Online Appearance Model Learning and Generation for Adaptive Visual Tracking

Peng Wang *Member, IEEE,* and Hong Qiao, *Senior Member, IEEE*

***Abstract*—Several adaptive visual tracking algorithms have been recently proposed to capture the varying appearance of target. However, adaptability may also result in the problem of gradual drift, especially when the target appearance changes drastically. This paper gives some theoretical principles for online learning of target model, and then presents a novel adaptive tracking algorithm which is able to effectively cope with drastic variations in target appearance and resist gradual drift. Once target is localized in each frame, the patches sampled from target observation are first classified into foreground and background using an effective classifier. Then the adaptive, pure and time-continuous target model is extracted online through two processes: absorption process and rejection process, through which only the reliable features with high separability are absorbed in the new target model, while the "dangerous" features which may cause interfusion of background patterns are rejected. To minimize the influence of background and keep the temporal continuity of target model, two collaborative models dominant model and continuous model are designed. The proposed learning and generation mechanisms of target model are finally embedded in an adaptive tracking system. Experimental results demonstrate the robust performance of the proposed algorithm under challenging conditions.**

***Index Terms*—Adaptive visual tracking, appearance variation, collaborative models, gradual drift, model learning.**

I. Introduction

THE MAIN challenge of visual tracking with a moving camera can be attributed to two aspects: intrinsic appearance variability of target and extrinsic disturbance of environment. Intrinsic appearance variability includes pose variation, scale change, partial or complete occlusion, complex motion, and shape deformation. Extrinsic disturbance includes illumination variation, cluttered background, image noise, camera vibration, and the existence of similar objects. Therefore, a robust visual tracking algorithm should be able to adapt to the varying appearance of target, and meanwhile eliminate or reduce the influence of environment.

Manuscript received July 2, 2009; revised December 10, 2009 and May 11, 2010; accepted August 28, 2010. Date of publication January 13, 2011; date of current version March 2, 2011. This work was supported in part by the NNSF of China, under Grant 90820007, the 863 Program of China, under Grant 2007AA04Z228, the Outstanding Youth Fund of the NNSF of China, under Grant 60725310, and the 973 Program of China, under Grant 2007CB311002. This paper was recommended by Associate Editor D. Schonfeld.

The authors are with the Institute of Automation, Chinese Academy of Sciences, Beijing 100029, China (e-mail: peng_wang@ia.ac.cn; hong.qiao@ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2011.2105598

Numerous approaches for visual tracking with a moving camera have been proposed, and these methods can be mainly categorized as follows.

1) Active contour-based methods [1]–[3] describe objects exactly and effectively, even under disturbance or partial occlusion. These methods are suitable for tracking complex nonrigid shapes, and objects are usually tracked by contour evolution.
2) Template matching methods [4], [5] estimate the transformation which maps the images of a given object between two consecutive frames. In order to resist lighting variation or partial occlusion, the template can be smoothed temporally by Kalman filters [4], or divided into a suitable set of parts [5].
3) Kernel-based methods [6]–[9], which track nonrigid objects using viewpoint-insensitive histograms, are widely used due to their high efficiency of computation and robustness to rotation and occlusion.
4) Statistical methods for tracking [10]–[13] use the state space approach to model object properties, and visual tracking can be formulated as a state estimation problem by taking measurement and model uncertainties into account.
5) View-based or appearance-based representation is widely used to model the variations of object appearance and track moving object undergoing appearance variation by using linear [14]–[17] or nonlinear [18] dimensionality reduction techniques.
6) Multistrategy association methods combine rich object representations together [19], [20], or integrate different tracking methods in one framework [21], to enhance the robustness property. However, these methods significantly increase computation cost, due to the increase of dimensionality of the state space or the feature vector.

Most of the existing tracking algorithms in dynamic scenes focus on the appearance of target alone, while the importance of background is often ignored. Therefore, these trackers are vulnerable to be distracted by nontarget regions with similar appearance. These algorithms usually assume that target appearance or environment would not change drastically and it is easy to distinguish between foreground and background. However, in practice, target appearance and environment usually change with time, especially when target moves freely and the camera also moves to follow the tracked object (e.g., the camera mounted on a mobile robot platform). Therefore,

these methods can only work well for short duration and in well-controlled environments.

Several adaptive tracking algorithms have been proposed to cope with the variations of target appearance and environment. These algorithms automatically select the best features which have high distinguishing abilities between foreground and surrounding background [22]–[27], or treat tracking as a binary classification problem [28]–[30]. During tracking, the appearance models of both foreground and background are usually updated with time to adapt to the variation of target appearance and environment. Existing adaptive tracking algorithms have been proved to be effective under many challenging conditions, and have achieved great success. However, the remaining problems, including the difficulty in tracking object with drastic variation of appearance and the potential threat of gradual drift, are yet to be solved.

The model evolving scheme is able to help the tracker to capture the varying appearance of target. However, adaptability may also result in the problem of gradual drift. Gradual drift of target model is mainly caused by the interfusion of background patterns in target model during model evolution, due to the approximate representation of target. In other words, there are also some background pixels existing in target observation which is used to update target model. With the purpose of reducing gradual drift, most of the updating methods use the initial target appearance [22], [23] or key frames [36] as prior knowledge. However, these methods may be ineffective when target appearance changes drastically.

This paper presents a novel adaptive tracking system which is able to effectively cope with the drastic variations of target appearance and environment and resist gradual drift. The key contributions of this paper can be summarized as follows.

1) Some theoretical principles for online target model learning are given. Based on the proposed theoretical principles, a novel online target model learning and generation method is proposed.
2) An effective classifier is constructed online using the past consecutive target models as positive samples and the surrounding background of target observation as negative samples, and the patches sampled from target observation are classified into foreground and background by using the obtained classifier.
3) Then target model is extracted through two processes: absorption process in which only the reliable features that have high distinguishing abilities between foreground and background are absorbed in the new target model, and rejection process in which the "dangerous" features that may cause interfusion of background patterns in the new target model are rejected. The obtained target model has three properties of: adaptability, purity, and time-continuity.
4) Two collaborative target models: dominant model and continuous model are designed to minimize the influence of background and keep the temporal continuity of target models.

The remaining part of this paper is organized as follows. The most relevant algorithms that motivate this paper are reviewed in Section II. In Section III, some theoretical principles for online target model learning and an overview of the proposed adaptive tracking system are presented. Section IV provides a detailed description of the method for online target model learning. In Section V, generation of collaborative target models and target localization are presented. Experiments and discussion are given in Section VI. Finally, Section VII concludes this paper.

II. Related Work

A. Discriminative Feature Selection for Tracking

Tracking with discriminative feature selection are based on the point of view that the features that best distinguish between foreground and background are the best features for tracking [22]. Usually, a set of candidate features, such as the linear combinations of the basic R, G, B pixel values [22], and other multiple cues [23], [26], are chosen initially. Features can be evaluated and ranked in different ways, such as the variance ratios [22], principal component analysis [24], variance of mutual information [25], and Kullback–Leibler distance [27]. Then the best ones which make foreground most discriminative against surrounding background are selected for tracking.

