Pattern Recognition 46 (2013) 2187–2201

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Visual tracking by proto-objects

Zhidong Li [a,b], Weihong Wang [a,b], Yang Wang [a,b,*], Fang Chen [a,b], Yi Wang [a,b]

[a] National ICT Australia[1], Level 5, 13 Garden Street, Eveleigh NSW 2015, Australia
[b] The University of New South Wales, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 10 April 2012
Received in revised form
31 October 2012
Accepted 14 January 2013
Available online 26 January 2013

Keywords:
Tracking
Proto-object
Saliency
Gibbs sampling
Bayesian

ABSTRACT

In this paper, we propose a biologically inspired framework of visual tracking based on proto-objects. Given an image sequence, proto-objects are first detected by combining saliency map and topic model. Then the target is tracked based on spatial and saliency information of the proto-objects. In the proposed Bayesian approach, states of the target and proto-objects are jointly estimated over time. Gibbs sampling has been used to optimize the estimation during the tracking process. The proposed method robustly handles occlusion, distraction, and illumination change in the experiments. Experimental results also demonstrate that the proposed method outperforms the state-of-the-art methods in challenging tracking tasks.

1. Introduction

Visual tracking, as a fundamental step to explore videos, is important in many computer vision based applications, such as security and surveillance, video compression, and robotic vision systems. During the tracking process, state of the target is estimated over time by associating its representation in the current frame with those in previous frames. Although research on visual tracking has lasted for decades, various noisy factors including background clutter, occlusion, distraction, and illumination variance still cause problems in practice, which makes visual tracking a challenging task [1,2]. Most of the problems will cause unpredictable appearance changes during tracking, especially when the target is a non-rigid object.

In order to improve the robustness of visual tracking, the method of tracking by detection has been proposed by translating visual tracking into classification task [3–5]. Besides, a number of tracking algorithms employ object model updating to deal with appearance changes, such as the incremental learning based target updating in [6] and the on-line sparse principal component analysis in [7]. However, it can be observed that appearance based methods, including those with on-line model updating, may cause drift or even loss of target under relatively complex conditions such as large pose variation and sudden illumination change.

An alternative method is saliency based tracking, which is inspired by biological vision systems [8–10]. Saliency represents the extent of visual attention paid to a place in the scene according to its standing out to surrounding [11]. For saliency detection, a saliency map is built to represent the saliency values of all the regions in the given image [12]. Usually the target is more likely to appear in the places with higher saliency values. Different kinds of methods have been proposed to compute saliency values of an image. Most approaches of saliency detection are based on low-level features [13–17], which focus on detecting salient regions in an image by bottom-up mechanism. In these approaches, different regions are ranked and selected according to the contrast of low-level features between each individual region and its surrounding. Bottom-up saliency detection can be easily applied to various scenarios. However, without task related knowledge, salient regions detected by such methods might be insufficient to comprehensively describe the target or separate the target from the background, as sometimes background areas may also have high contrast with low-level features. Also, some salient regions such as corners and edges could appear and then disappear in a short time when illumination condition or target appearance changes drastically. Therefore, during the tracking process the detected salient regions may lack consistency over time. Other approaches of saliency detection make use of object-level representation through the top-down mechanism [18–20]. That is, with explicit conceptual

* Corresponding author at: National ICT Australia, Level 5, 13 Garden Street, Eveleigh NSW 2015, Australia. Tel.: +61 2 9376 2200.
E-mail addresses: zhidong.li@nicta.com.au (Z. Li), weihong.wang@nicta.com.au (W. Wang), yang.wang@nicta.com.au (Y. Wang), fang.chen@nicta.com.au (F. Chen), yi.wang@nicta.com.au (Y. Wang).

[1] National ICT Australia (NICTA) is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Center of Excellence program.

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.020

knowledge about the target, the saliency value over a region measures its chance of belonging to an object, so that the target can be separated from the background based on the object-level saliency. However, such approach is inevitably limited to detecting an already learned target, for example, saliency based human detection in [19]. Compared to bottom-up methods, detecting targets by object-level representation during tracking is usually inconvenient due to the difficulties of obtaining the target conceptual knowledge.

It has been observed that, even when the conceptual knowledge about actual objects is not explicitly presented, humans can still percept a region that is plausible to be an object. Representation level for such kind of plausible objects is between the low-level feature based representation and the object-level semantic representation [21]. In psychology, it can be explained by the coherency theory by Rensink [22] that an amount of attentional regions (regions of the plausible objects) can be selected and bundled together into actual objects. These regions, referred to as proto-objects, are volatile units of visual information that can be accessed by selective attention and subsequently validated as actual objects [22]. Compared to low-level feature based salient regions, proto-objects are determined by simultaneously considering the top-down mechanism and the bottom-up mechanism [21,23,24], so that their correlations to the target are much more stable during the tracking process. In addition, detection of the proto-objects does not require conceptual knowledge about the target.

In this paper, we introduce a framework of visual tracking based on proto-objects. Given an image sequence, proto-objects around the target region are first detected by combining bottom-up saliency and top-down topic model. The target is then tracked over time based on spatial and saliency information of the proto-objects. Sampling based optimization algorithm is utilized to infer the states of both the target and the proto-objects during the tracking process. Experimental results demonstrate that the proposed approach outperforms the state of the art, and it robustly handles occlusion, distraction, as well as illumination variation. To authors' knowledge, the approach of visual tracking by proto-objects is first introduced in this paper.

The paper is organized as follows. Section 2 reviews the related work. Section 3 presents our method for proto-object detection. Section 4 introduces the framework of tracking by proto-objects and the optimization algorithm. Experimental results are discussed in Section 5. The conclusion is drawn in Section 6.

2. Related work

Several saliency based approaches have been proposed for visual tracking. Mahadevan and Vasconcelos proposed salient feature based tracking in [8]. The features for target tracking are selected based on its saliency values, which are estimated according to the power of discriminating the target from its surrounding areas [25]. The technique has also been extended to detect salient regions in spatiotemporal space [26]. In addition, the framework of tracking by salient regions has been proposed in [9,10]. In their work salient regions are first detected and then tracked in video frames based on low-level features using the bottom-up mechanism. Target location is determined through the weighted combination of all available salient regions. Tracking by salient regions often relies on the co-occurrence of a large number of salient regions [10], which is computationally expensive; besides, the target must be large enough [9]. Tracking by proto-objects generally requires less number of proto-objects and it works for small target as well, since the whole target can be detected as salient proto-object.

Yang et al. has proposed an approach of context-aware visual tracking that uses a set of auxiliary objects to support the target tracking [27]. The auxiliary objects are tracked collaboratively with the target during the tracking process. Auxiliary object refers to image region that exhibits persistent co-occurrence and consistent motion with the target. Compared with auxiliary object, proto-object is a biologically inspired concept, and saliency information is used to detect proto-objects in this work. Moreover, in practice most auxiliary objects are detected outside the target region, while proto-objects are usually detected within the target region. Hence the two methods overall utilize different sources of context information, and they could be complementary to each other during the tracking process.

The problem of proto-object detection has also attracted researchers from both psychology and computer vision societies. Existing work on proto-object detection includes the biologically plausible visual attention models [21,23]. For computational implementation, to know how likely a region is a proto-object, the saliency value of the region is determined by the combination of both bottom-up and top-down mechanisms [24]. As an example, proto-objects are detected using the feed-forward and feedback links in [23]. In addition, some work in the field of psychology carried out cognitive experiments to discover various phenomena about proto-objects, such as [22].

3. Salient proto-object detection

Given a video sequence, our method first detects proto-objects in the initial frame based on saliency detection and topic model. For visual tracking, the initialized target region (usually represented by a rectangle bounding box) contains certain knowledge about the target although it is far from comprehensively representing the target. Such knowledge is represented by a top-down topic model in this work. Considering various noisy factors such as appearance or illumination change during the tracking process, the impact of top-down knowledge sometimes becomes weak and implicit. Hence for the robustness of proto-object detection, it is helpful to combine the top-down topic model with the bottom-up saliency.

First, a saliency map is built for the target region and surrounding areas based on low-level features. The low-level features we use consist of color, intensity, and orientation, which are the same as those features used in [12]. For each low-level feature f, a feature map in scale c can be obtained for the image. Then a saliency map is calculated based on the spectral phase information of the feature map [14]. The scale parameter c is determined by the input image size [13], for which we use 32×32, 64×64, and 128×128 in this work. The combination of multi-scale saliency maps helps discover more proto-objects and produces better object boundaries [28]. Hence saliency maps with different features and scales are summed up to form a bottom-up saliency map. Fig. 1b shows an example of saliency map obtained using our method.

Meanwhile, the image is segmented into a set of partitions, and topic model is used to group the partitions with high saliency values to form proto-objects. To ensure that each partition only belongs to one proto-object, we employ an over-segmentation algorithm proposed in [29] (see Fig. 1c for an example). Here we use visual words [30] to represent a partition and PLSA model [31] to group the partitions with consistent representation. The visual words are composed by clustering low-level feature vectors using K-means. The PLSA model assumes that there exists a set of topics for both salient and non-salient regions. An image partition can be assigned to either a salient region topic or a non-salient region topic, while each topic is characterized by certain

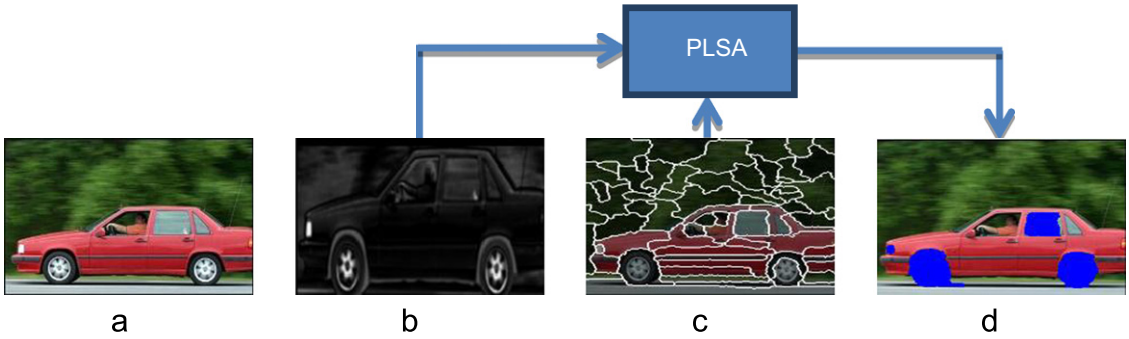

Fig. 1. Proto-object detection. (a) An example image. (b) The bottom-up saliency map. (c) The over-segmentation result. (d) The detected proto-objects.

