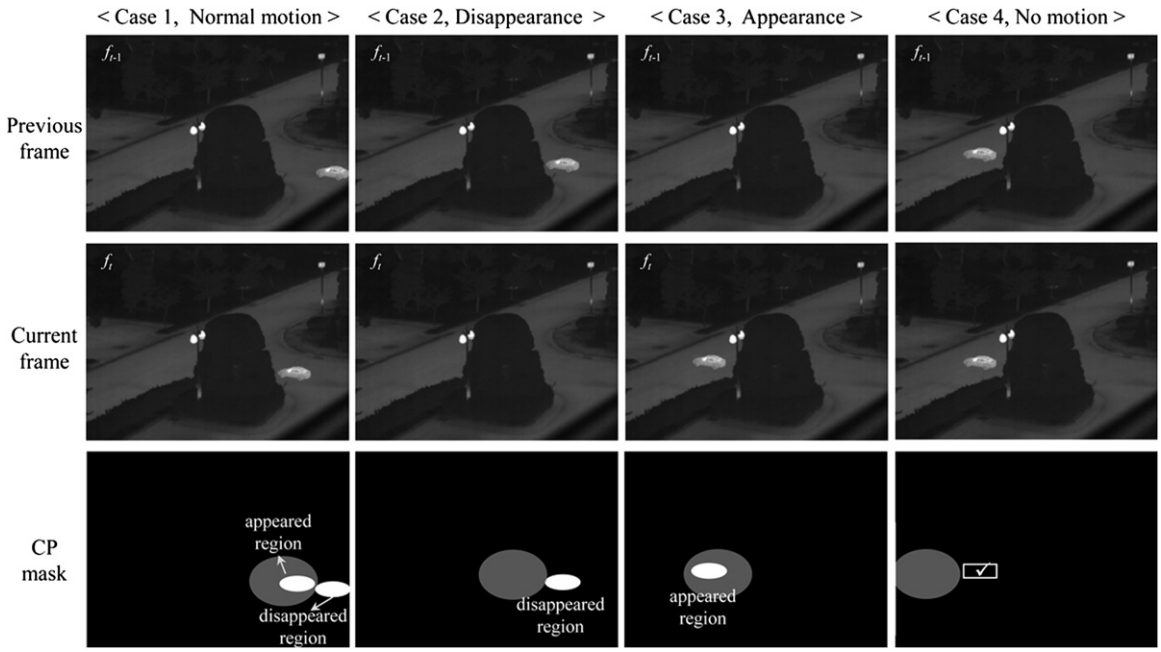


Fig. 7. Motion type classification of cases 1 to 4 according to d_{CP}.

Fig. 8. Change perception filtering of the MLF: (a) MLF, (b) CP mask, (c) Filtered MLF with the CP mask of (b).

decision, whereas a candidate having the largest similarity score becomes a superior candidate if the *RoS* is close to 0.

In order to define the distinguishability of an object, we investigate the characteristics of the d_{SS} and the *RoS* for correct and incorrect matches between two object pairs. 280,000 training samples of data association that include correct and incorrect matches were obtained from training image sets, consisting of 1,100 images of vehicles and pedestrians acquired with both CCD and IR sensors. To simulate the effect of rapid motion, the training images were undersampled in time axis. Small objects (less than 15×10 pixels) were also trained to simulate the effect of long range sequences. In addition, partial occlusion cases were used for the training.

Fig. 9(a) shows the training results, where the horizontal and vertical axes represent the *RoS* and the d_{SS}, respectively. While the correct matches are mostly at the top and the incorrect matches are near the bottom right, some correct and incorrect matches are mixed at the top right of the plot. To classify the correct and incorrect matches in Fig. 9(a), the Fisher linear discriminant analysis (FLDA) [26] is adopted and employed to find a *canonical direction*, marked with a gray line, r. By projecting the data onto the canonical direction, the two-parameter plot can be simplified to a single parameter value, r. Fig. 9(b) shows the probabilities of correct matches, p_C, and incorrect matches, p_I, with respect to r. The samples are clearly distinguished in regions where either $p_C(r)$ or $p_I(r)$ is dominant. Hence, the proposed distinguishability, d_D, is defined as follows,

$$d_D(r) = 1 - \frac{\min(p_C(r), p_I(r))}{\max(p_C(r), p_I(r))}. \tag{10}$$

By using the Lorentz fitting [27], a continuous function is modeled as illustrated in Fig. 9(c) from the measured d_D of Eq. (10), as follows,

$$d_D(r) = 1 - y_0 + 2\frac{A}{\pi} \cdot \frac{w}{4(r-r_c)^2 + w^2}, \tag{11}$$

where $A = 0.1255$, $y_0 = -0.001$, $w = 0.0883$, and $r_c = 0.6641$.

3.5. The proposed likelihood model for data association

The proposed likelihood model for data association, which consists of the four components explained in the previous sections, is defined as follows,

$$d(\mathbf{O}_t^i, \mathbf{O}_{t-1}^j) = \frac{1}{d_{CP}(x_t^i, y_t^i) + d_D(r)} (d_{CP}(x_t^i, y_t^i) \cdot d_{MLF}(\mathbf{O}_t^i, \mathbf{O}_{t-1}^j) + d_D(r) \cdot d_{SS}(\mathbf{O}_t^i, \mathbf{O}_{t-1}^j)). \tag{12}$$

Data association is performed on the ith observation in the current frame at t and the jth object in the previous frame at $t-1$, and each association has d_{SS} adjusted by $d_D(r)$ and d_{MLF} filtered by d_{CP}. Each component has a value in a range of [0,1] and $d(\mathbf{O}_t^i, \mathbf{O}_{t-1}^j)$