These methods usually use predefined finite candidate features, and the target models usually rely on the initial appearance of target, which will result in invalidation when target appearance becomes completely different from the initial appearance. In the proposed method, no predefined candidate feature set is used, and the high separability of the tracker is obtained by adjusting the feature distributions of target model. The target model evolves with time not depending on the initial target appearance, but depending on the past consecutive target models with time decay forgetting factors and the current observations of both foreground and background. Therefore, the proposed method can follow up the real target appearance promptly and resist gradual drift effectively.

B. Foreground-Background Discrimination for Tracking

There are also other adaptive tracking algorithms which treat tracking as a foreground-background classification problem [28]–[31], [33]. Nguyen *et al.* [29] proposed a method that detects target through foreground-background texture discrimination using modified linear discriminant analysis (LDA). However, the algorithm may fail in case of poor separability between foreground and background, and it also suffers from gradual drift, due to the shortcoming of LDA and the wrong updating of foreground and background. Other tools used for foreground-background discrimination include AdaBoost [30], support vector machine [28], fast relevance vector machine [31], and online multiple instance learning Boosting [32].

Some of these methods need an off-line process to train the classifiers [28], and they can only work in well-controlled environment. The online classifiers [29]–[31], [33] usually directly use the current foreground/background observations as the new training samples, and the impurity of the training samples (e.g., the target observations usually contain some

background pixels) usually results in gradual deterioration of classifiers. In this paper, we propose an effective online classifier to classify the patches sampled from target observation into foreground and background. The history of the target model is introduced to the training process, i.e., the past consecutive target models which contain almost no background patterns are used as positive samples, and this guarantees the accuracy and robustness of the classifier.

C. Updating and Learning Appearance Models for Tracking

To adapt to the varying appearance of target, the tracking algorithms should update or learn the target model online with time. The existing model updating methods have achieved good performance, and can reduce gradual drift under some conditions. However, most of them use the initial appearance of target [22], [23], [34], [35] or key frames [36] as prior knowledge, and they are vulnerable to be ineffective when target appearance changes drastically.

There are also some tracking algorithms that learn the target model online by adapting the model parameters using online expectation-maximization (EM) algorithm [37], [38], or incrementally learning a low-dimensional subspace representation [16]. Jepson *et al.* [37] proposed a *WSL* tracker which involves stable model, wandering component, and outlier process to model the responses of wavelet filters. The parameters of appearance model are adjusted with time using an online EM algorithm. Ross *et al.* [16] combined subspace representations of appearance models, particle filters, and online updating schemes together to incrementally learn a low-dimensional subspace representation.

The target models obtained through these methods are usually not the exact targets but approximations of the real targets. In these methods, the background patterns may gradually get interfused into the new target model, which will deteriorate the tracking results especially under cluttered background. For example, the target model in [37] also learns the background structure when the target consistently moves with its surrounding background, and gradual drift happens.

In the proposed method, we first classify the patches sampled from the current target observation into foreground and background. Then two processes: absorption process and rejection process, are used to learn an adaptive, pure and time-continuous target model which can effectively cope with the varying target appearance and resist gradual drift. A combination of dominant model and continuous model is used to represent the target appearance, with the purpose of minimizing the influence of background and preventing sudden variation of target model. No prior knowledge or off-line training process, such as key frames obtained before tracking and initial appearance of target, is used in the proposed algorithm.

III. Theoretical Principles for Online Target Model Learning and Overview of the Proposed Method

This section will give some theoretical principles for online target model learning and an overview of the proposed tracking system which is designed based on the proposed principles.

A. Theoretical Principles for Online Target Model Learning

The aim of object tracking is to estimate the target state, denoted by $\boldsymbol{s}$, in each frame, including position coordinates, scale and orientation. Let $\boldsymbol{f}^t_{\text{target}}$ denotes the feature vector of target model at time t, then this tracking problem can be reduced to an estimation problem of finding the best match to the target model, that is

$$\mathbf{s}_t = \arg\min_{\mathbf{s}} \boldsymbol{O}(\boldsymbol{f}(\mathbf{s}), \boldsymbol{f}^t_{\text{target}}) \tag{1}$$

where $\boldsymbol{s}_t$ is the estimated target state at time t, and $\boldsymbol{O}(\cdot,\cdot)$ denotes the matching objective function, which is usually defined using the matching error between target model and the candidate region, such as the Bhattacharrya coefficient [6], the sum of squared error [34] and the Mahalanobis distance [4]. $\boldsymbol{f}(\mathbf{s})$ is the feature vector of the candidate region defined by $\mathbf{s}$ in the current image, and $\mathbf{s}$ is inside the region with the center $\mathbf{s}_{t-1}$ and the disturbance $\Delta\mathbf{s}$, that is $\|\mathbf{s}-\mathbf{s}_{t-1}\| \leqslant \|\Delta\mathbf{s}\|$.

In this paper, we mainly consider the problem that how to online learn the target model for a robust tracker.

1) *No Adaptation in Target Model:* Some tracking algorithms [6], [28] build the initial target model $\boldsymbol{f}^1_{\text{target}}$ initially and then use it for tracking during the whole tracking process, that is

$$\boldsymbol{f}^{t+1}_{\text{target}} = \boldsymbol{f}^1_{\text{target}}. \tag{2}$$

In these cases, the target model is independent of the target observation, and consequently becomes no longer representative of the target appearance when target appearance changes with time. Therefore, the tracker tends to deviate from the real target, and tracking fails.

2) *Learning Target Model Using Target Observation and Drifting Problem:* To adapt to the varying appearance of target, the target model should evolve with time according to the target observation $\boldsymbol{f}(\mathbf{s}_t)$ (region inside the tracking window), i.e., the target model should meet the *adaptability* requirement. Some methods [22], [23], [34], [35] adaptively update or learn the target model with time by

$$\boldsymbol{f}^{t+1}_{\text{target}} = \boldsymbol{g}(\boldsymbol{f}^1_{\text{target}}, \boldsymbol{f}^t_{\text{target}}, \boldsymbol{f}(\mathbf{s}_t)) \tag{3}$$

where $\boldsymbol{g}()$ is the updating or learning function. Through (3), the target observation is introduced into the updating or learning process of target model, and the obtained new target model should adaptively represent the real appearance of target, that is

$$\boldsymbol{f}^{t+1}_{\text{target}} = \arg\min_{\boldsymbol{f}} \boldsymbol{O}(\boldsymbol{f}, \boldsymbol{f}^{\text{real},t+1}_{\text{target}}) \tag{4}$$

where $\boldsymbol{f}^{\text{real},t+1}_{\text{target}}$ is the feature vector of the real target appearance at time $t+1$.