Fig. 2. Examples of proto-objects detected in the target region.

distribution of the visual words. The EM algorithm is used to reveal the topics and optimize the topic assignment through maximum log-likelihood. The image partitions with the same salient region topic are then grouped as a proto-object. More details could be found in [24]. Fig. 2 gives some examples of detected proto-objects from the video sequences used in the experiment section. For computational simplicity, we use rectangles to represent proto-objects during the tracking.

4. Tracking by proto-objects

4.1. Model representation

Given a video sequence, the target state x^t at each time instant t can be represented by a set of elements including its horizontal position and vertical position. The observation at time t is represented as $z^t=\{z_r^t, z_s^t\}$, which includes the proto-object measurement z_r^t and the saliency map z_s^t. We decompose the proto-object measurement into $z_r^t=\{z_{r,1}^t, z_{r,2}^t \ldots z_{r,M}^t\}$, where M is the number of proto-objects, and $z_{r,i}^t$ is the measurement of the i-th proto-object. We use y^t to represent the hidden state for proto-objects. It consists of spatial and saliency information for each proto-object, which can be written as $y^t=\{r^t,s^t\}$. Here $r^t=\{r_1^t,r_2^t \ldots r_M^t\}$ represents the spatial state of individual proto-objects, and $s^t=\{s_1^t,s_2^t \ldots s_M^t\}$ represents the saliency state of the proto-objects. $s_i^t=1$ if the i-th proto-object becomes salient, and $s_i^t=0$ otherwise. Given observations up to current time t, the MAP (maximum a posterior) state estimation of the target and proto-objects becomes:

$$(\hat{x}^t,\hat{y}^t)=\underset{(x^t,y^t)}{\operatorname{argmax}}\, p(x^t,y^t|z^{1:t}) \tag{1}$$

Here the joint posterior probability $p(x^t, y^t|z^{1:t})$ at time t is factorized as

$$p(x^t, y^t|z^{1:t})=p(x^t|y^t)p(y^t|z^{1:t}) \tag{2}$$

For the first term on the right, the target state x^t can be estimated from those of the proto-objects. The state of each proto-object contains its spatial and saliency information. For human vision system, higher saliency indicates stronger stimulation from the corresponding proto-object to the tracker. Thus salient proto-objects tend to have higher impact on the tracking process than non-salient ones. In this work, salient proto-objects are used to support the target tracking though generalized Hough Transform [32]. Given the spatial position of salient proto-objects, the target state can be voted probabilistically. The conditional probability $p(x^t|y^t)$ of target state is calculated as:

$$p(x^t|y^t)=p(x^t|r^t,s^t)=\frac{1}{\sum_i s_i^t}\sum_i s_i^t p(x^t|r_i^t) \tag{3}$$

The probabilistic vote $p(x^t|r_i^t)$ describes the distribution of possible target location with regard to the i-th proto-object, which is represented by a Gaussian distribution $N(x^t; r_i^t+\mu_i^t,(\sigma_i^t)^2)$ learned from previous frames. For the Gaussian distribution, μ_i^t is the mean distance from the center of the i-th proto-object to the target, and σ_i^t is the standard deviation. It can be known that the probabilistic vote takes effect only when the corresponding proto-object becomes salient ($s_i^t=1$).

For the second term in (2), the posterior probability of proto-objects could be formulated as:

$$p(y^t|z^{1:t})=p(r^t,s^t|z^{1:t})=p(s^t|r^t,z^t)p(r^t|z^{1:t}) \tag{4}$$

Conditional independence among proto-objects is assumed to simplify the computation. The posterior distribution of proto-object position is factorized as:

$$p(r^t|z^{1:t})=\prod_i p(r_i^t|z_{r,i}^{1:t}) \tag{5}$$

Using Bayesian rule, for each proto-object the term $p(r_i^t|z_{r,i}^{1:t})$ is computed as:

$$p(r_i^t|z_{r,i}^{1:t})\propto p(z_{r,i}^t|r_i^t)\int p(r_i^t|r_i^{t-1})p(r_i^{t-1}|z_{r,i}^{1:t-1})dr_i^{t-1} \tag{6}$$

It can be seen that the posterior $p(r_i^t|z_{r,i}^{1:t})$ could be iteratively estimated through commonly used Bayesian filtering framework, such as Kalman filter in this work. Thus for each proto-object, the posterior at time t is represented by a Gaussian distribution $N(r_i^t;\mu_{r,i}^t,(\sigma_{r,i}^t)^2)$, where $\mu_{r,i}^t$ and $\sigma_{r,i}^t$ are the mean position and the corresponding standard deviation respectively. The conditional probability of proto-object saliency at time t is also factorized as:

$$p(s^t|r^t,z^t)=\prod_i p(s_i^t|r_i^t,z_s^t), \tag{7}$$

where

$$p(s_i^t=1|r_i^t,z_s^t)=z_s^t(r_i^t) \tag{8}$$

Here $z_s^t(r_i^t)\in[0,1]$ is obtained from the normalized saliency map z_s^t at position r_i^t.

Combining the above, it can be known that:

$$\begin{aligned}p(x^t,r^t,s^t|z^{1:t}) &= \left[\frac{1}{\sum_i s_i^t}\sum_i s_i^t p(x^t|r_i^t)\right]\left[\prod_i p(s_i^t|r_i^t,z_s^t)p(r_i^t|z_{r,i}^{1:t})\right]\\ &= \left[\frac{1}{\sum_i s_i^t}\sum_i s_i^t N\left(x^t;r_i^t+\mu_i^t,(\sigma_i^t)^2\right)\right]\\ &\quad\times\left[\prod_i z_s^t(r_i^t)^{s_i^t}(1-z_s^t(r_i^t))^{1-s_i^t}N(r_i^t;\mu_{r,i}^t,(\sigma_{r,i}^t)^2)\right]\end{aligned} \tag{9}$$

4.2. Optimization algorithm

Sampling based technique is employed to optimize the states of the target and proto-objects during the tracking process. Since sampling on the joint posterior probability of (x^t,r^t,s^t) is difficult, we iteratively sample the values of x^t, r^t, and s^t by Gibbs Sampling [33]. As one typical class of MCMC algorithm, the Gibbs sampling approximates the unknown joint posterior distribution of multiple variables by iteratively computing the conditional distribution of these variables. At each time t, values of the variables are firstly initialized for the current frame. Then each variable will be iteratively inferred by sampling it from the conditional probability on other variables. Although Gibbs sampling can accept randomized initialization, to achieve better convergence rate, we initialize the proto-object positions $\{\hat{r}_i^t\}$ by Mean Shift algorithm [1]. Then at each position $\hat{r}_i^t$, the value of bottom-up saliency map is used to initialize $\hat{s}_i^t$. In this work, the conditional probabilities of x^t, r^t, and s^t must be estimated, which are shown in the following paragraphs.

The conditional distribution of target state x^t on the estimated proto-object saliency $\hat{s}^t$ and position $\hat{r}^t$, and observations $z^{1:t}$ can be used to sample x^t. As x^t is conditionally independent on $z^{1:t}$, according to (3) the conditional distribution can be written as:

$$p\left(x^t|,\hat{r}^t,\hat{s}^t\right)\propto\sum_i \hat{s}_i^t N\left(x^t;\hat{r}_i^t+\mu_i^t,(\sigma_i^t)^2\right) \tag{10}$$

Then r^t is sampled according to its conditional distribution given estimated $\hat{x}^t$ and $\hat{s}^t$.The conditional distribution can be written as:

$$p\left(r^t|z^{1:t},\hat{x}^t,\hat{s}^t\right)\propto p\left(\hat{x}^t|\hat{s}^t,r^t\right)p\left(\hat{s}^t,r^t|z^{1:t}\right) \tag{11}$$

Each r_i^t can be sampled from its conditional distribution:

$$\begin{aligned}p\left(r_i^t|\hat{r}_{-i}^t,z_i^{1:t},\hat{x}^t,\hat{s}^t\right)\propto&\left[\hat{s}_i^t N\left(\hat{x}^t;r_i^t+\mu_i^t,(\sigma_i^t)^2\right)+\sum_{j\neq i}\hat{s}_j^t N\left(\hat{x}^t;\hat{r}_j^t+\mu_j^t,(\sigma_j^t)^2\right)\right]\\ &z_s^t(r_i^t)^{\hat{s}_i^t}(1-z_s^t(r_i^t))^{1-\hat{s}_i^t}N(r_i^t;\mu_{r,i}^t,(\sigma_{r,i}^t)^2).\end{aligned} \tag{12}$$

Here we use r_{-i}^t to represent the states of all the proto-objects except the i-th proto-object. To simplify the computation, $z_s^t(r_i^t)$ is approximated by first-order Taylor expansion.

ALGORITHM 1. Tracking by proto-objects

INPUT
A video sequence for visual tracking;
The initial target state x^1 and detected proto-objects for the first frame;
OUTPUT
Target state x^t and proto-object state $y^t=(r^t,s^t)$ for each frame t;
WHILE (t<=T)
Obtain z_s^t around the $\hat{x}^{t-1}$ as bottom-up saliency map;
FOR EACH proto-object i
Obtain observation $z_{r,i}^t$ given $\hat{r}_i^{t-1}$;
Initialize $\hat{r}_i^t$ by Mean Shift algorithm;
Initialize $\hat{s}_i^t$ by $z_s^t(\hat{r}_i^t)$;
END FOR
$k=1$
WHILE ($k\leq Iteration\_Max$)
$\hat{x}^t\leftarrow$ sample x^t by (10);
FOR EACH proto-object i
$\hat{r}_i^t\leftarrow$ sample r_i^t by (12);
END FOR
$\hat{r}^t=\{\hat{r}_1^t,\ldots,\hat{r}_M^t\}$;
FOR EACH proto-object i
$\hat{s}_i^t\leftarrow$ sample s_i^t by (14);
END FOR
$\hat{s}^t=\{\hat{s}_1^t,\ldots,\hat{s}_M^t\}$;
k++;
END WHILE
Select optimal (x^t,r^t,s^t) from the obtained samples.
t++;
END WHILE

Fig. 3. The algorithm of visual tracking by proto-objects.