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

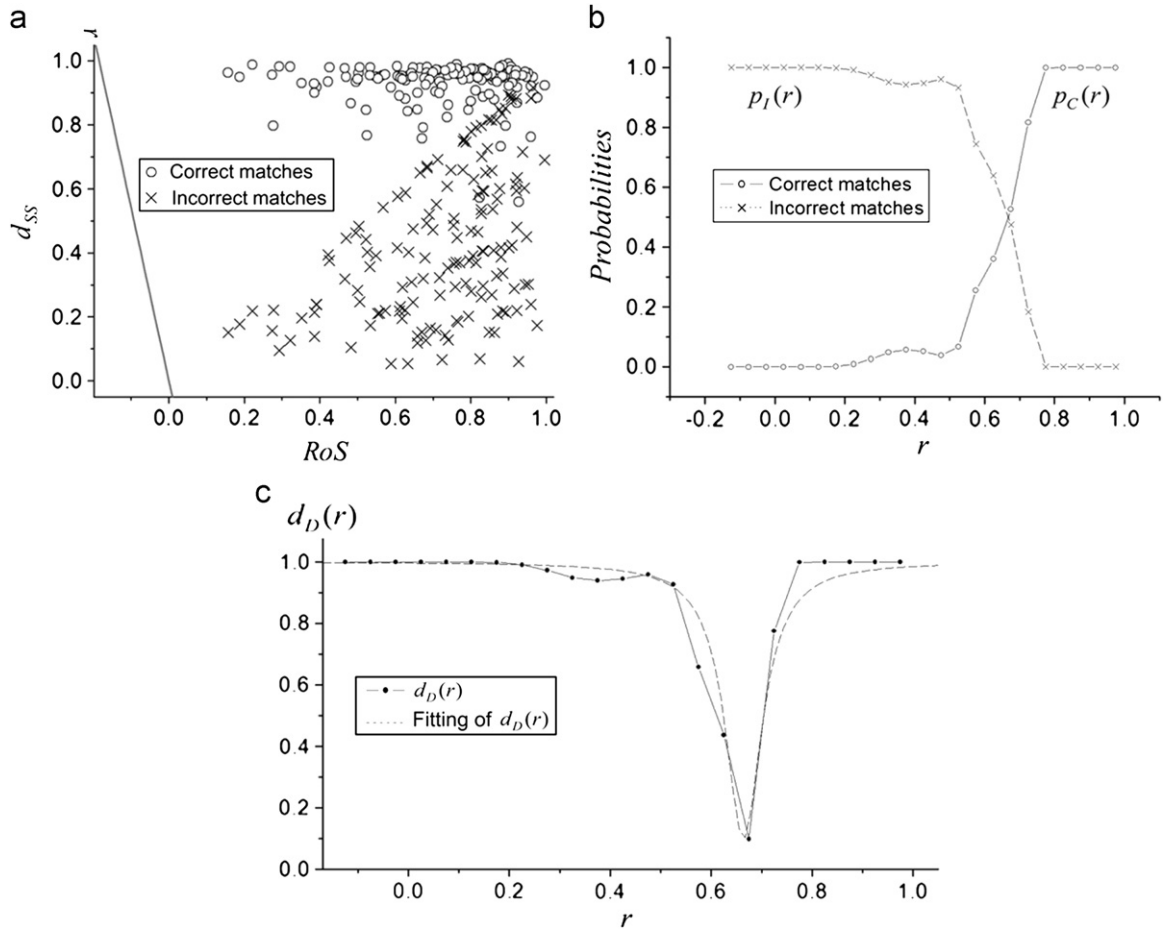


Fig. 9. Modeling of the distinguishability using training samples: (a) Plot of training samples, (b) probability distribution of the correct and incorrect matches, (c) the proposed distinguishability, $d_D(r)$.

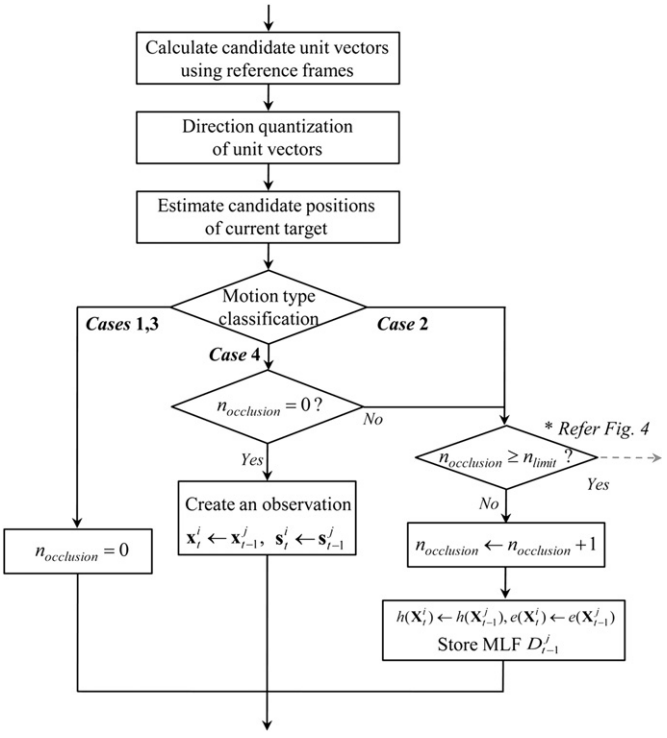


Fig. 10. Procedure of the motion modeling and occlusion handling.

also has a range of [0,1] due to the normalization constant $(1/(d_{CP}(x_t^i, y_t^i) + d_D(r)))$.

As explained in the beginning of Section 3, the overall tracking algorithm follows the procedure illustrated in Fig. 4. The motion modeling and occlusion handling part of Fig. 4 can be performed as illustrated in Fig. 10. Then, target IDs can be assigned using the proposed likelihood model for data association.