However, in practice, due to the approximate representation of target (e.g., using a rectangle or an ellipse to represent the human head) and the uncertainty in tracking, there are usually some background pixels interfused in the target observation $\boldsymbol{f}(\mathbf{s}_t)$. Therefore, the target model obtained through (3) is usually not a pure target but a combination of the real target with an error term

$$\boldsymbol{f}^{t+1}_{\text{target}} = \boldsymbol{f}^{\text{real},t+1}_{\text{target}} + \sum_{i=1}^{t} \boldsymbol{f}^{i\to i+1}_{\varepsilon} \tag{5}$$

TABLE I
THEORETICAL PRINCIPLES FOR ONLINE TARGET MODEL LEARNING

1) *adaptability*, i.e., the model should adaptively account for variations in target appearance

$$f_{\text{target}}^{t+1} = \arg\min_{f} O(f, f_{\text{target}}^{\text{real},t+1})$$

2) *purity*, i.e., the model is not 'polluted' by background patterns, and is able to resist gradual drift during tracking

$$f_{\text{target}}^{t+1} = \arg\min_{f} \boldsymbol{O}(f, f_{bac1}^{t})$$

3) *time-continuity*, i.e., the model can inherit the nature of past consecutive models, and can effectively cope with sudden tracking failure

$$f_{\text{target}}^{t+1} = \boldsymbol{g}(\{f_{\text{target}}^{t-k}\}_{k=0}^{t-1}, f(s_t), f_{bac1}^{t})$$

To design a robust tracker which can adapt to the varying appearance of target and resist gradual drift, the obtained target models should meet the requirements of adaptability, purity, and time-continuity.

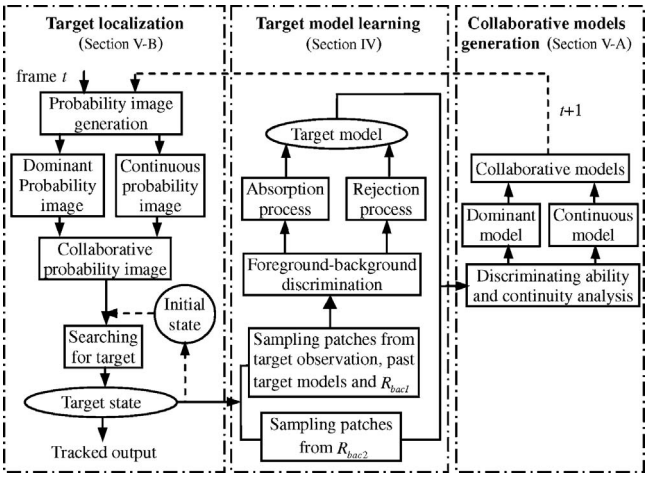


Fig. 1. Framework of the proposed adaptive tracking system.

where $f_{\varepsilon}^{i\to i+1}$ is the accumulated error in target model from time i to i+1, which is mainly caused by the gradual interfusion of the immediate surrounding background. Then the state of target can be estimated by

$$\mathbf{s}_{t+1} = \arg\min_{\mathbf{s}} \boldsymbol{O}(f(\mathbf{s}), f_{\text{target}}^{\text{real},t+1} + \sum_{i=1}^{t} f_{\varepsilon}^{i\to i+1}) \tag{6}$$

and the accumulated errors in target model will cause deviation in the estimated state of target

$$\mathbf{s}_{t+1} = \mathbf{s}_{t+1}^{\text{real}} + \sum_{i=1}^{t} \mathbf{s}_{i\to i+1}^{\varepsilon} \tag{7}$$

where $\mathbf{s}_{t+1}^{\text{real}}$ is the real state of target at time $t+1$, and $\mathbf{s}_{i\to i+1}^{\varepsilon}$ denotes the accumulated error in the estimated state of target from time i to $i+1$. The tracker gradually deviates from the real target, i.e., gradual drift happens.

3) *Online Adaptive, Pure, and Time-Continuous Target Model Learning:* Gradual drift is mainly caused by the interfusion of background patterns in target model during the model evolving process. Therefore, to resist gradual drift, the target model should be *pure* enough and the immediate surrounding background of target observation should be involved in the process of target model learning, that is

$$f_{\text{target}}^{t+1} = \boldsymbol{g}(f_{\text{target}}^{t}, f(s_t), f_{bac1}^{t}) \tag{8}$$

where f_{bac1}^{t} denotes the immediate surrounding background of target observation. This is under the assumption that the surrounding background f_{bac1}^{t} could well represent the interfused background in target observation, that is

$$\| f_{bac1}^{t} - f_{\varepsilon}(s_t) \| \approx 0 \tag{9}$$

where $f_{\varepsilon}(s_t)$ is the interfused background in target observation $f(s_t)$.

To obtain a *pure* target model, the background patterns should be rejected from the new target model, that is

$$f_{\text{target}}^{t+1} = \arg\max_{f} \boldsymbol{O}(f, f_{bac1}^{t}). \tag{10}$$

When sudden tracking failure happens at time t, the target state error $\mathbf{s}_{t\to t+1}^{\varepsilon}$ will increase sharply, and f_{bac1}^{t} is not able to represent the interfused background in target observation. This will result in a completely inaccurate target model, and the tracker has no chance to reacquire the target. Therefore, the target model should keep some *time-continuity* properties and inherit the nature of the past consecutive models, then

$$f_{\text{target}}^{t+1} = \boldsymbol{g}(\{f_{\text{target}}^{t-k}\}_{k=0}^{t-1}, f(s_t), f_{bac1}^{t}). \tag{11}$$

The past target models should have different influences in the new target model learning process, and the latest ones should have the most significant influence.

Based on the above analysis, some theoretical principles for online target model learning are given in Table I.

B. Overview of the Proposed System

According to the proposed principles, we design a novel adaptive tracking system which is able to adapt to the varying appearance of target and resist gradual drift under various challenging conditions.

The framework of the proposed system is shown in Fig. 1. Three steps are taken in the proposed algorithm: online target model learning, collaborative models generation, and target localization.

1) *Online target model learning:* An effective classifier is constructed online to classify the patches sampled from target observation into foreground and background. Then the new target model is extracted through two processes: absorption process and rejection process (Section IV). Through this way, we can get an adaptive, pure, and time-continuous appearance model which is able to follow up the varying appearance of target and resist gradual drift.
2) *Collaborative models generation:* The collaborative target models are generated by combining a dominant model with a continuous model (Section V-A). Through this, the influence of background is reduced, and this mechanism also prevents sudden changes in target model effectively.

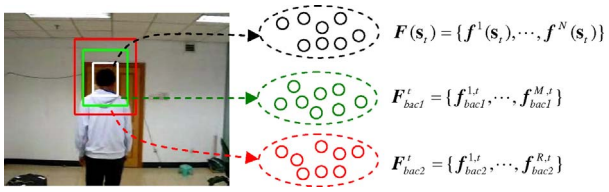


Fig. 2. Image patches sampled from target observation and local backgrounds at time t.