Then the saliency state s^t of proto-objects can be sampled according to its conditional distribution given estimated $\hat{x}^t$ and $\hat{r}^t$, which can be written as:

$$p\left(s^t|z^{1:t},\hat{x}^t,\hat{r}^t\right)\propto p\left(\hat{x}^t|\hat{r}^t,s^t\right)p\left(s^t|\hat{r}^t,z^t\right) \tag{13}$$

For each proto-object, s_i^t is sampled according to the following conditional probability:

$$\begin{aligned}p\left(s_i^t|\hat{s}_{-i}^t,z_i^{1:t},\hat{x}^t,\hat{r}^t\right)\propto&\frac{1}{s_i^t+\sum_{j\neq i}\hat{s}_j^t}\left[s_i^t N\left(\hat{x}^t;\hat{r}_i^t+\mu_i^t,(\sigma_i^t)^2\right)\right.\\ &\left.+\sum_{j\neq i}\hat{s}_j^t N(\hat{x}^t;\hat{r}_j^t+\mu_j^t,(\sigma_j^t)^2)\right]z_s^t(\hat{r}_i^t)^{s_i^t}(1-z_s^t(\hat{r}_i^t))^{1-s_i^t}\end{aligned} \tag{14}$$

Overall, the tracking process can be described by the algorithm shown in Fig. 3.

5. Experiments

To test the effectiveness of our tracking method, the Head Tracking Dataset [34] and CAVIAR dataset [35] have been used for evaluation in the experiments. Especially, tracking results for 12 changeling video sequences are shown in this work. Seven of them (named Head1–Head7) are from the Head Tracking Dataset. These videos contain various difficult situations including fast movement, occlusion to the target and deformation caused by head rotation. Browse, Fight, Reading, and Meeting are four videos from CAVIAR dataset. These sequences are captured with distorted camera view and contain abrupt movement and background distraction. Skating is an even challenging sequence which includes large pose variation, partial occlusion, and illumination change. Without code optimization in our Matlab implementation, the videos are processed about one frame per second with the resolution of 384 by 288 pixels; variance of the speed exists due to different number and size of proto-objects. The experiments are conducted on a Pentium 4 computer with 1 GB of RAM.

Fig. 4 shows the results of salient region [36] based tracking and proto-object based tracking for the video Skating. From Fig. 4a, it can be seen that most salient regions are gradually drifted or even lost during the tracking due to the variation of the actress' pose and illumination condition. Meanwhile, the

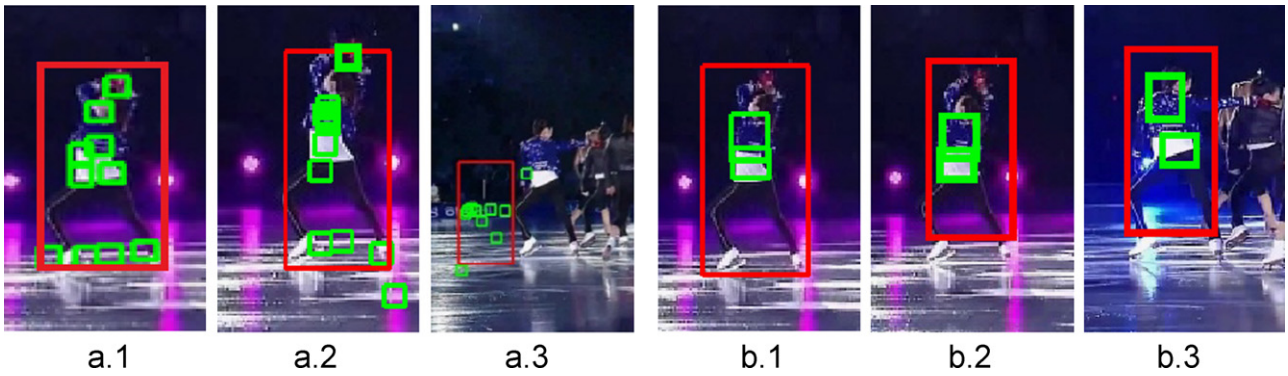

Fig. 4. (a) The salient region based tracking results, where the green rectangles represent the tracked salient regions. (b) The proto-object based tracking results, where the green rectangles represent the tracked salient proto-objects. The red rectangles represent corresponding tracking results in both (a) and (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

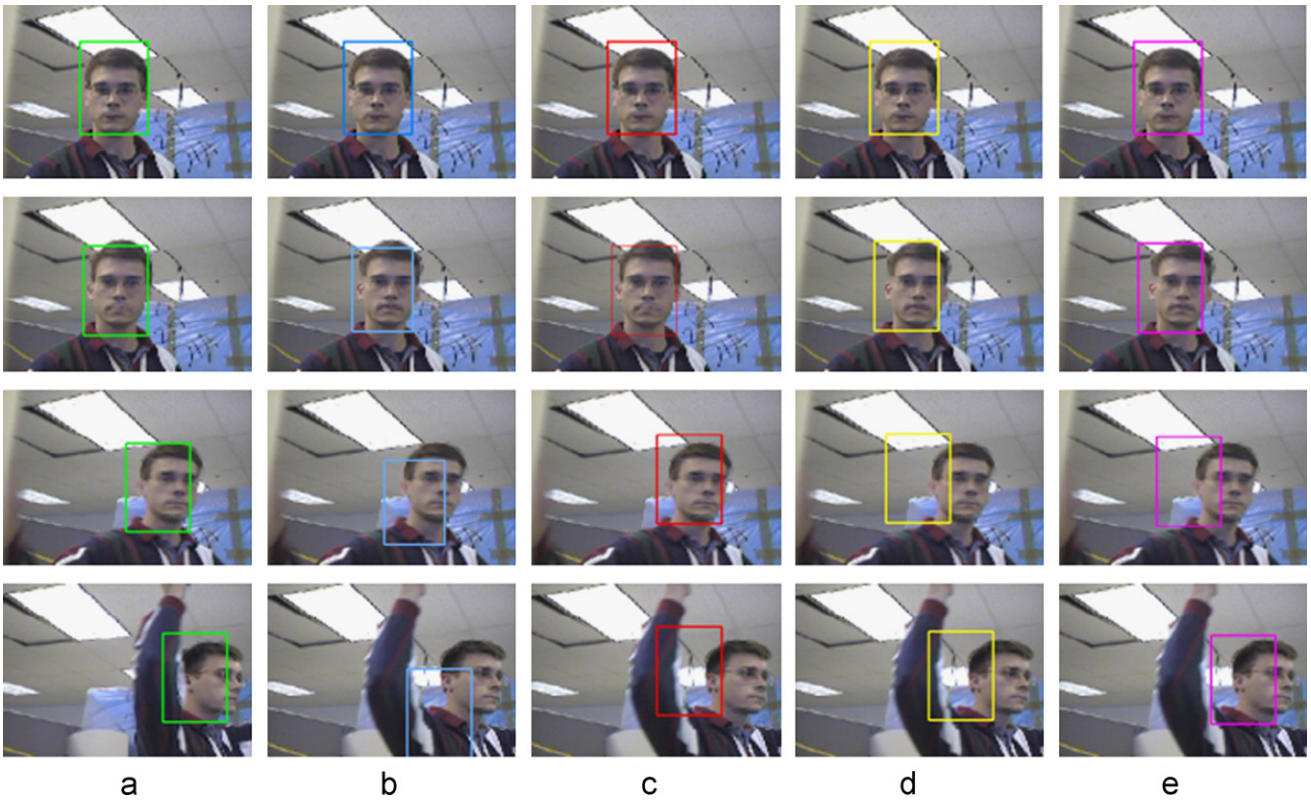

Fig. 5. Tracking results of video Head1 by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

proto-objects are detected by combining bottom-up saliency map and top-down topic model. In the video sequence, even though the target appearance deforms drastically, the salient proto-objects still exhibit high coherence over time (see Fig. 4b). In addition, usually the number of salient regions detected during tracking is higher than the number of detected proto-objects. Hence our method requires a smaller number of sub-trackers and still achieves satisfactory results. The experiment results demonstrate that tracking by proto-objects is more effective and stable than tracking by salient regions under complex visual environment.

The proposed tracking method has been tested on all the video sequences. To make the task even more challenging, the targets are tracked every three frames in all videos. We compare our method with four state-of-art methods. The first two methods are saliency based, using attentional region (AR) [9] and salient feature (SF) [8] respectively. The third method is multiple instance learning based tracking (MIL) [5], and the fourth method is tracking by decomposition (VTD) [7]. The authors' implementation of MIL, and VTD is available online.