In Fig. 10, the first three blocks correspond to the MLF estimation process, which is followed by the motion type classification discussed in Section 3.3. For 'Case 1' and 'Case 3', the parameter $n_{occlusion}$, indicating the cumulative number of occlusion status, is set to zero and the data association is performed using the proposed likelihood model of Eq. (12). An object of 'Case 4' remains at the same position as in the previous frame, where the CP mask cannot find the change region even though the object is apparent in the current frame. Hence, an observation for the current frame is created at the same position in order for the standstill object to be correctly tracked. For an object of 'Case 2', which has no corresponding observation in the current frame and is considered as an occluded object, $n_{occlusion}$ is increased by 1 and the previous MLF is stored for the data association of the next frame. The normalized histogram $h(\mathbf{O}_t^i)$ and the normalized edgy energy $e(\mathbf{O}_t^i)$ are also copied from the object in the previous frame for processing of the next frame. If $n_{occlusion}$ exceeds a predefined threshold n_{limit} in 'Case 2' pathway, the observation at the next appearance is considered as a new object and a new ID is assigned, because successive MLFs without any observation can be unreliable. The paths of this case are marked with a dashed arrow in Figs. 4 and 10.

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

4. Experimental results

Experiments were performed to evaluate the proposed method. In the paper, s_h and s_e of Eq. (2) are set to 0.7 and 0.3, respectively, and the number of previous frames for the MLF estimation and n_{limit} are set to 3. For the experiments, video sequences with vehicles and pedestrians were used, where the challenging problems mentioned in Section 2 were present, including small targets with a cluttered background, targets with abrupt motion change, and low frame rate sequences provoking unstable MLF. Two public sequences in CAVIAR and PETS data sets were also used to demonstrate the consistent tracking performance of the proposed method. Fig. 11 shows the first frames of the test sequences, where the solid rectangles in the frames denote the selected regions. The tracking results of the selected regions are shown in Section 4.3.

4.1. The filtered MLF

Fig. 12 shows parts (335×90 pixels) of a video sequence which has a frame size of 640×480 pixels, containing small targets with cluttered background. The tracking results are shown in Fig. 12(a) with corresponding target ID numbers ('#1' and '#2'). The MLFs and the filtered MLFs for crossing targets are shown in Fig. 12(b) and (c), respectively, where the MLFs of the targets are marked with the corresponding numbers. '#1' target abruptly changes its direction, where the MLF of '#1' has the maximum value marked with an arrow, as shown in the third row. Since the

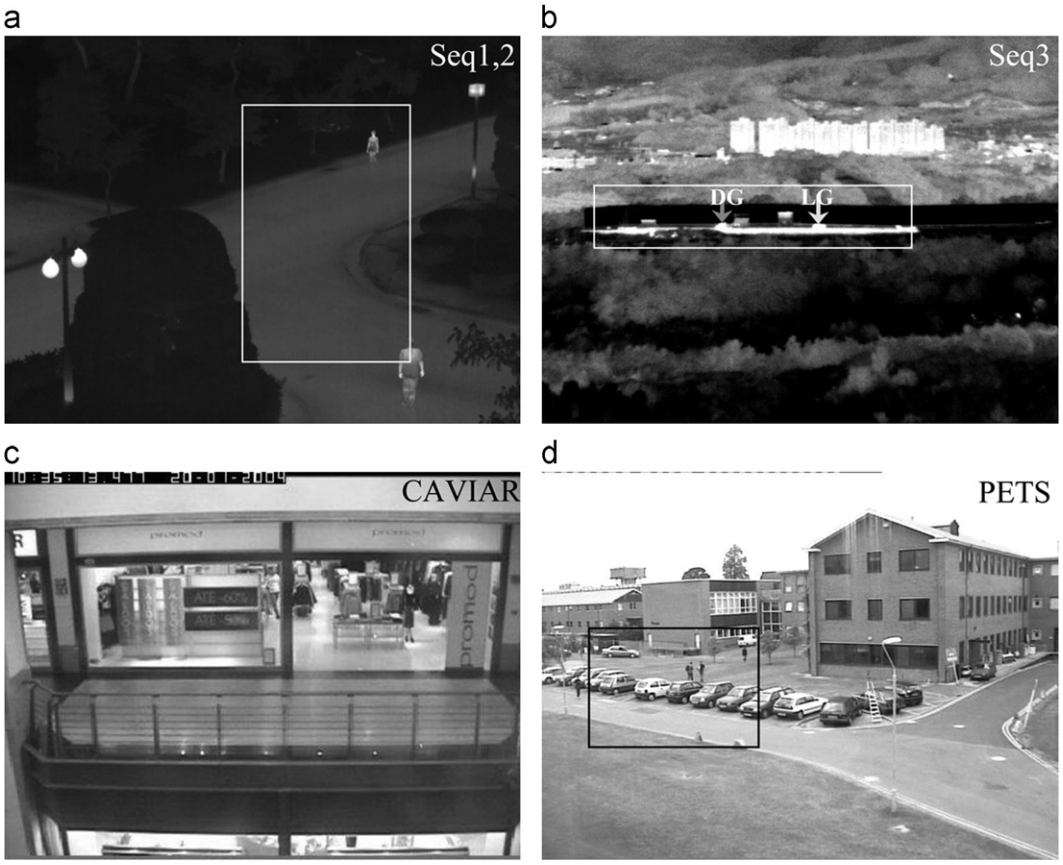


Fig. 11. First frames of the test sequences, where the solid rectangles in the frames denote the selected regions: (a) Seq. 1 and Seq. 2, (b) Seq. 3 where the initial positions of vehicle 'DG' and vehicle 'LG' are marked with a dark gray arrow and a light gray arrow, respectively, (c) CAVIAR sequence, and (d) PETS sequence.