3) *Target localization:* The obtained collaborative target models will be used to localize target in the next frame (Section V-B).

For each frame, target is localized using the appearance models obtained in the previous frame, and the estimated target state, such as location and scale, is saved as initial state for next frame. As illustrated in Fig. 2, the white rectangle is the tracking window which is used to represent the target approximately. Note that the tracking window also includes some background pixels which are mistakenly treated as foreground. The region between the tracking window and the green rectangle is denoted by R_{bac1}, which is used to sample background patches as negative samples for foreground-background discrimination and target model extraction (explained in Section IV). The region between the green rectangle and the red rectangle is denoted by R_{bac2} which is used for finding discriminative features and generating collaborative models (Section V has the details). Considering the displacement of target between two consecutive frames, R_{bac2} is usually set with larger size than R_{bac1}.

IV. Online Learning of Target Appearance Model

In a tracking scenario, target is usually approximately represented by primitive geometric shapes, such as rectangle or ellipse. However, if the tracked object is not an exact rectangle or ellipse then the tracking window will include some background pixels which will be mistaken as foreground. In other words, target observation (region inside the tracking window) is not a pure target, but an approximate representation of target. The existing online model updating or learning processes usually directly use target observation to evolve target model [22], [23], [29], [34], [35], [37]. These methods may cause interfusion of background patterns in the new target model gradually, and tracking accuracy decreases until the tracker mistakenly captures another object finally, i.e., gradual drift happens.

In this section, a novel online target model learning method is proposed. In each frame, once target is localized, we first classify the patches sampled from target observation into foreground and background using an effective classifier which is constructed online. Then two interactive processes: absorption process and rejection process are introduced to extract the adaptive, pure, and time-continuous target model which can effectively follow up the varying appearance of target and meanwhile resist gradual drift.

A. Representation of Foreground/Background

Object appearance can be represented in many ways, such as probability densities [6] and templates [4]. In this paper, target and background are represented by H-bin histograms, due to their invariance to rotation and translation. Then the feature distribution of target model can be represented by

$$\boldsymbol{f}^t_{\text{target}}=\{f^t_{\text{target},h}\}_{h=1,\ldots,H},\ \sum_{h=1}^{H} f^t_{\text{target},h}=1. \tag{12}$$

Based on the proposed principles in Table I, the past consecutive target models should be involved in the learning process of target model. The set of past consecutive target models can be denoted by

$$\boldsymbol{F}_{\text{target}}=\{\boldsymbol{f}^{t-k}_{\text{target}},\ldots,\boldsymbol{f}^{t}_{\text{target}}\},\ \boldsymbol{f}^{t-k}_{\text{target}}\in\mathbb{R}^H,k=1,\ldots,K \tag{13}$$

where K is the number of the past consecutive target models used. Then, we randomly extract W patches of the same size from each target model

$$\boldsymbol{F}^{t-k}_{\text{target}}=\{\boldsymbol{f}^{1,t-k}_{\text{target}},\ldots,\boldsymbol{f}^{W,t-k}_{\text{target}}\},\ \boldsymbol{f}^{w,t-k}_{\text{target}}\in\mathbb{R}^H,w=1,\ldots,W \tag{14}$$

where $\boldsymbol{f}^{w,t-k}_{\text{target}}$ denotes the wth patch sampled from the past target model $\boldsymbol{f}^{t-k}_{\text{target}}$.

Once target is localized in frame t, the estimated target state can be defined by $\mathbf{s}_t=(x_t,y_t,w_t,h_t)$, including the position coordinates (x_t,y_t) and scale (w_t,h_t). We randomly extract N patches of the same size from target observation $\boldsymbol{f}(\mathbf{s}_t)$ (Fig. 2)

$$\boldsymbol{F}(\mathbf{s}_t)=\{\boldsymbol{f}^1(\mathbf{s}_t),\ldots,\boldsymbol{f}^N(\mathbf{s}_t)\},\ \boldsymbol{f}^n(\mathbf{s}_t)\in\mathbb{R}^H,n=1,\ldots,N \tag{15}$$

where $\boldsymbol{f}^n(\mathbf{s}_t)$ denotes the nth patch sampled from target observation $\boldsymbol{f}(\mathbf{s}_t)$, and $\boldsymbol{f}^n(\mathbf{s}_t)=\{f^n_h(\mathbf{s}_t)\}_{h=1,\ldots,H}$, $\sum_{h=1}^{H} f^n_h(\mathbf{s}_t)=1$. Most of these patches are formed by pure foreground pixels. However, the patches far from the center of tracking window may also contain some background pixels which will be gradually interfused in target model during model learning.

Similarly, we randomly extract M and R patches from background regions R_{bac1} and R_{bac2}, respectively (Fig. 2)

$$\begin{aligned}&\boldsymbol{F}^t_{bac1}=\{\boldsymbol{f}^{1,t}_{bac1},\ldots,\boldsymbol{f}^{M,t}_{bac1}\},\ \boldsymbol{f}^{m,t}_{bac1}\in\mathbb{R}^H,m=1,\ldots,M\\&\boldsymbol{F}^t_{bac2}=\{\boldsymbol{f}^{1,t}_{bac2}\ldots,\boldsymbol{f}^{R,t}_{bac2}\},\ \boldsymbol{f}^{r,t}_{bac2}\in\mathbb{R}^H,r=1,\ldots,R\end{aligned} \tag{16}$$

where $\boldsymbol{f}^{m,t}_{bac1}$ denotes the mth patch sampled from background R_{bac1}, and $\boldsymbol{f}^{m,t}_{bac1}=\{f^{m,t}_{bac1,h}\}_{h=1,\ldots,H}$, $\sum_{h=1}^{H} f^{m,t}_{bac1,h}=1$; $\boldsymbol{f}^{r,t}_{bac2}$ denotes the rth patch sampled from background R_{bac2}, and $\boldsymbol{f}^{r,t}_{bac2}=\{f^{r,t}_{bac2,h}\}_{h=1,\ldots,H}$, $\sum_{h=1}^{H} f^{r,t}_{bac2,h}=1$. The set of background patches $\boldsymbol{F}^t_{bac1}$ is designed to provide negative samples for foreground-background discrimination, and $\boldsymbol{F}^t_{bac2}$ is designed with the purpose of generating collaborative target models (as described in Section V).

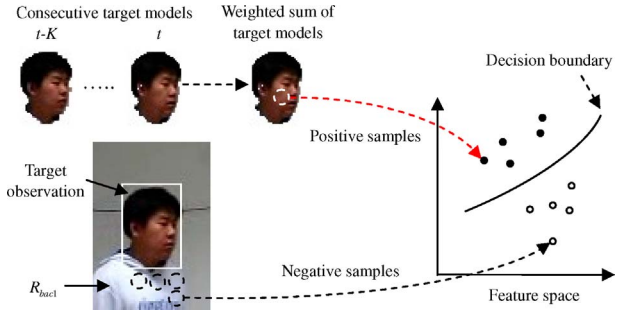


Fig. 3. Construction of the classifier. The patches sampled from the weighted sum of consecutive target models are used as positive samples, and patches sampled from R_{bac1} are used as negative samples.