Figs. 5 and 6 show the tracking results of different methods for video Head1 and Head2. In both video sequences the target's appearance changes drastically due to the rotation of the head. Compared to our method, the results of AR, SF, MIL and VTD have larger drifts when the target is translating and rotating simultaneously. Figs. 7 and 8 show the tracking results by all five methods for video Head3 and Head4. Both videos contain target occlusion and distraction. It can be seen that the proposed method and AR are more robust to occlusion than the other methods. In Fig. 8, the hands act as distractor as the

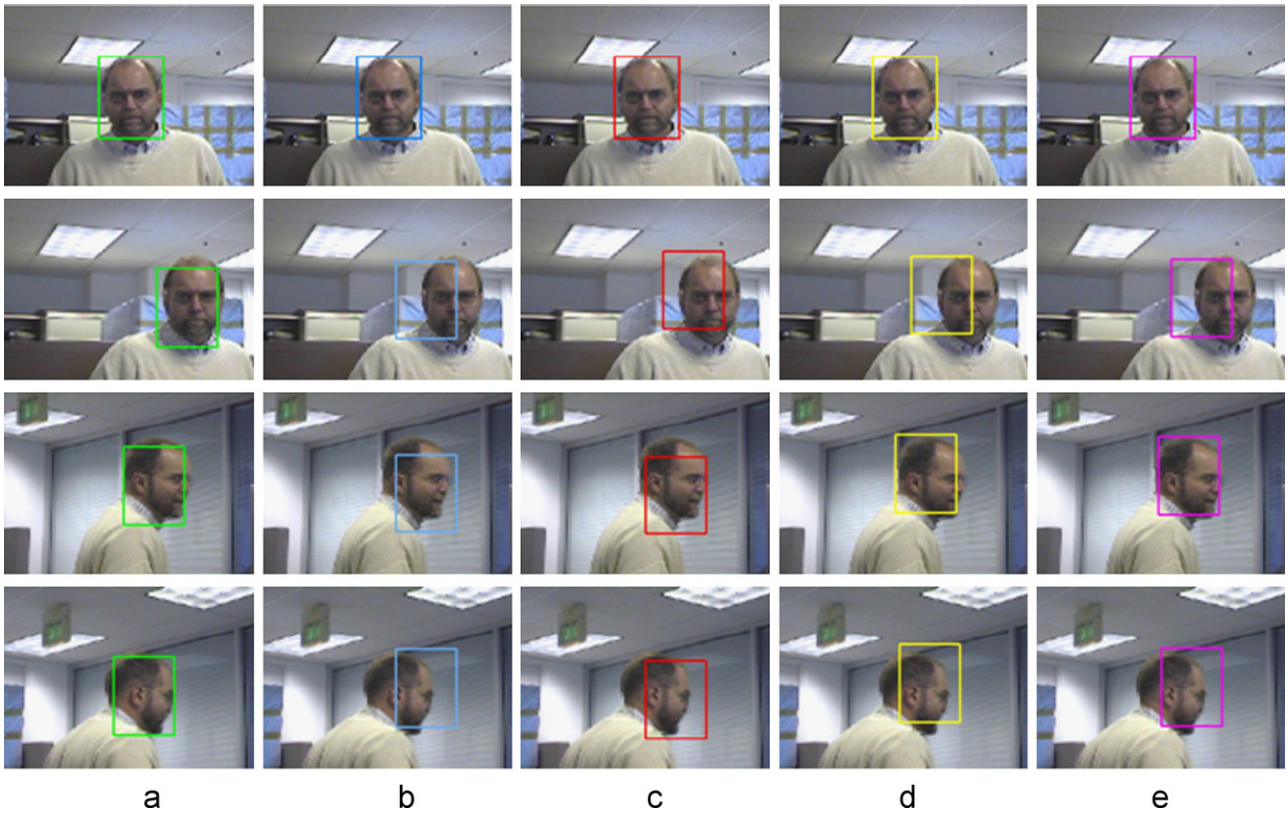

Fig. 6. Tracking results of video Head2 by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

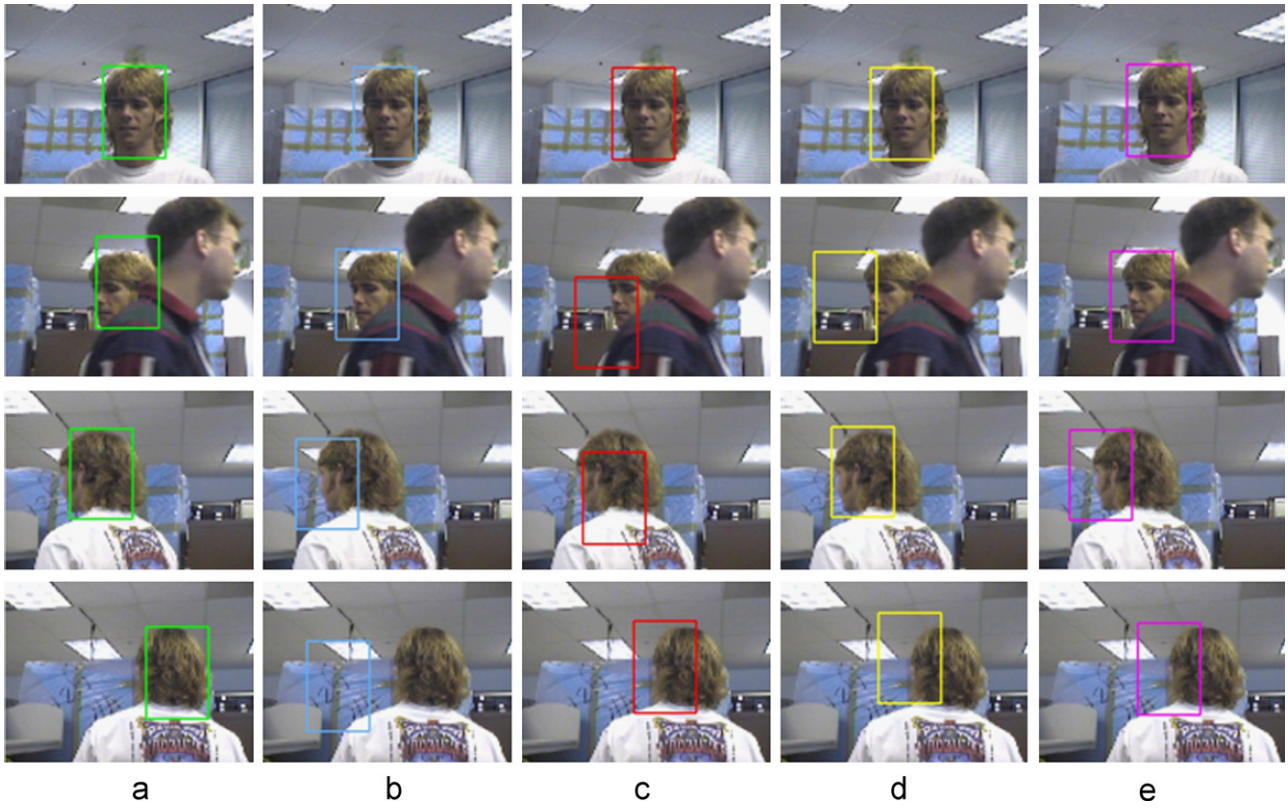

Fig. 7. Tracking results of video Head3 by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

appearance is similar to that of the face. The two methods still can track the face with satisfactory performance, while the other trackers are confused by the moving hands.

Fig. 9 shows the tracking results by different method for video Browse. The video contains pose variation, which can be overcome by most of the methods. The AR method fails to track the target due to

Fig. 8. Tracking results of video Head4 by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

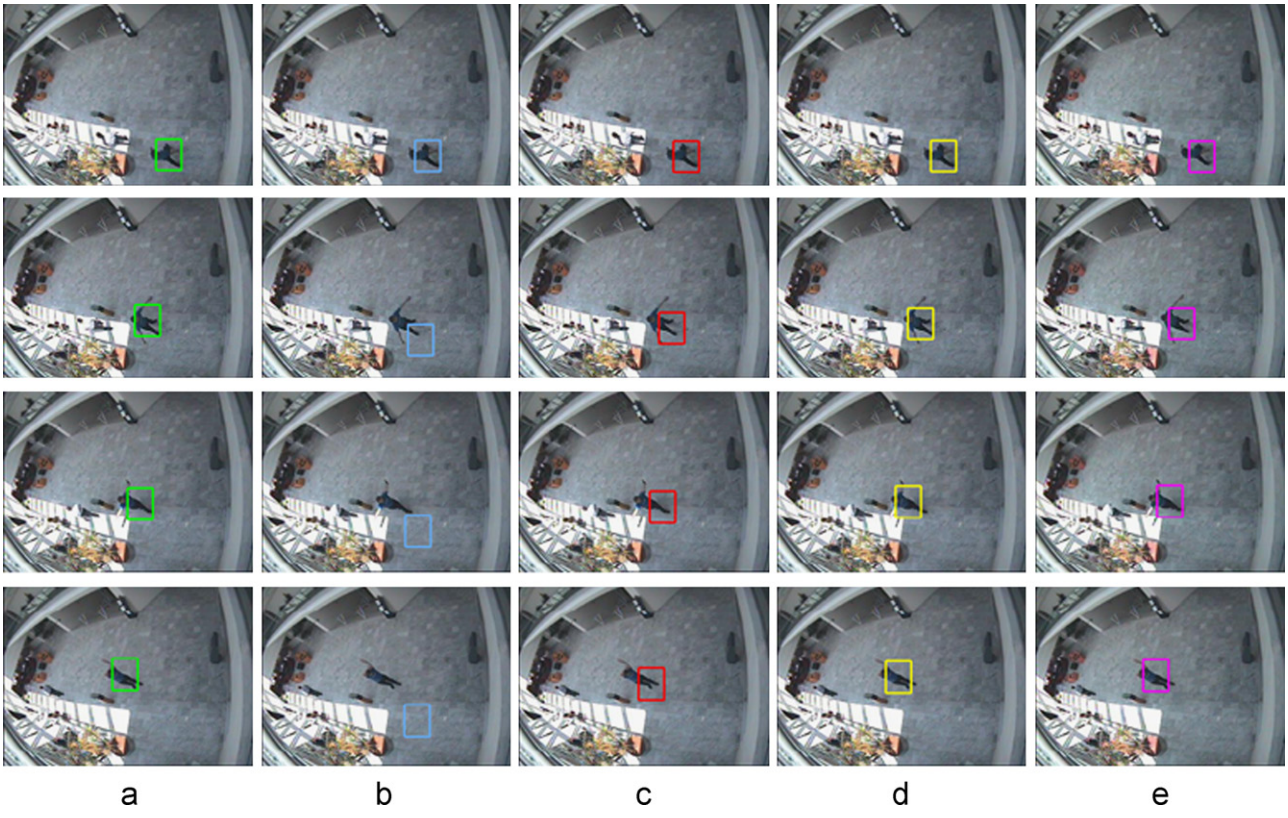

Fig. 9. Tracking results of video Browse by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

the small target size, which limits the number of salient regions. Fig. 10 shows the tracking results by all the methods for video Fight. In the video there are two persons who meet and fight against each other, then run away separately. It can be seen in Fig. 10a that our method consistently tracks the correct person throughout the entire video sequence. The other trackers are distracted by the other person