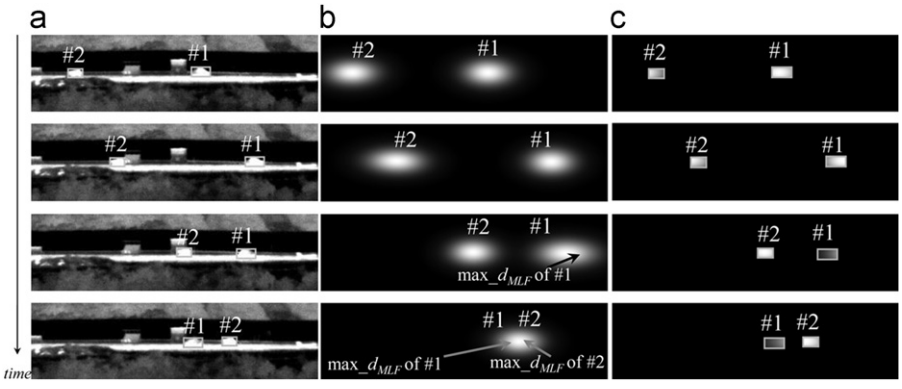


Fig. 12. Tracking results of crossing targets from the MLF and filtered MLF: (a) Tracking results, (b) MLF, (c) Filtered MLF.

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

targets are featureless, it is difficult to distinguish the target from the background region. However, the filtered MLFs successfully eliminate the clutter and help tracking the targets well at each frame. Moreover, the targets do not miss the corresponding IDs for the crossing objects.

4.2. Distinguishability

Fig. 13 shows data association examples in real video sequences. The second and third rows denote objects at time $t-1$ and observations at time t, respectively. The dashed lines present the data association in columns (a)–(g), and their (*RoS*, d_{SS}), r, and $d_D(r)$ values are shown in the third, fourth, and fifth rows, respectively. When objects at $t-1$ are visually distinguishable, low *RoS* and high d_{SS} values are assigned for correct matches ((a), (b)) and both *RoS* and d_{SS} are low for incorrect matches ((c)). When there are two rival objects and we consider the third object at $t-1$ in (d), high *RoS* and low d_{SS} values are assigned, thereby resulting in high $d_D(r)$. Therefore, if the objects at $t-1$ are distinguishable from each other, the likelihood model has high value of $d_D(r)$, so that it can be strongly affected by visual similarity. For rival objects in (e), high *RoS* and d_{SS} values result in low $d_D(r)$, which means that d_{MLF} becomes more dominant in the likelihood model. To investigate the robustness of the distinguishability, image patches which were not used in the training, were tested as shown in (f) and (g). (*RoS*, d_{SS}) of distinguishable objects in CAVIAR sequence (f) and indistinguishable objects in PETS sequence (g) are similar to (b) and (e), respectively. From $d_D(r)$ values of (f) and (g), we can understand that non-training images also have consistent distinguishability to determine the contribution level of the visual features.

4.3. Tracking results

The tracking performance was evaluated using Seq. 1 in Fig. 11(a) which had a frame size of 640×480 pixels. Two pedestrians in Seq. 1 were large enough and had stable motion, resulting in a successful tracking outcome during all frames. The proposed method was compared with three previous tracking methods, such as the particle filter [6] which adopted a scale space model to estimate the similarity between the background and the object; WLMCE [10] which dealt with abrupt motion using Wang Landau Monte Carlo Estimation (WLMCE); and TLD method [18] which adopted P–N learning for training a binary classifier. Fig. 14(a) and (b) show the trajectories of the two pedestrians, which indicate center positions of the detected

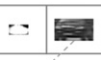
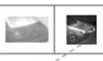
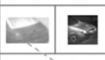
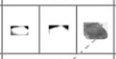
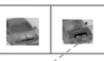
	a	b	c	d	e	f	g
Obj. in f_{t-1}							
Obs. in f_t							
(RoS, d_{SS})	(0.16, 0.9)	(0.47, 0.92)	(0.47, 0.43)	(0.91, 0.14)	(0.91, 0.87)	(0.43, 0.9)	(0.92, 0.82)
r	0.856	0.819	0.338	-0.028	0.690	0.811	0.641
$d_D(r)$	0.955	0.933	0.984	0.998	0.331	0.926	0.295

Fig. 13. Examples of the distinguishability values, $d_D(r)$, for various *RoS* and d_{SS}.