B. Foreground-Background Discrimination for Patches Sampled from Target Observation

As illustrated in Fig. 2, if target is not an exact rectangle, then the tracking window will include some immediate surrounding background pixels. These background patterns will be gradually interfused in target model during model evolving with time, and gradual drift happens. In view of this, this subsection aims to construct a classifier to classify the image patches sampled from target observation (i.e., $\boldsymbol{F}(\mathbf{s}_t)$) into two classes: foreground (w_{target}) and background (w_{bac}).

Based on Bayes' rule, the posterior probability that the nth target observation patch $\boldsymbol{f}^n(\mathbf{s}_t)$ that comes from foreground is

$$P(w_{\text{target}}|\boldsymbol{f}^n(\mathbf{s}_t)) = \frac{P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{\text{target}})P(w_{\text{target}})}{P(\boldsymbol{f}^n(\mathbf{s}_t))} \tag{17}$$

where $P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{\text{target}})$ denotes the probability that $\boldsymbol{f}^n(\mathbf{s}_t)$ is observed as foreground. $P(\boldsymbol{f}^n(\mathbf{s}_t))$ is the prior probability of $\boldsymbol{f}^n(\mathbf{s}_t)$ being observed in the nth patch, and $P(w_{\text{target}})$ is the prior probability of $\boldsymbol{f}^n(\mathbf{s}_t)$ belonging to foreground.

Similarly, the posterior probability that the nth patch $\boldsymbol{f}^n(\mathbf{s}_t)$ comes from background is

$$P(w_{bac}|\boldsymbol{f}^n(\mathbf{s}_t)) = \frac{P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{bac})P(w_{bac})}{P(\boldsymbol{f}^n(\mathbf{s}_t))}. \tag{18}$$

To estimate the conditional probabilities $P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{\text{target}})$ and $P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{bac})$, the patches sampled from the past consecutive target models are used as positive samples (foreground) and patches sampled from the current immediate surrounding background R_{bac1} are used as negative samples (background). This is to keep the temporal continuity of the target model and meanwhile identify and eliminate the background patterns from current target observation. The similarities between the nth patch $\boldsymbol{f}^n(\mathbf{s}_t)$ and the labeled samples can be computed as

$$\begin{aligned}\rho(\boldsymbol{f}^n(\mathbf{s}_t),\, \boldsymbol{f}^{t-k}_{\text{target}}) &= \max_{w} \sum_{h=1}^{H} \sqrt{f^n_h(\mathbf{s}_t) f^{w,t-k}_{\text{target},h}} \\ \rho(\boldsymbol{f}^n(\mathbf{s}_t),\, \boldsymbol{f}^{m,t}_{bac1}) &= \sum_{h=1}^{H} \sqrt{f^n_h(\mathbf{s}_t) f^{m,t}_{bac1,h}}.\end{aligned} \tag{19}$$

The past consecutive target models have different influences on the construction of the classifier, and the latest one has the most important effect. The influence of the kth consecutive target model is given by a time-decaying forgetting factor

$$\lambda_k = e^{-\alpha k} / \sum_{k=0}^{K} e^{-\alpha k} \tag{20}$$

where α is the decay coefficient. Then the conditional probability $P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{\text{target}})$ can be estimated by

$$P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{\text{target}}) = C_1 \sum_{k=0}^{K} \lambda_k \rho(\boldsymbol{f}^n(\mathbf{s}_t),\, \boldsymbol{f}^{t-k}_{\text{target}}) \tag{21}$$

where C_1 is a normalization constant, which is used to impose the condition $P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{\text{target}}) + P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{bac}) = 1$.

Equation (21) is formulated to enhance the adaptability and the temporal continuity of the classifier. It is equivalent to

$$P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{\text{target}}) = C_1 \rho(\boldsymbol{f}^n(\mathbf{s}_t),\, \sum_{k=0}^{K} \lambda_k \boldsymbol{f}^{t-k}_{\text{target}}) \tag{22}$$

where $\sum_{k=0}^{K} \lambda_k \boldsymbol{f}^{t-k}_{\text{target}}$ is the weighted sum of the consecutive target models.

Similarly, the conditional probability $P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{bac})$ can be estimated by

$$P(\boldsymbol{f}^n(\mathbf{s}_t)|w_{bac}) = C_1 \max_{m} \rho(\boldsymbol{f}^n(\mathbf{s}_t),\, \boldsymbol{f}^{m,t}_{bac1}). \tag{23}$$

The prior probabilities $P(w_{\text{target}})$ and $P(w_{bac})$ are estimated by the proportions of foreground and background pixels in the target observation of previous frame, respectively. Then the image patches $\{\boldsymbol{f}^n(\mathbf{s}_t)\}_{n=1,\ldots,N}$ can be classified by

$$\boldsymbol{f}^n(\mathbf{s}_t) \in \begin{cases} w_{\text{target}}, & \text{if } P(w_{\text{target}}|\boldsymbol{f}^n(\mathbf{s}_t)) > P(w_{bac}|\boldsymbol{f}^n(\mathbf{s}_t)) \\ w_{bac}, & \text{otherwise.} \end{cases} \tag{24}$$

Fig. 3 shows the construction of the classifier. For each frame, once target is localized, we randomly extract patches from each of the consecutive target models which are obtained before current frame and use them as positive samples. The patches from surrounding background R_{bac1} are used as negative samples. Using these samples, we can construct an effective classifier to classify $\{\boldsymbol{f}^n(\mathbf{s}_t)\}_{n=1,\ldots,N}$ into foreground and background [Fig. 4(a)]. In the following subsection, the labeled patches sampled from target observation will be used to extract the new target model.

C. Online Extraction of Robust Target Appearance Model

Based on the theoretical principles proposed in Section III-A, the target models should meet the requirements of *adaptability*, *purity* and *time-continuity*. To achieve this goal, we first classify the patches inside tracking window into two classes: foreground and background using (24), and then extract the new target appearance model through two processes: absorption process and rejection process [Fig. 4(a)].