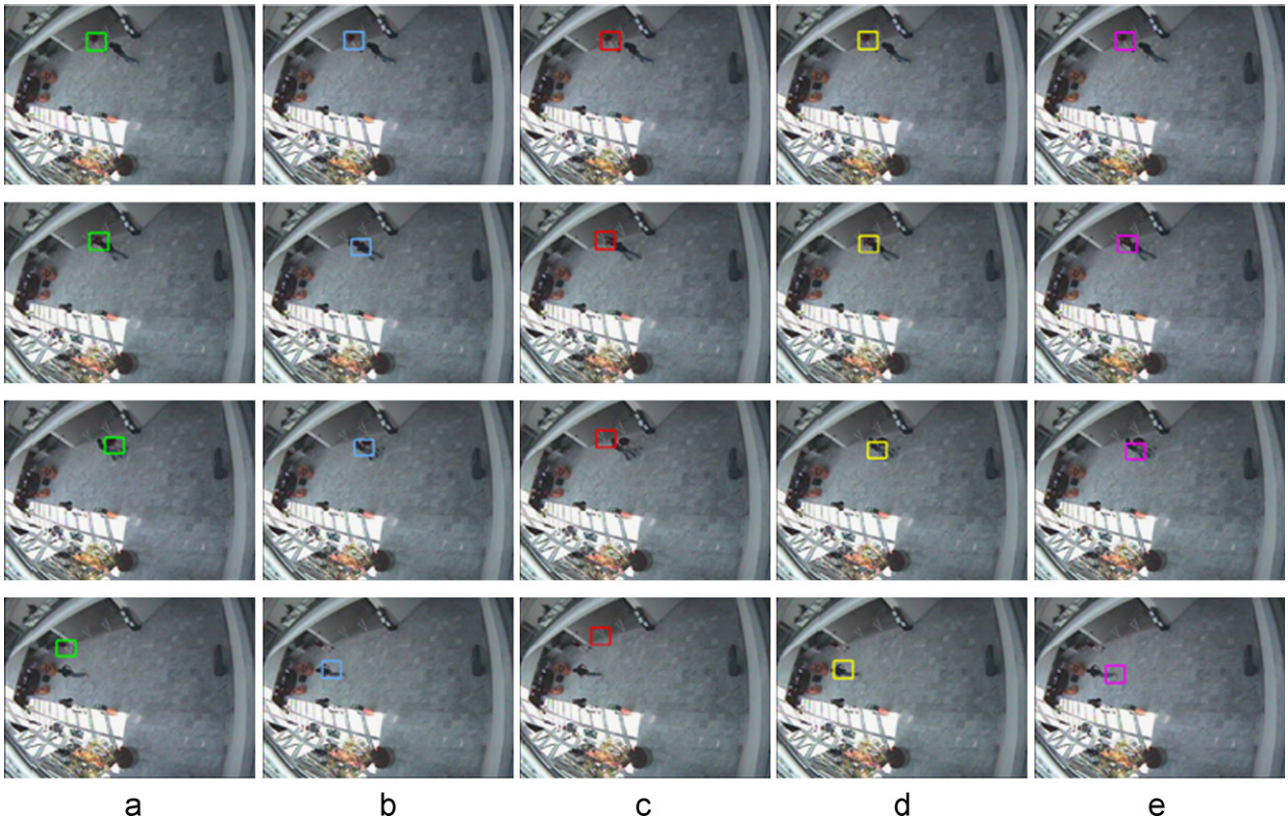

Fig. 10. Tracking results of video Fight by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

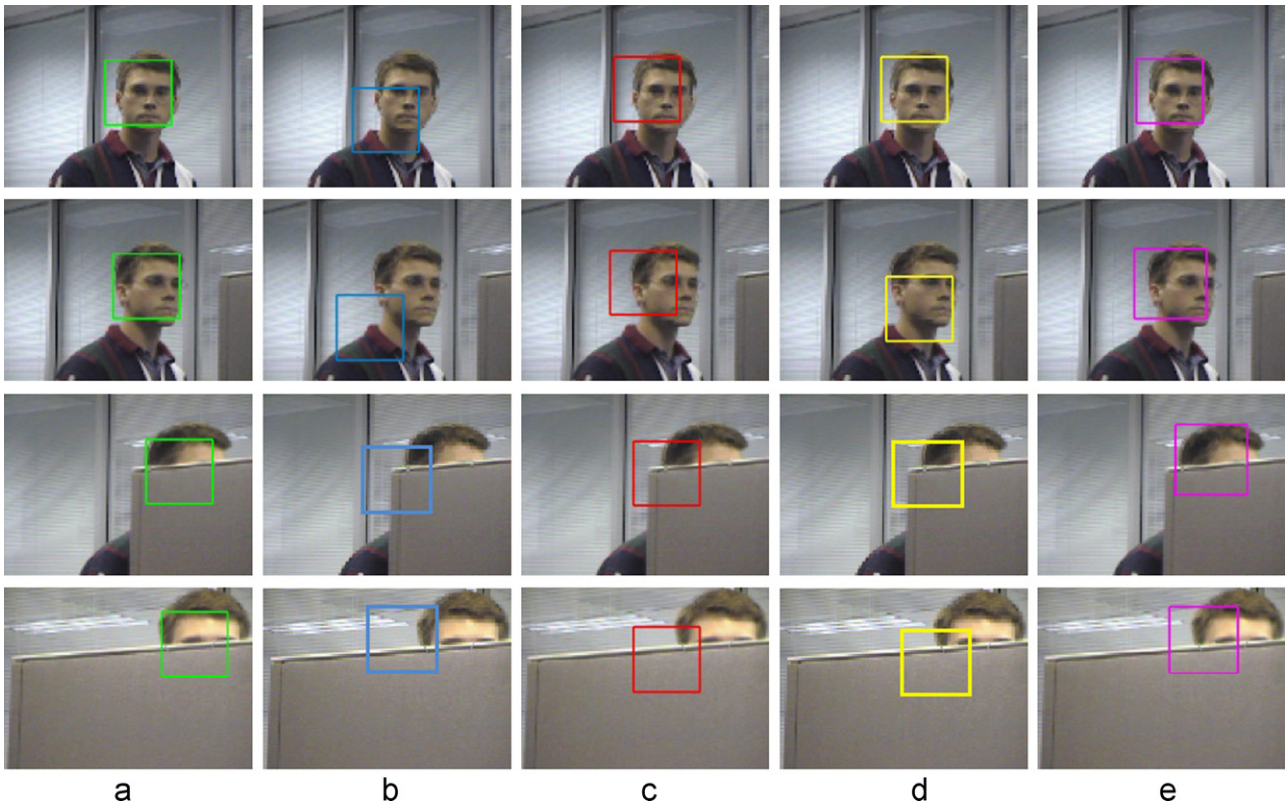

Fig. 11. Tracking results of video Head5 by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

when they stay close, as the person actually occupies more area in the scene than the target. Our method overcomes this issue by tracking salient proto-objects, such as the upper body of the target.

The tracking results of another five representative video sequences are exhibited in the experiment. Figs. 11–13 show the tracking results by different methods for video Head5, Head6, and Head7. These video

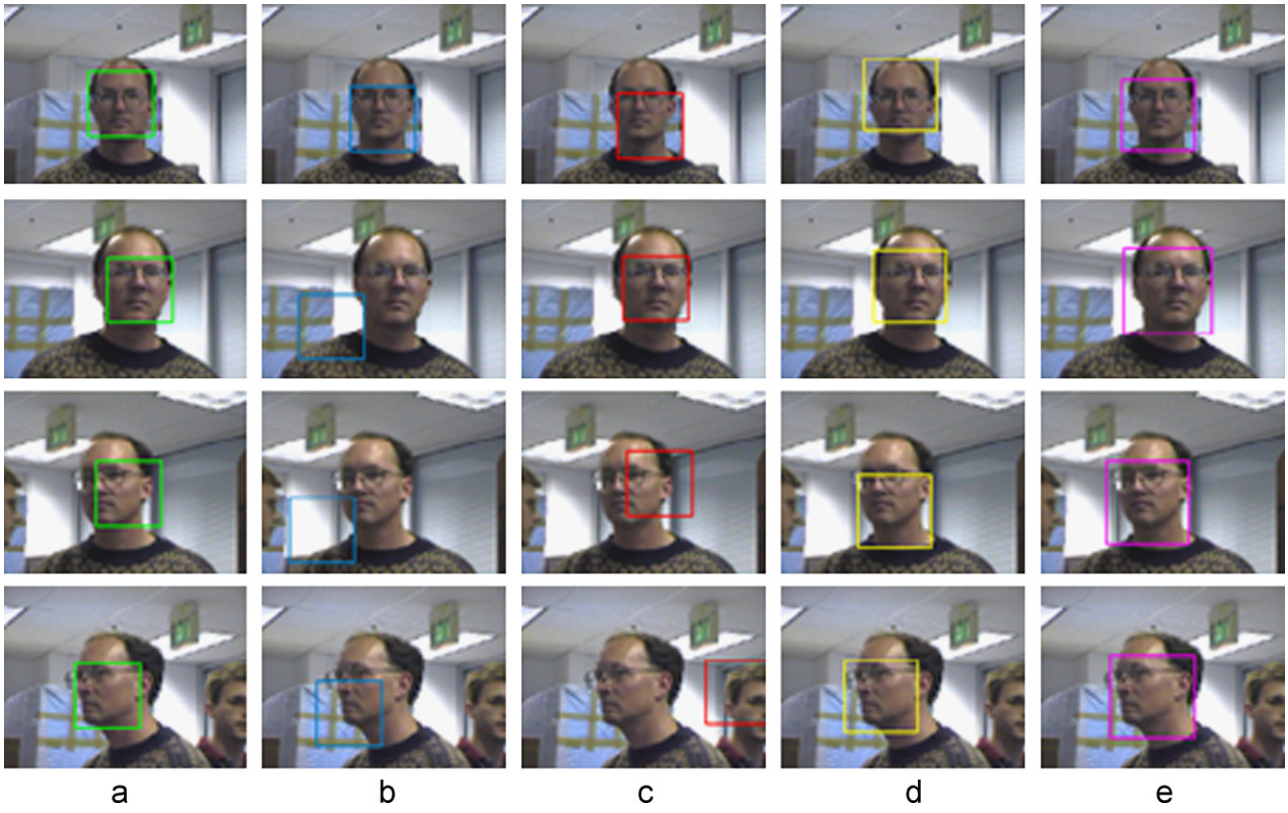

Fig. 12. Tracking results of video Head6 by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