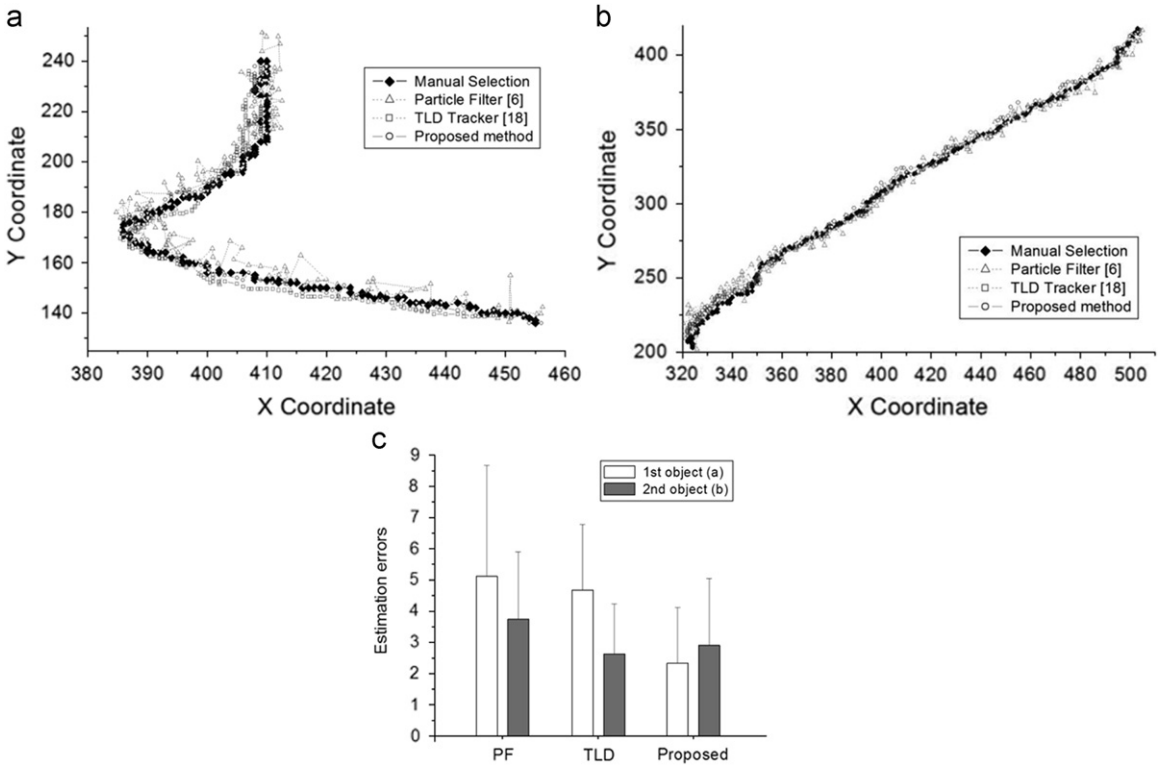


Fig. 14. Tracking results from the proposed method and previous methods: (a) Trajectory of the first objects, (b) Trajectory of the second objects, (c) Tracking performance comparison using an error analysis.

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

objects. Since the WLMCE deals with abrupt motion by providing a way to escape from local maxima, frequent ID switching of the two similar pedestrians were observed. Thus, the trajectory comparison of the WLMCE for Seq. 1 was excluded in this paper. To evaluate the accuracy of the resultant trajectories, the estimation errors were obtained with respect to the ground truth as follows,

$$e = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_{est}^{i}-x_{GT}^{i})^2+(y_{est}^{i}-y_{GT}^{i})^2}, \tag{13}$$

where N is the frame number, $(x_{est}^{i},y_{est}^{i})$ is the estimated center position, and (x_{GT}^{i},y_{GT}^{i}) is the ground truth center position that was manually selected. Although all trajectories estimated by the proposed and previous methods are similar to the ground truth trajectory, the proposed method yields relatively small estimation errors in both the means and standard deviations as shown in Fig. 14(c).

The second test sequence of Seq. 2 has a low frame rate sequence to evaluate the effects of rapid motion, which was obtained by downsampling Seq. 1 in time domain with a sampling interval of 30 frames. Fig. 15 shows the selected region of Seq. 2 with the tracking results from the proposed and previous methods. Since similar pedestrians had relatively rapid motion and passed by in frame 6, the particle filter and the WLMCE yield incorrect ID assignments. After frame 6, the two pedestrians were tracked with the switched IDs as shown in the first and second rows. As shown in the third and fourth rows, on the other hand, the TLD and the proposed method successfully tracked the pedestrians because they utilized the P–N learning and the distinguishability in deciding IDs for rival candidates, respectively.

In addition, the tracking methods were applied to small object sequence of Seq. 3 in Fig. 11(b), which had a frame size of 640×480 pixels and contained small vehicles of less than 15×10 pixels. In Fig. 11(b), the initial positions of vehicle 'DG' and vehicle 'LG' were marked with a dark gray arrow and a light gray arrow, respectively. Fig. 16 shows the selected regions of the tracking results from the proposed and previous methods, while the arrows denote the true position of each vehicle. The first column shows the tracking results from the particle filter. When 'LG' abruptly decelerated and changed its motion direction in frames 9 and 10, its position was misestimated at the cluttered background, as the light gray ellipse illustrates. The missed ID of 'LG' could not be recovered until the end of the sequence. In addition, the position of 'LG', which lost its ID at frame 9, was misestimated as the current position of 'DG' in frame 17, while 'DG' was lost. The second and third columns show the tracking results of a single vehicle 'LG' by using the WLMCE and the TLD, respectively. Although the WLMCE encouraged sampling of less-visited regions in the state space to deal with abrupt motion, incorrect estimation of the position such as the position of 'DG' or cluttered background were frequently selected due to insufficient visual features of the small object. Although the P–N constraints of the TLD was effective for the rival objects of Seq. 2, the TLD also suffered from insufficient visual features, resulting in detection failure and ID switching as indicated by the circles. The tracking results from the proposed method are indicated by the rectangles in the fourth column. Since the filtered MLF dealt with the abrupt motion and the distinguishability adjusted the level of visual similarity for ID assignments of similar objects, the proposed method produced stable and correct tracking results for all frames.

The proposed method was also applied to '*EnterExitCrossingPaths1front*' sequence from CAVIAR dataset whose resolution is 384×288 and '*PETSdata4_training_camera1*' sequence from PETS dataset whose resolution is 768×576. Fig. 17(a) shows the tracking results of CAVIAR sequence. Since 'Case 2' pathway of Fig. 10 dealt with the occlusion in frame 19, object #1 could be well tracked with the correct ID in frame 42. n_{limit} was set to 20 for CAVIAR sequence since CAVIAR sequence has small amount of