In absorption process, the features with high separability between foreground and background are absorbed in the new target model, to meet the requirement of *adaptability*. Meanwhile, in rejection process, the "dangerous" features which may cause interfusion of background patterns in the new target

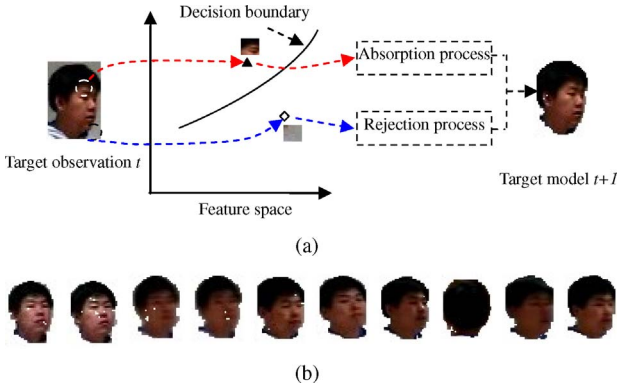


Fig. 4. (a) Illustration of online learning of target model. The patches sampled from target observation are classified into two classes: foreground and background, and then the new target model is extracted through two processes: absorption process and rejection process. (b) Varying target models obtained online through the proposed method. Note that the target models have no gradual drift (without interfusion of background pixels) and can follow up the varying target appearance well.

model are rejected, to meet the requirement of *purity*. The past consecutive target models are also involved in learning the new target model, to meet the requirement of *time-continuity*. The two processes are described in detail as below.

1) *Absorption Process:* For the nth patch $\boldsymbol{f}^n(\mathbf{s}_t)$ sampled from target observation, if $\boldsymbol{f}^n(\mathbf{s}_t) \in w_{\text{target}}$, it means that most of the pixels in this patch belong to foreground. However, we cannot absorb all components of $\boldsymbol{f}^n(\mathbf{s}_t)$ into the new target model, due to the risk of interfusion of immediate surrounding background patterns, although the proportion of background pixels inside this patch may be small. Different from the traditional model updating or learning methods which produce new target model directly using the target observation, the proposed method only absorbs the reliable feature components with high separability into the new target model.

The separability of feature value $h = 1, \ldots, H$ between foreground and surrounding background can be measured by

$$D_t(h) = \frac{\max_m f_{bac1,h}^{m,t}}{\max\{\sum_{k=0}^{K} \lambda_k \max_w f_{\text{target},h}^{w,t-k}, \delta\}} \tag{25}$$

where δ is a small positive value which is designed to prevent division by zero.

Let $1 > \delta_1 > 0$ denotes a threshold value. For patch $\boldsymbol{f}^n(\mathbf{s}_t) \in w_{\text{target}}$, if $D_t(h) < \delta_1$, it means that the corresponding feature value h has high separability and it rarely exists in surrounding background. Then we absorb it in the new target model, and the corresponding feature distribution of the new target model can be computed as

$$f_{\text{target},h}^{t+1} = \frac{1}{N} \sum_{n=1}^{N} k(\|x_n\|^2) f_h^n(\mathbf{s}_t) + \sum_{k=0}^{K} \gamma_k f_{\text{target},h}^{t-k} \tag{26}$$

where $k(x)$ is a convex and monotonic decreasing kernel function, such as Epanechnikov function. x_n is the normalized location of the nth patch inside the tracking window which is centered at zero. The patches farther from the center of tracking window are assigned with smaller weights, because these patches may be "polluted" by background pixels partly. The parameters $\{\gamma_k\}_{k=0,\ldots,K}$ are forgetting factors which are defined in (20). The current target observation has the most important effect on the new target model, and it is intuitively reasonable in practice. The time decaying forgetting factor γ_k controls the influence of the consecutive target models, and it enables the tracker to gradually forget the target appearance which emerged long time ago.

Through the absorption process, the new target model can automatically absorb the features which emerge in the target appearance with high separability between foreground and background. The obtained target model can adaptively follow up the varying target appearance. In the following step, we will introduce a rejection process to keep the purity and completeness of target model.

2) *Rejection Process:* For the nth patch $\boldsymbol{f}^n(\mathbf{s}_t)$ sampled from target observation, if $\boldsymbol{f}^n(\mathbf{s}_t) \in w_{bac}$, it means that most of the pixels in this patch belong to background. However, the new target model cannot arbitrarily reject the patterns which exist in the patches labeled as background, because some features in these patches may also exist in foreground. In order to guarantee the completeness of target model, only the "dangerous" features, which may cause interfusion of background in the new target model, are rejected.

Let $\delta_2 > 0$ denotes a threshold value, and $\delta_2 > \delta_1$. For patch $\boldsymbol{f}^n(\mathbf{s}_t) \in w_{bac}$, if $D_t(h) > \delta_2$, it means that the corresponding feature value h has a high probability of belonging to the surrounding background, and hence an outlier rejection scheme is needed. To keep the time-continuous nature of target model and cope with sudden tracking failure, the distribution of the rejected feature in the new target model can be inherited from the weighted sum of the past consecutive target models. Then the corresponding distribution of the rejected feature in the new target model can be computed as

$$f_{\text{target},h}^{t+1} = \alpha \sum_{k=0}^{K} \lambda_k f_{\text{target},h}^{t-k} \tag{27}$$

where parameter $\alpha \in [0, 1]$ controls the temporal continuity of the new target model. If α is set to zero, it means that the corresponding feature is absolutely excluded from the new target model. Conversely, if $\alpha = 1$, it means that the distribution of the rejected feature is a weighted sum of the distributions of the past consecutive target models. At the beginning of tracking, due to the lack of past target models, α is set to zero, and after a few frames α is set to nonzero to keep the time-continuity property of target model.

Through the rejection process, the "dangerous" features which may cause interfusion of background patterns in the new target model are rejected. The features, which lie between reliable features and "dangerous" features, keep the same distribution values as in the previous target model. The obtained target model is finally normalized to impose the condition $\sum_{h=1}^{H} f_{\text{target},h}^{t+1} = 1$.

Fig. 4(a) illustrates the process of target model learning without any off-line training process and prior knowledge. For each frame, we first classify the patches sampled from target observation into two classes: foreground and background using

the classifier constructed online (Fig. 3). Then two novel processes: absorption process and rejection process are used to extract the new target model. To keep the continuous nature of target model, the past consecutive target models are involved with different weights. Through the proposed online target model learning mechanism, we can get an adaptive, pure and time-continuous target model which can effectively follow up the varying appearance of target and resist gradual drift. Fig. 4(b) shows the variation of the target models with time. Note that there is almost no interfusion of background pixels in each target model, and the target models are able to well represent the varying target appearance.

D. Discussion

Similar to (9), the above online target model learning method works under the assumption that there should be at least one patch sampled from the surrounding background $\{\boldsymbol{f}_{bac1}^{m,t}\}_{m=1,\dots,M}$ which could well represent the interfused background in target observation, that is

$$\exists\ \boldsymbol{f}_{bac1}^{m,t} \quad s.t. \quad \|\boldsymbol{f}_{bac1}^{m,t} - \boldsymbol{f}_{\varepsilon}(\boldsymbol{s}_t)\| \approx 0 \tag{28}$$

where $\boldsymbol{f}_{\varepsilon}(\boldsymbol{s}_t)$ is the interfused background in target observation $\boldsymbol{f}(\boldsymbol{s}_t)$. The failure cases which cannot meet the above assumption will be shown in Section VI-E.