Fig. 13. Tracking results of video Head7 by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

sequences from the Head Tracking dataset contain various challenging situations including fast movement, occlusion, distraction to the target, and appearance change caused by head rotation and body movement. Figs. 14 and 15 show the tracking results by different methods for video Reading and Meeting. These video sequences from the CAVIAR dataset are captured with distorted camera view and

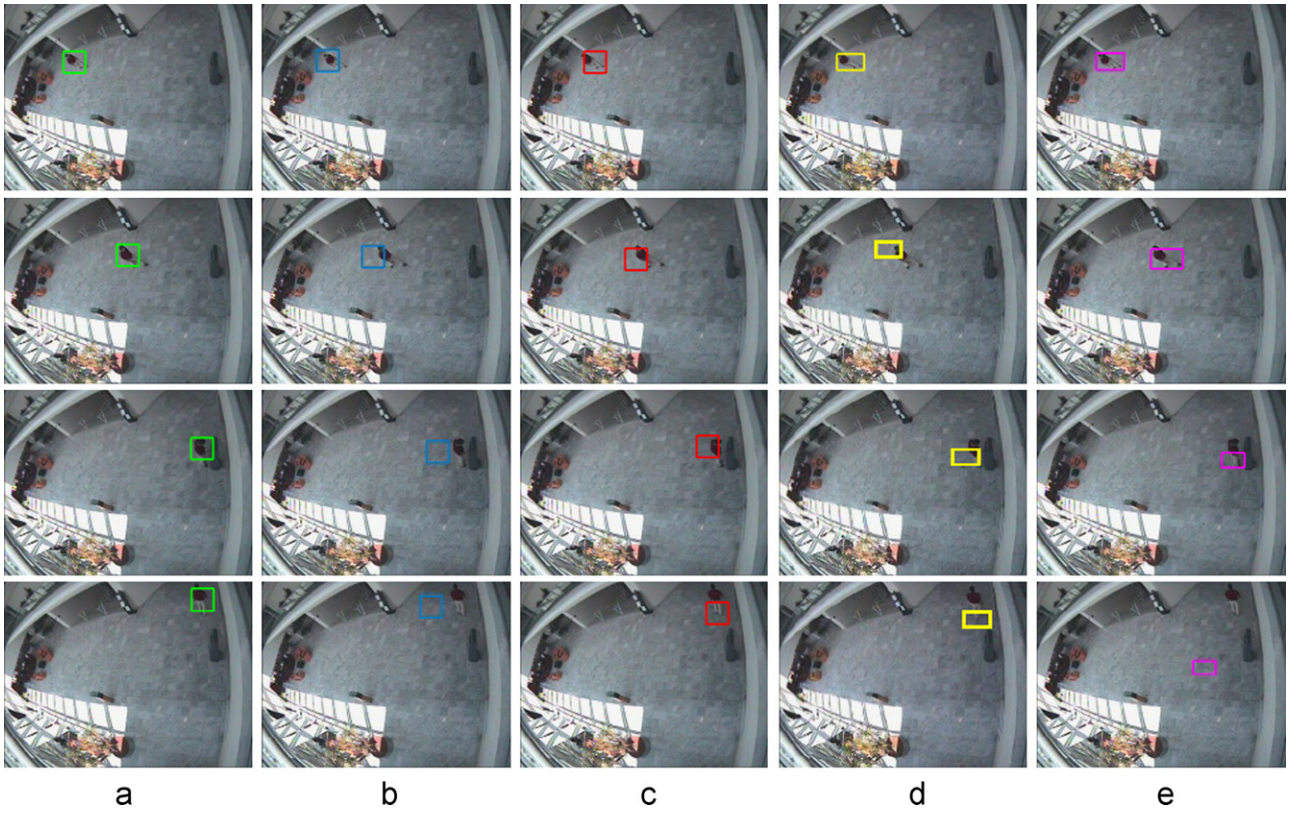

Fig. 14. Tracking results of video Reading by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

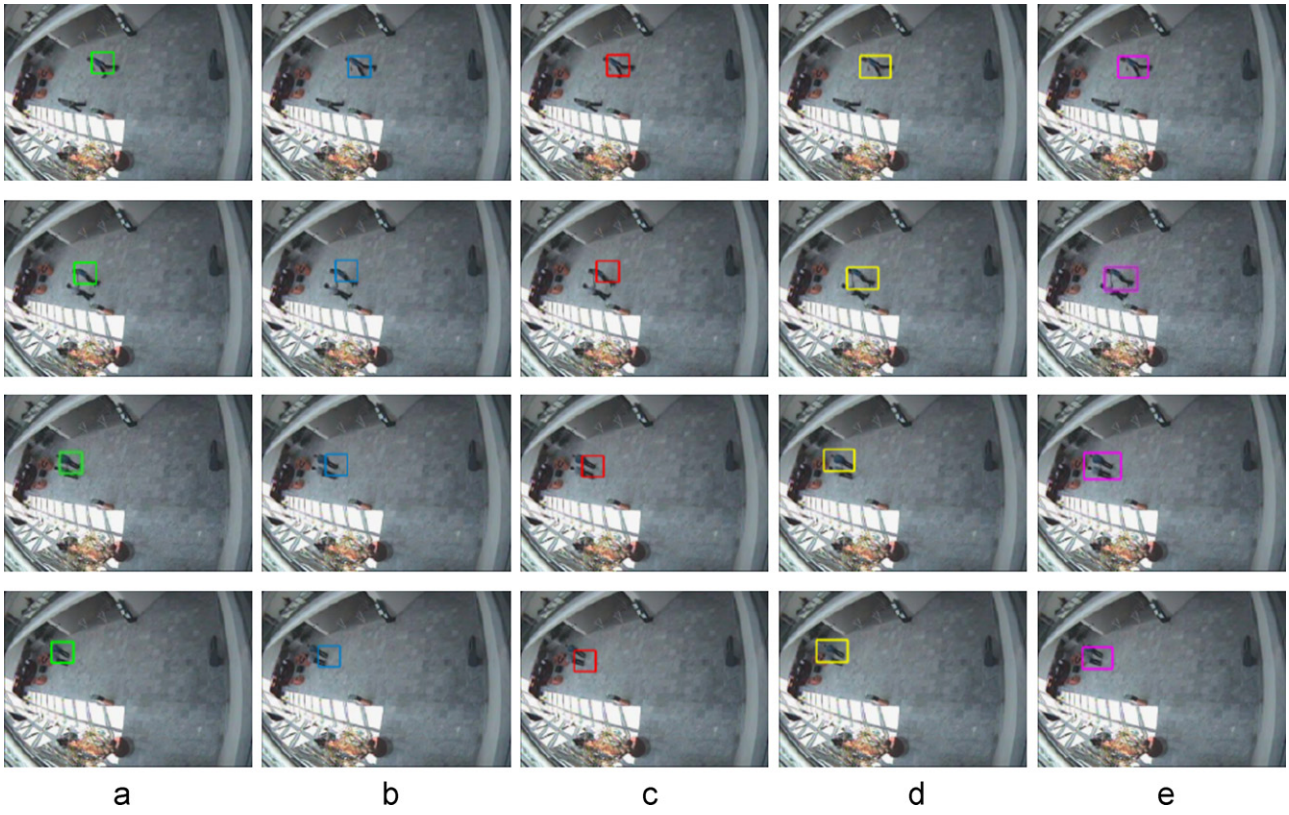

Fig. 15. Tracking results of video Meeting by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

contain abrupt movements, target deformation, and background distraction. It can be seen that the proposed method overall outperforms the other methods in the experiment.

Fig. 16 shows the tracking results by all the methods for video Skating. The video exhibits high variation of target's posture and illumination condition. Both AR and SF methods fail after a few