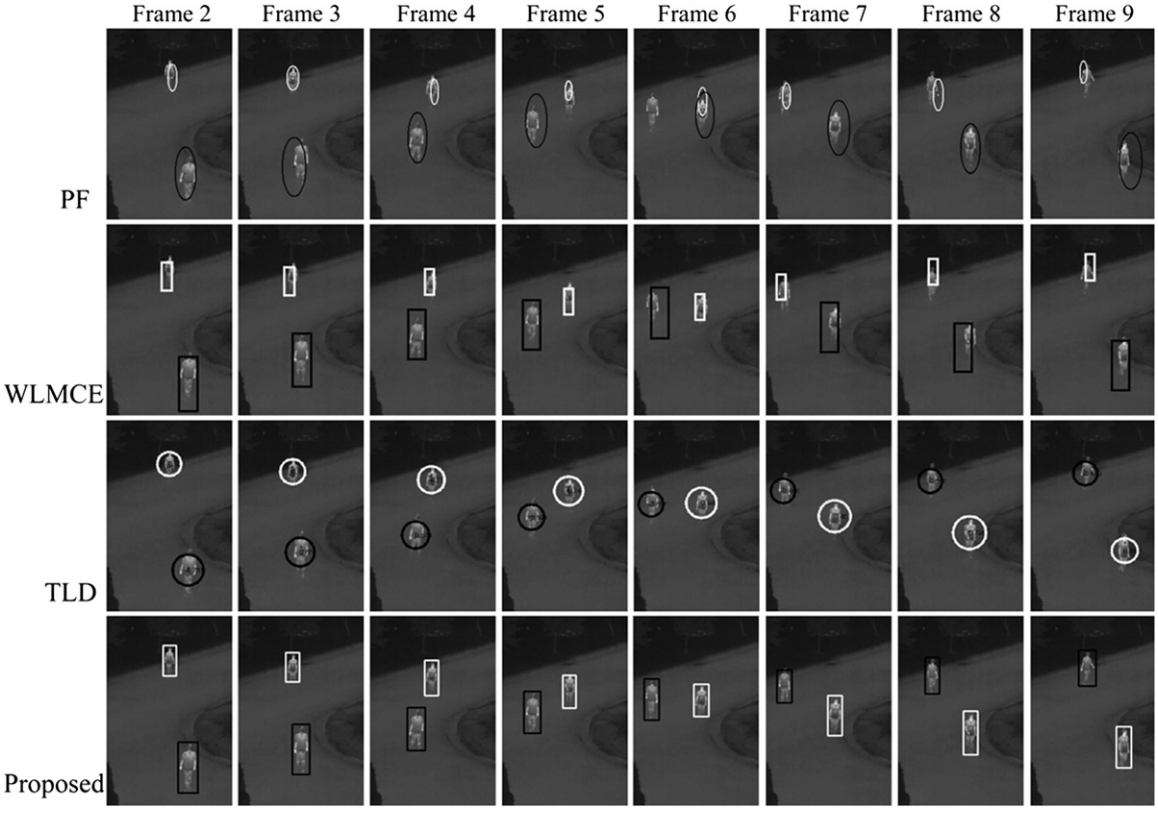


Fig. 15. Selected frames of Seq. 2, which show the tracking results from the proposed method and previous methods.

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

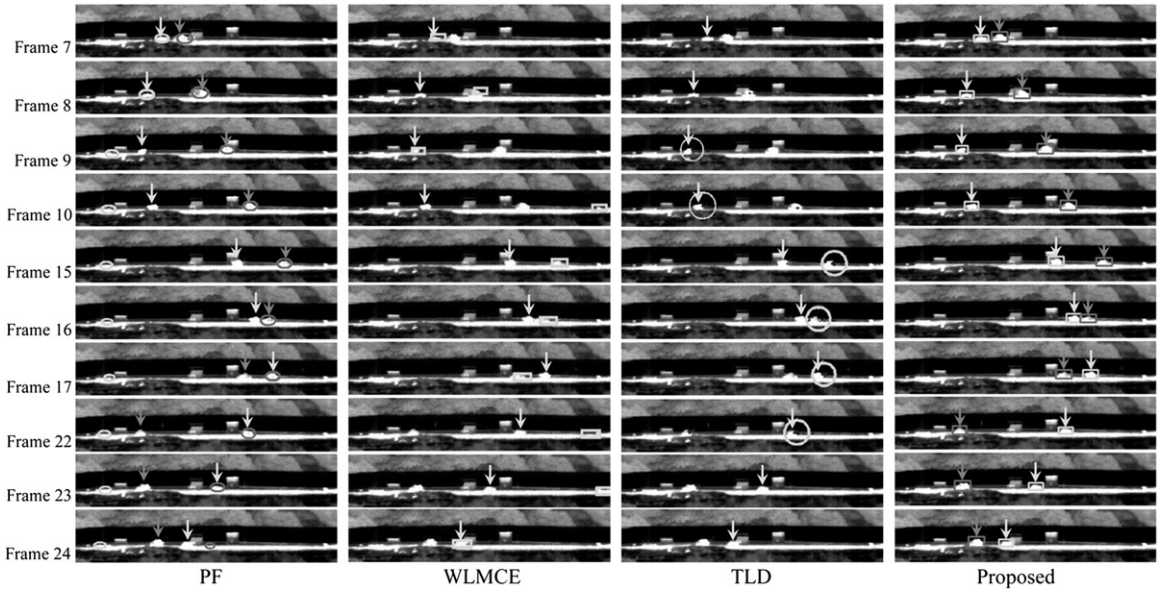


Fig. 16. Selected frames of Seq. 3, which show the tracking results from the proposed method and previous methods.