V. Collaborative Target Models Generation and Target Localization

A. Collaborative Target Models Generation

For each frame, target can be localized using the target model obtained in the previous frame, and all parts of target appearance usually have the same significance during tracking. However, feature distributions of target and surrounding background usually overlap, i.e., some features are shared by both target and background. Therefore, it is difficult to separate them clearly, and this will result in decrease of tracking accuracy or even failure in tracking. Furthermore, different parts of target appearance usually have different discriminating abilities during tracking. The parts which have larger difference with surrounding background are more discriminative, and such parts play an important role in developing a robust tracker. Meanwhile, target appearance in consecutive frames usually does not change completely, and has some continuous parts which change slowly with time.

Therefore, based on the target model obtained in the previous section, we introduce a collaboration of two models: dominant model and continuous model, to represent target appearance. The dominant model enhances the feature distributions which can clearly distinguish target from surrounding background, and suppresses the ones which are shared by both target and background. The continuous model is designed to prevent sudden tracking failure and keep the temporal continuity of target model.

For each feature value $h = 1, \dots, H$, a log-likelihood ratio based on the histograms of current target model and local background R_{bac2} is designed to describe the difference between target model and surrounding background

$$L_t(h) = \max_r \max_w \log \frac{\max\{f_{\text{target},h}^{w,t+1}, \delta\}}{\max\{f_{bac2,h}^{r,t}, \delta\}}. \tag{29}$$

The log-likelihood ratio function maps distributions of salient features in target model into positive values and maps distributions of salient features in surrounding background into negative values. Distributions of features which are shared by both target model and background are with small values close to zero. The range of $L_t(h)$ is $(-\infty, +\infty)$. In order to exclude the negative values and prevent too high contrast between different parts of target model, we transform $L_t(h)$ into the range of $[0, 1)$ by

$$\hat{f}_{\text{target},h}^{t+1} = \max\{\frac{2}{\pi} \arctan L_t(h), 0\}. \tag{30}$$

If $\hat{f}_{\text{target},h}^{t+1}$ tends toward zero, it means that the feature value h is shared by both target model and background. We select the salient features to construct a dominant target model

$$f_{\text{target},h}^{D,t+1} = \begin{cases} \frac{1}{C_2} f_{\text{target},h}^{t+1}, & \text{if } \hat{f}_{\text{target},h}^{t+1} \geqslant \delta_T^{t+1} \\ 0, & \text{otherwise} \end{cases} \tag{31}$$

where δ_T^{t+1} is a time-varying threshold value with range of $(0, 1)$, and it is estimated using the naive median estimator $\delta_T^{t+1} = \text{median}(\hat{f}_{\text{target},1}^{t+1}, \dots, \hat{f}_{\text{target},Z}^{t+1})$, where $\{\hat{f}_{\text{target},z}^{t+1}\}_{z=1,\dots,Z}$ are the nonzero components of $\hat{\boldsymbol{f}}_{\text{target}}^{t+1}$. The normalization constant C_2 is derived by imposing the condition $\sum_{h=1}^{H} f_{\text{target},h}^{D,t+1} = 1$. The dominant model has powerful discriminating ability between foreground and background, and it plays an important role in robust tracking. The main advantage of the dominant model is that the tracker can easily find and enhance the salient parts of target in challenging conditions and effectively reduce the influence of environment.

However, the tracker only with a dominant model may capture local parts of target appearance and neglect most of the other parts, and this will result in the decrease of tracking accuracy. Therefore, we introduce a collaborative continuous model to cope with this limitation.

Due to the time-continuous nature, the target appearance should keep some similarity and continuity between consecutive frames. Therefore, a continuous target model is introduced as

$$f_{\text{target},h}^{C,t+1} = \begin{cases} \frac{1}{C_3}(\sum_{k=0}^{K} \gamma_k f_{\text{target},h}^{t-k} + f_{\text{target},h}^{t+1}), & \text{if } \hat{f}_{\text{target},h}^{t+1} < \delta_T^{t+1} \text{ and } \sum_{k=0}^{K} \gamma_k f_{\text{target},h}^{t-k} > \delta_\gamma^{t+1} \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

where the normalization constant C_3 is derived by imposing the condition $\sum_{h=1}^{H} f_{\text{target},h}^{C,t+1} = 1$. The parameter δ_γ^{t+1} is derived by guaranteeing the continuity of target model, and features which rarely exist in past consecutive target models are excluded from the continuous target model. The continuous model improves the temporal continuity of the proposed tracker, and effectively avoids the tracker from capturing

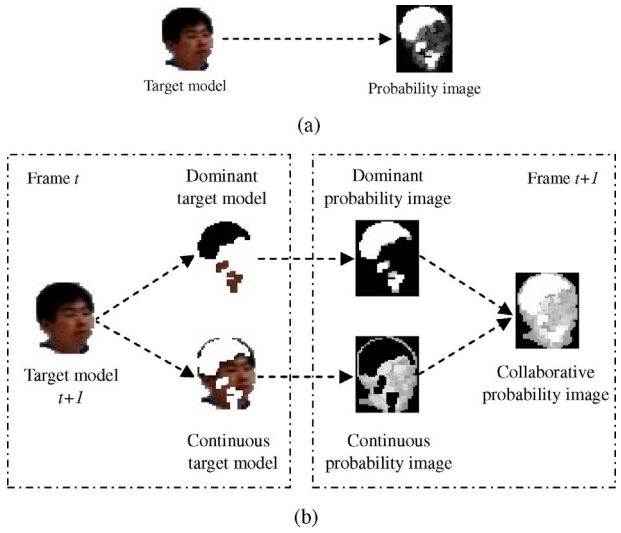


Fig. 5. (a) Probability image generated by the extracted target model with dominant feature selection. (b) Illustration of the generation of the collaborative target models and the collaborative probability image.

local parts of target. Meanwhile, the continuous model can prevent sudden tracking failure, and enhance the robustness and stability of the tracker.

For the input frame $t+1$, these two collaborative models will generate a dominant probability image $p^{D,t+1}$ and a continuous probability image $p^{C,t+1}$ independently

$$p^{D,t+1}(x, y) = f^{D,t+1}_{\text{target},h(x,y)} / \sum_{(x,y)\in ROI} f^{D,t+1}_{\text{target},h(x,y)} \tag{33}$$

$$p^{C,t+1}(x, y) = f^{C,t+1}_{\text{target},h(x,y)} / \sum_{(x,y)\in ROI} f^{C,t+1}_{\text{target},h(x,y)} \tag{34}$$

where $p^{D/C,t+1}(x, y)$ denotes the probability of the pixel in (x, y), and ROI is the region of interest for tracking. $h(x, y)$ is a binning function that maps the color value of (x, y) onto a histogram bin, and $h(x, y) \in \{1, \ldots, H\}$, $H = H_R \cdot H_G \cdot H_B$

$$h(x, y) = h_R(x, y) \cdot H_G \cdot H_B + h_G(x, y) \cdot H_B + h_B(x, y) \tag{35}$$

where h_R, h_G, and h_B denote the bin indices for R, G, and B channels, respectively.