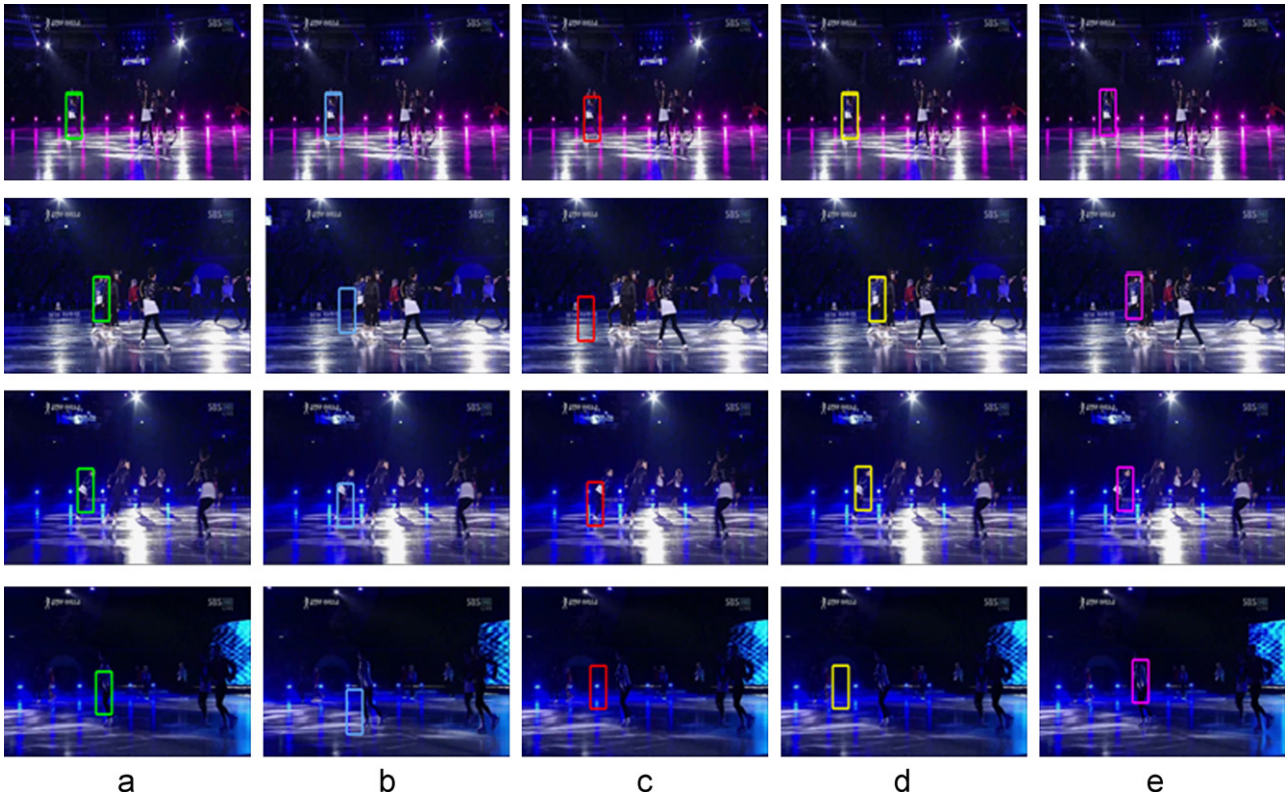

Fig. 16. Tracking results of video Skating by (a) the proposed method, (b) AR method, (c) SF method, (d) MIL method, and (e) VTD method.

Table 1
Tracking error rates for test video sequences.

	Proposed	AR	SF	MIL	VTD
Head1	**0.21**	0.47	0.36	0.49	0.32
Head2	**0.07**	0.24	0.24	0.19	0.19
Head3	**0.13**	0.59	0.29	0.28	0.36
Head4	**0.07**	0.11	0.40	0.17	0.36
Head5	**0.55**	0.78	0.67	0.75	0.63
Head6	**0.24**	0.53	0.41	0.27	0.27
Head7	0.22	0.48	0.35	0.56	**0.21**
Browse	**0.17**	0.52	0.38	0.21	0.19
Reading	**0.14**	0.52	0.53	0.45	0.73
Meeting	**0.15**	0.33	0.46	**0.15**	0.22
Fight	**0.14**	0.33	0.62	0.41	0.30
Skating	0.20	0.63	0.76	0.39	**0.13**

frames due to the complex visual environment and quick movement of the target. Our method can robustly track the target over time as long as some of proto-objects remain salient. It can be seen from the last frame in Fig. 16 that both the proposed method and VTD can accurately track the target, while MIL loses the target as the overall appearance has almost been completely changed.

We also quantitatively compare the results based on tracking error rate E:

$$E = \frac{1}{T}\left(\sum_{t=1}^{T}\frac{|rx_t - gx_t|}{w_t} + \sum_{t=1}^{T}\frac{|ry_t - gy_t|}{h_t}\right) \tag{15}$$

where rx_t and ry_t are the target central position of the tracking result; gx_t and gy_t are the target central position of the ground truth; w_t and h_t are the width and height of the ground truth; T is the total number of frames that have been tracked. Empirically, the error rate below 30% is considered as effective tracking. The quantitative results are summarized in Table 1, with the best results shown in bold. It can be found that our method outperforms the other four methods overall.

The proposed approach detects proto-objects in the first video frame during the initialization. The performance of the proposed approach has been evaluated for all the head sequences under different initialization conditions. In the experiment, the size of initialized target region varies from 50% to 150% of its original size (see Figs. 17 and 18). Table 2 shows the tracking error rates of the proposed method under different initialization situations. The percentage in the first row represents the ratio of the initialized size over its original size. It can be seen that the proposed method exhibits robust performance in most cases. However, when the initialized region becomes too large, proto-objects belonging to the background might be erroneously included, leading to biased tracking result. Details about the performance of proto-object detection in individual images can also be found in [24].

Figs. 19 and 20 show two examples of tracking failure by the proposed approach. As the proposed method uses proto-objects

to support the target tracking, it may not work well when all of the proto-objects are missed or become non-salient for a long period. For example, in Fig. 19d tracking failure occurs when the whole target is totally occluded and changes its motion at the same time. In Fig. 20d, the proto-objects lose their saliency when the light is suddenly dimmed out. When it is difficult to detect

Fig. 17. Tracking results of video Head2 with inaccurate initialization. The size of initialized region (the white rectangle in (a)) is 125% of its original size (the green rectangle in (a)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

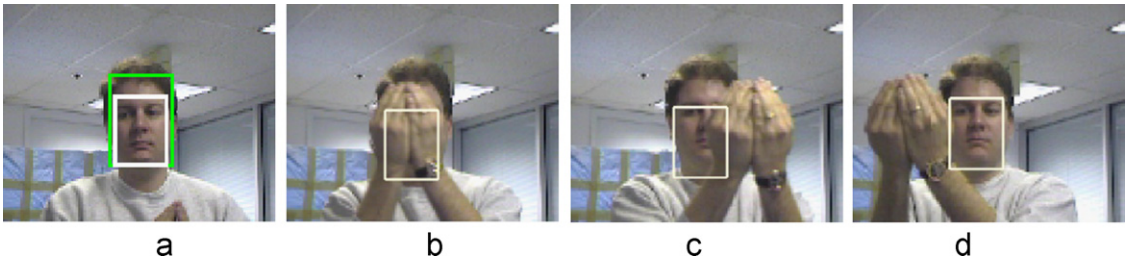

Fig. 18. Tracking results of video Head3 with inaccurate initialization. The size of initialized region (the white rectangle in (a)) is 75% of its original size (the green rectangle in (a)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Tracking error rates for inaccurate initialization.

	50%	75%	100%	125%	150%
Head1	0.35	0.17	0.21	0.24	0.37
Head2	0.18	0.16	0.07	0.15	0.17
Head3	0.11	0.16	0.13	0.26	0.21
Head4	0.14	0.26	0.07	0.14	0.08
Head5	0.61	0.61	0.55	0.62	0.65
Head6	0.21	0.13	0.24	0.22	0.24
Head7	0.21	0.22	0.22	0.36	0.45

Fig. 19. Tracking results for video Cup by the proposed method.

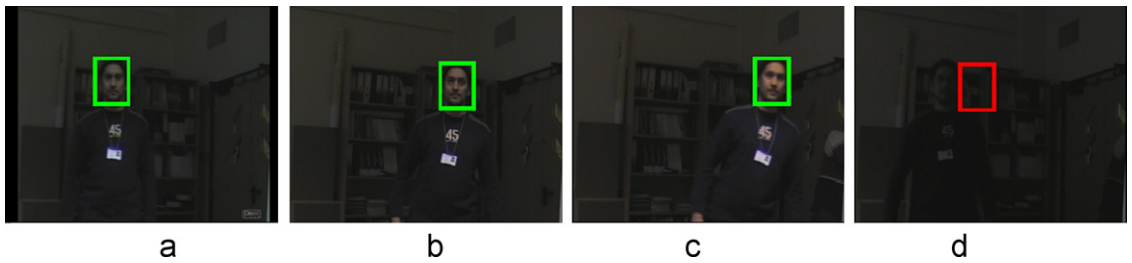

Fig. 20. Tracking results for video Change Illumination by the proposed method.

proto-objects, context features (such as auxiliary objects used in [27]) could be helpful to track the target. For example, in Fig. 19 the watch can be used as an auxiliary object to support the target tracking.

The independence assumption of proto-objects is used to reduce the computational complexity in this work. However, such independence assumption may not be valid in some cases. For example, in Fig. 21 irrelevant proto-object will be erroneously included during the tracking when the initialized region is oversized, leading to biased tracking result with the accumulation of error. To reduce the influence of irrelevant proto-objects, we could weigh the contribution of each proto-object according to their distance to the target center (inverse proportion in this work). For more complicated situations, with the trade-off in computational complexity, sophisticated techniques of joint tracking such as [37] could be exploited to further improve the robustness of tracking. Figs. 21 and 22 show the tracking results by the proposed method with and without the weighing scheme. It can be seen that the performance of target tracking has been improved for both video sequences. The tracking error rate drops from 0.40 to 0.11 for the first sequence and from 0.21 to 0.15 for the second one.

Figs. 23–25 show the results of multi-target tracking for video sequences Fight, Meeting, and Head7. Different targets are labeled with different colors in each video. It can be seen that our method can track the targets throughout the video sequence even under

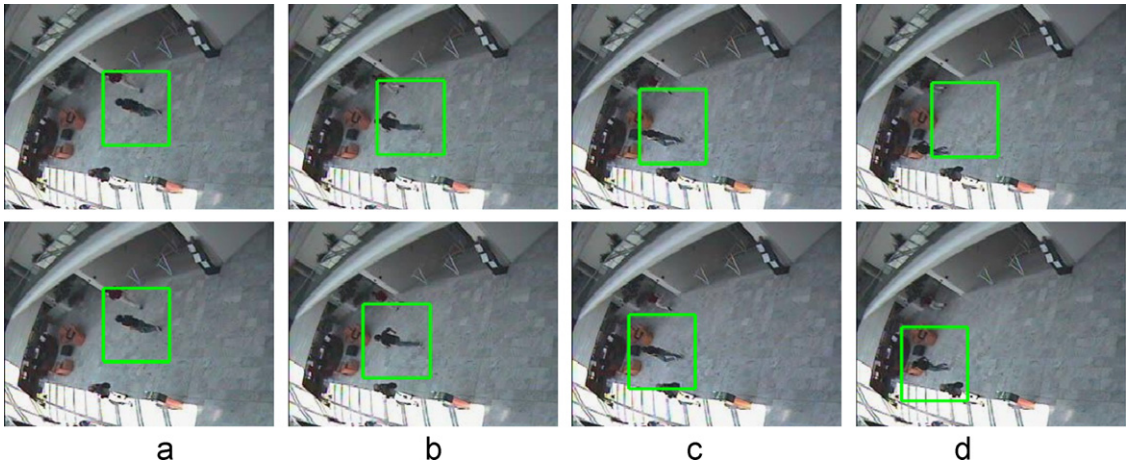

Fig. 21. Tracking results for video Fight by the proposed method with and without weighing proto-objects according to their distance to the target center. Tracking results without using the weighing scheme are shown in the first row. Tracking results using the weighing scheme are shown in the second row.