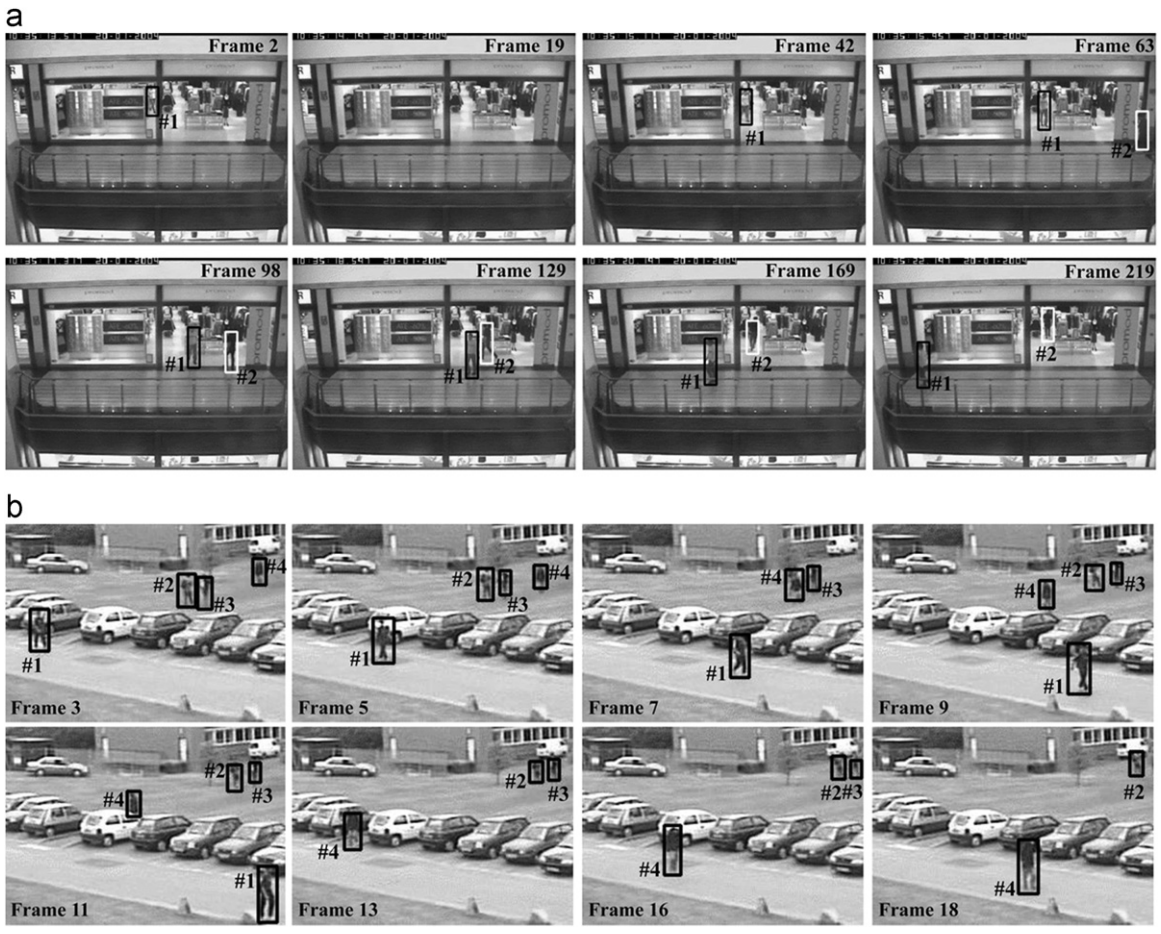


Fig. 17. Selected frames including the tracking results: (a) CAVIAR sequence, (b) PETS sequence.

motion between consecutive frames. Both object #1 and #2 (appeared in frame 64) were correctly tracked for all frames because they had stable motion and distinguishable visual features as partly shown in Fig. 13(f). Fig. 17(b) shows the selected region of PETS sequence (Fig. 11(d)) with the tracking results from the proposed method. PETS sequence which contained four similar objects was downsampled in the same way as Seq. 2 to simulate the effects of rapid motion. As shown in the experiments, the proposed method can successfully track several small objects that are occluded or crossed with each

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

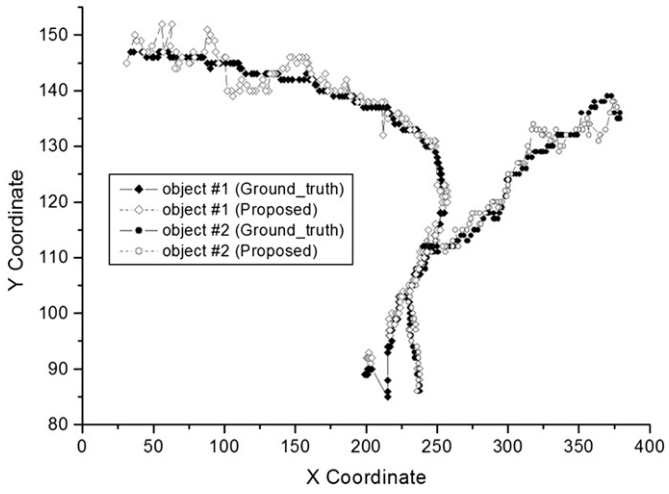


Fig. 18. Tracking results of CAVIAR sequence from the proposed method in comparison with the ground truth.

other. Fig. 18 shows the center positions of the tracked objects in CAVIAR sequence, where white diamonds and circles denote the trajectories determined by the proposed method and the black diamonds and circles denote the ground truths which are available at "http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/". The means and standard deviations of the estimation errors of Eq. (13) are 2.79 (2.39) and 1.48 (1.47) for object #1 (object #2), respectively, which denote the accuracy of the resultant trajectories.

The tracking process runs at 5.72–13.3 fps for various frame sizes of 768×576–384×288 on a 2.67 GHz Pentium Core processor with un-optimized C++ code. All experimental results of the video sequences are available at http://athena.kaist.ac.kr/MTT.html.

In this work, the size of the objects was assumed to be small because many conventional detection algorithms were developed to track big objects that generally had plenty of features representing their visual characteristics. Since features of small objects are less sufficient than those of big objects, humans get their attention to the change region that is highlighted by the frame difference between consecutive frames during tracking. Hence, information of the change region is used in the proposed method to help successful tracking of small objects.

5. Conclusion

We presented a robust tracking method for multiple targets using four spatial and temporal features based on the human object recognition process. In addition, the likelihood model for data association was also proposed to use the proposed four features more efficiently. Consequently, the proposed algorithm is more robust than the previous methods in deteriorated conditions such as tracking small targets with a cluttered background or with abrupt motion changes. We confirmed with a number of experiments that the proposed method provided consistent and robust tracking performance.

Conflict of interest

None

Acknowledgments

The authors would like to thank the Agency for Defense Development, Korea, for providing image sequences used in our experiments. This work was partly supported by the Agency for Defense Development under contract UD080032ID.

References

[1] Q. Yu, G. Medioni, Multiple-target tracking by spatiotemporal Monte Carlo Markov Chain Data Association, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (12) (2009) 2196–2220.
[2] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Transactions on Signal Processing 50 (2) (2002) 174–188.
[3] J. Czyza, B. Risticb, B. Macq, A particle filter for joint detection and tracking of color objects, Image and Vision Computing 25 (8) (2007) 1271–1281.
[4] C. Shan, T. Tan, Y. Wei, Real-time hand tracking using a mean shift embedded particle filter, Pattern Recognition 40 (7) (2007) 1958–1970.
[5] B. Zhang, W. Tian, Z. Jin, Robust appearance-guided particle filter for object tracking with occlusion analysis, International Journal of Electronics and Communications 62 (1) (2008) 24–32.
[6] P. Heo, S. Park, S. Jin, B. Yeou, H. Park, An object tracking method using particle filter and scale space model, IEEE International Conference on Image Processing, 2009, pp. 4081–4084.
[7] Z. Khan, T. Balch, F. Dellaert, An MCMC-based particle filter for tracking multiple interacting targets, in: Proceedings. of the European Conference on Computer Vision, 2004, pp. 279–290.
[8] P. Kumar, M.J. Brooks, A. Dick, Adaptive multiple object tracking using colour and segmentation cues, in: Proceedings of Asian Conference on Computer Vision, 2007, pp. 853–863.
[9] K. Smith, D. Gatica-Perez, J.-M. Odobez, Using particles to track varying numbers of interacting people, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 962–969.
[10] J. Kwon, K. Lee, Tracking of abrupt motion using Wang-Landau Monte Carlo estimation, in: Proceedings. of the European Conference on Computer Vision, 2008, pp. 387–400.
[11] Y. Huang, J. Llach, Tracking the small object through clutter with adaptive particle filter, in: International Conference on Audio, Language and Image Processing, 2008, pp. 357–362.
[12] B. Zou, X. Peng, L. Han, Particle filter with multiple motion models for object tracking in diving video sequences, 2008 Congress on Image and Signal Processing (2008) 224–228.
[13] H. Wang, D. Suter, A consensus-based method for tracking: modeling background scenario and foreground appearance, Pattern Recognition 40 (3) (2007) 1091–1105.
[14] D. Tsai, S. Lai, Independent component analysis-based background subtraction for indoor surveillance, IEEE Transactions on Image Processing 18 (1) (2009) 158–167.
[15] T. Celik, H. Kusetogullari, Solar-powered automated road surveillance system for speed violation detection, IEEE Transactions on Industrial Electronics 57 (9) (2010) 3216–3227.
[16] B. Wu, R. Nevatia, Detection and tracking of multiple, partially occluded humans by Bayesian combination of Edgelet based part detectors, International Journal of Computer Vision 75 (2) (2007) 247–266.
[17] S. Saravanakumar, A. Vadivel, C.G. Saneem Ahmed, Multiple human object tracking using background subtraction and shadow removal techniques, 2010 International Conference on Signal and Image Processing (ICSIP), 2010, pp.79–84.
[18] Z. Kalal, J. Matas, K. Mikolajczyk, P–N learning: Bootstrapping binary classifiers by structural constraints, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 49–56.
[19] Z. Kalal, K. Mikolajczyk, J. Matas, Face-TLD: Tracking-Learning-Detection applied to faces, IEEE International Conference on Image Processing, 2010, pp. 3789–3792.
[20] Q. Yu, T.B. Dinh, G. Medioni, Online tracking and reacquisition using co-trained generative and discriminative trackers, in: Proceedings. of the European Conference on Computer Vision, 2008, pp. 678–691.
[21] D.-J. Yi., N.B. Turk-Browne, J.I. Flombaum, M.-S. Kim, B.J. Scholl, M.M. Chun, Spatiotemporal object continuity in human ventral visual cortex, Proceedings of the National Academy of Sciences 105 (26) (2008) 8840–8845.
[22] A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Journal of Computing Surveys 38 (4) (2006).
[23] X. Zhang, M.L. King, R.J. Hyndman, A Bayesian approach to bandwidth selection for multivariate kernel density estimation, Computational Statistics and Data Analysis 50 (11) (2006) 3009–3031.
[24] A.K. Jain, Fundamentals of Digital Image Processing, Prentice Hall, NJ, 1989.
[25] N. Otsu, A. Threshold, Selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979).
[26] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., John Wiley & Sons, USA, 2001.
[27] 〈http://mathworld.wolfram.com/LorentzianFunction.html〉.

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005

Yeol-Min Seong received the B.S. degree in electrical engineering and computer science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea in 2005 and the Ph.D. degree in electrical engineering from KAIST in 2012. Now he is a senior engineer of SAMSUNG Electronics. Co., Ltd., Suwon, Korea. His research interests include image processing, pattern recognition, and object tracking.

HyunWook Park is a professor of the department of electrical engineering at KAIST. He received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea in 1981 and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Seoul, Korea in 1983 and 1988, respectively. He has been a professor of electrical engineering department since 1993 and an adjunct professor of biosystem department since 2003, KAIST, Korea. He was a research associate at the University of Washington from 1989 to 1992 and was a senior executive researcher at the Samsung Electronics Co. Ltd., from 1992 to 1993. He is a senior member of IEEE and a member of SPIE. He has served as Associate Editor for International Journal of Imaging Systems and Technology. His current research interests include image computing system, image compression, medical imaging, and multimedia system.

Please cite this article as: Y.-M. Seong, H. Park, Multiple target tracking using cognitive data association of spatiotemporal prediction and visual similarity, Pattern Recognition (2012), doi:10.1016/j.patcog.2012.03.005