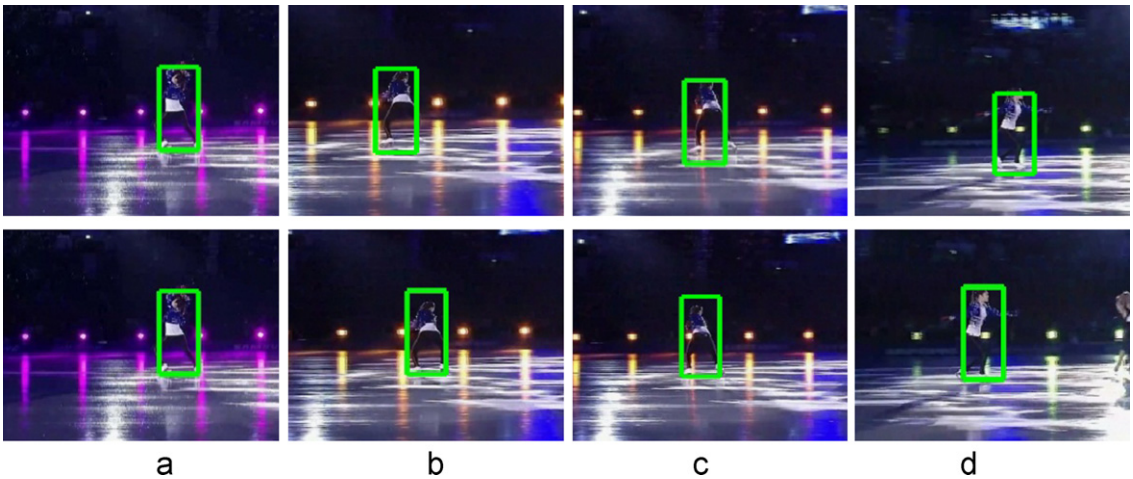

Fig. 22. Tracking results for video Skating by the proposed method with and without weighing proto-objects according to their distance to the target center. Tracking results without using the weighing scheme are shown in the first row. Tracking results using the weighing scheme are shown in the second row.

Fig. 23. Results of multi-target tracking in video Fight by the proposed method. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Fig. 24. Results of multi-target tracking in video Meeting by the proposed method. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Fig. 25. Results of multi-target tracking in video Head7 by the proposed method. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

various challenging conditions including appearance change, fast movement, object occlusion, and background distraction. Moreover, multiple segmentation [38] and multi-scale saliency detection [28] are used for proto-object detection in this work. When the number of segmentation/scale is high enough, the proposed method is able to handle small target even at pixel level (in such case the target may only consist of single proto-object). For example, Figs. 23 and 24 show the results of tracking small targets, where the size of the target could be as small as 20 pixels.

6. Conclusion

In this work we have presented a biologically inspired framework of visual tracking based on proto-objects. The proposed approach simultaneously estimates the states of the target and proto-objects over time. Tracking by proto-objects leads to more stable and robust performance than saliency region based methods. In our experiments, the proposed approach achieves superior performance compared to the state-of-the-art methods in challenging visual tasks.

Conflict of interest

We declare that we have no conflict of interest.

Appendix A. Supplementary Information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.patcog.2013.01.020.

References

[1] D. Comaniciu, P. Meer, Mean Shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 603–619.

[2] R.T. Collins, Mean-shift blob tracking through scale space, in: IEEE Conference on Computer Vision and Pattern Recognition 2003, pp. 234–240.

[3] S. Avidan, Ensemble tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 261–271.

[4] R.T. Collins, Y. Liu, M. Leordeanu, Online selection of discriminative tracking features, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1631–1643.

[5] B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, in: IEEE Conference on Computer Vision and Pattern Recognition 2009, pp. 983–990.

[6] D. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, International Journal of Computer Vision 77 (2007) 125–141.

[7] J. Kwon, K.M. Lee, Visual tracking decomposition, in: IEEE Conference on Computer Vision and Pattern Recognition 2010, pp. 1269–1276.

[8] V. Mahadevan, N. Vasconcelos , Saliency based discriminant tracking, in: IEEE Conference on Computer Vision and Pattern Recognition 2009, pp. 1007–1013.

[9] J. Fan, Y. Wu, S. Dai, Discriminative spatial attention for robust tracking, in: European Conference on Computer Vision 2010, pp. 480–493.

[10] M. Yang, J. Yuan, Y. Wu, Spatial selection for attentional visual tracking, in: IEEE Conference on Computer Vision and Pattern Recognition 2007.

[11] L. Itti, C. Koch, Computational modelling of visual attention, Nature Reviews Neuroscience 2 (2001) 194–203.

[12] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1254–1259.

[13] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition 2007.

[14] C. Guo, Q. Ma, L. Zhang, Spatio-Temporal saliency detection using phase spectrum of quaternion fourier transform, in: IEEE Conference on Computer Vision and Pattern Recognition 2008.

[15] Y. Hu, X. Xie, W.-Y. Ma, L.-T. Chia, D. Rajan, Salient region detection using weighted feature maps based on the human visual attention model, in: ACM International Conference on Multimedia 2005, pp. 993–1000.

[16] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, S.M. Hu, Global contrast based salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition 2011, pp. 409–416.

[17] R. Achantay, S. Hemamiz, F. Estraday, S. Süsstrunk, Frequency-tuned salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition 2009, pp. 1597–1604.

[18] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, SUN: a bayesian framework for saliency using natural statistics, Journal of Vision 8 (2008) 1–20.

[19] D. Gao, S. Han, N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 989–1005.

[20] L. Itti, Models of bottom-up and top-down visual attention, Ph.D. Thesis, California Institute of Technology, 2000.

[21] F. Orabona, G. Metta, G. Sandini, A proto-object based visual attention model, in: Attention in Cognitive Systems of Lecture Notes in Artificial Intelligence, 2007, pp. 198–215.

[22] R.A. Rensink, Seeing, sensing, and scrutinizing, Vision Research 40 (2000) 1469–1487.

[23] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (2006) 1395–1407.

[24] Z. Li, J. Xu, Y. Wang, G. Geers, J. Yang, Saliency detection based on proto-objects and topic model, in: IEEE Workshop on Applications of Computer Vision 2011,pp. 125–131.

[25] D. Gao, V. Mahadevan, N. Vasconcelos, On the plausibility of the discriminant center-surround hypothesis for visual saliency, Journal of Vision 8 (2008) 1–18.
[26] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in highly dynamic scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 171–177.
[27] M. Yang, Y. Wu, G. Hua, Context-aware visual tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 1195–1209.
[28] B. Alexe, T. Deselaers, V. Ferrari, What is an object?, in: IEEE Conference on Computer Vision and Pattern Recognition 2010, pp. 73–80.
[29] Y. Deng, B.S. Manjunath, Unsupervised segmentation of color-texture regions in images and video, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 800–810.
[30] F.-F. Li, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition 2005, pp. 524–531.
[31] T. Hofmann, Probabilistic latent semantic indexing, in: International ACM SIGIR Conference on Research and Development in Information Retrieval 1999, pp. 50–57.
[32] R.O. Duda, P.E. Hart, Use of the Hough transformation to detect lines and curves in pictures, Communications of the ACM 15 (1972) 11–15.
[33] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (2004) 5228–5235.
[34] S. Birchfield, Elliptical head tracking using intensity gradients and color histograms, in: IEEE Conference on Computer Vision and Pattern Recognition 1998, pp. 232–237.
[35] R. Maree, P. Geurts, J. Piater, L. Wehenkel, Random subwindows for robust image classification, in: IEEE Conference on Computer Vision and Pattern Recognition 2005, pp. 34–40.
[36] T. Kadir, M. Brady, Saliency, scale and image description, International Journal of Computer Vision 45 (2000) 83–105.
[37] 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 (2005) 174–188.
[38] B.C. Russell, A.A. Efros, J. Sivic, W.T. Freeman, A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, in: CVPR 2006.

Zhidong Li is currently pursuing Ph.D. degree at the University of New South Wales, Sydney, Australia. He received the M.E. degree in computer science from the University of New South Wales in 2006, and the B.S. degree in mathematics from the University of Xiamen in 2002. His research interests include pattern recognition, image processing and machine learning.

Weihong Wang is currently pursuing Ph.D. degree at the University of New South Wales, Sydney, Australia. He received the Master of Information Technology degree from the University of New South Wales in 2006, and the B.S. degree in computer science from the University of Sun Yat-Sen in 2003. His research interests include pattern recognition, image processing and machine learning.

Yang Wang received his Ph.D. degree in computer science from National University of Singapore in 2004. He is currently a researcher in Neville Roach Laboratory, National ICT Australia (NICTA). Before joining NICTA in 2006, he worked at Institute for Infocomm Research, Rensselaer Polytechnic Institute, and Nanyang Technological University. His research interests include video analysis, sensor networks, pattern classification, biomedical engineering, medical imaging, and computer vision.

Fang Chen holds a Ph.D. in Communications and Electronic Systems and an MBA. She is currently the research group manager for the Making Sense of Data research theme in National ICT Australia (NICTA), Sydney. She is also a Conjoint Associate Professor at the University of New South Wales. Her main research interests are human machine interaction, especially in multimodal systems, cognitive load modeling, speech processing, natural language processing, user interface design and evaluation.

Yi Wang is currently pursuing Ph.D. degree at the University of New South Wales, Sydney, Australia. He received the M.S. degree in Electrical Engineering from the University of North Dakota in 2008, and the B.E. degree in Communication Engineering from University of Electronic Science and Technology of China in 2005.