The collaborative probability image can be created through combining the obtained dominant probability image $p^{D,t+1}(x, y)$ and the continuous probability image $p^{C,t+1}(x, y)$

$$p^{\text{Coll},t+1}(x, y) = \frac{\tau_D p^{D,t+1}(x, y) + \tau_C p^{C,t+1}(x, y)}{\sum_{(x,y)\in ROI} (\tau_D p^{D,t+1}(x, y) + \tau_C p^{C,t+1}(x, y))} \tag{36}$$

where $p^{\text{Coll},t+1}$ denotes the collaborative probability image, and τ_D and τ_C are the combining coefficients.

Fig. 5(a) shows the probability image generated by the extracted target model using dominant feature selection. The distributions of salient features are enhanced, and the features with low discriminating abilities are suppressed. Only local parts of target appearance have higher probability values. Fig. 5(b) illustrates the generation of the collaborative target models and the collaborative probability image. Note that all parts of target appearance are with high probability values in the collaborative probability image. For the convenience of description, only the probability images of the region inside tracking window are shown in Fig. 5.

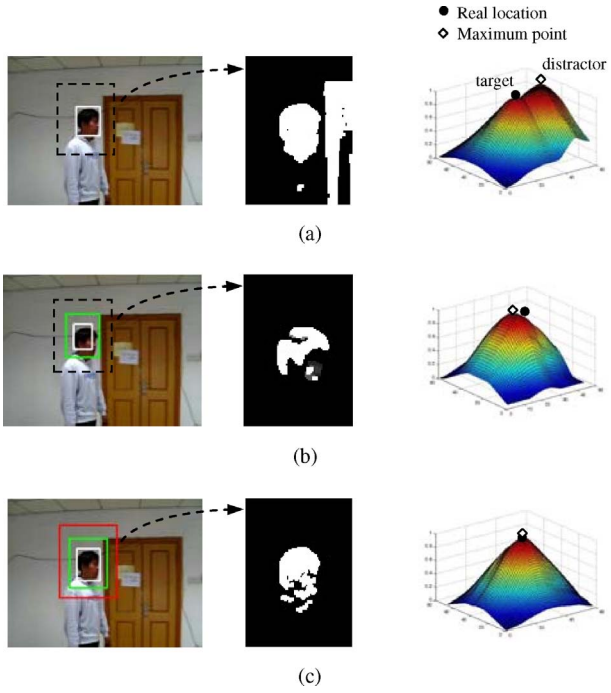


Fig. 6. Comparison between three different algorithms. From left to right in each row: the tracking result, the probability image, and the similarity surface of the corresponding region. (a) Results of the traditional mean-shift algorithm. The tracker will mistakenly capture the distractor in next frame. (b) Results of the improved mean-shift method with dominant feature selection. (c) Results of the proposed method. In the similarity surfaces, "•" denotes the real location of the target, and "◇" denotes the maximum point of the similarity surface.

B. Target Localization

The aforementioned online target model learning (Section IV) and collaborative models generation (Section V-A) mechanisms are embedded in an adaptive tracking system as shown in Fig. 1. For each frame, two probability images are generated independently using the collaborative target models obtained in the previous frame. Then these two probability images are combined to create a collaborative probability image [using (36)], where the value of each pixel characterizes the probability that the pixel belongs to target.

The tracked object is localized using mean-shift algorithm [6], and the initial position of target is determined in previous frame. The mean-shift process finds the nearest local maximum in the collaborative probability image iteratively, and converges to the local optimal location of target. Besides location, the scale of target is also adjusted adaptively with time, and the current state (position and scale) of target is saved as the initial state of next frame. Once target is localized, new patches are extracted from the past and current target models and also the current surrounding local backgrounds, to learn the new target appearance model and generate the collaborative target models for next frame.

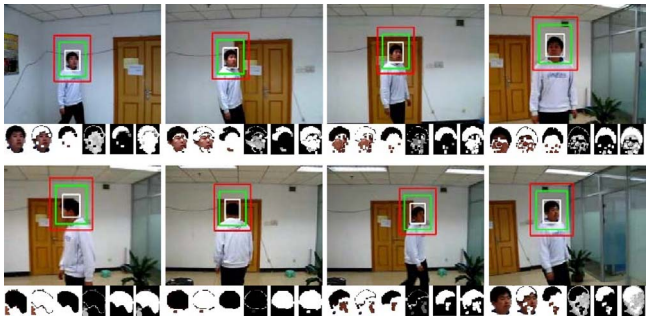


Fig. 7. Tracking results under severe changes of viewpoint and similar background. The first row of each panel shows the tracked object (enclosed in the *white rectangle*), and the second row shows (from left to right) the current target model, continuous model, dominant model, continuous probability image, dominant probability image, and the collaborative probability image. The frames 43, 72, 93, 137, 160, 185, 220, and 246 are shown.

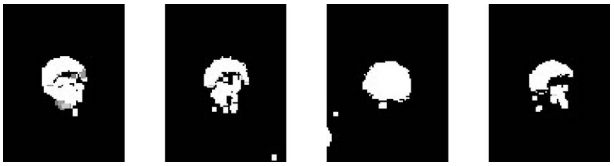


Fig. 8. Collaborative probability images of the corresponding regions (inside the red rectangle) for frames 72, 93, 185, and 220.

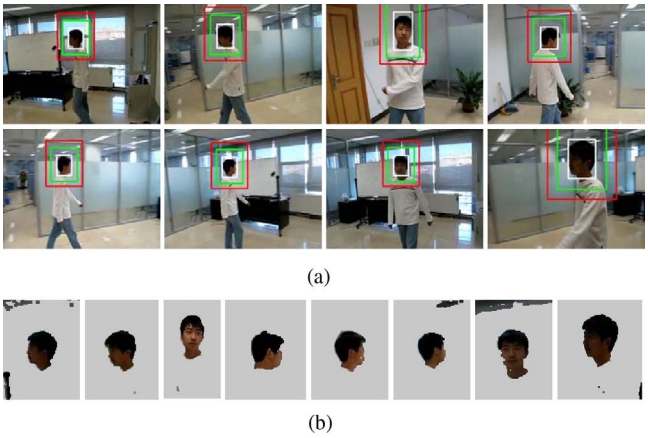
Fig. 6 shows the comparison between three different algorithms. The traditional mean-shift algorithm without feature selection [Fig. 6(a)] is sensitive to background disturbance, and the tracker will mistakenly capture the distractor (the door which has a similar color as that of skin) in the next frame. The improved mean-shift algorithm with dominant feature selection [Fig. 6(b)] can effectively suppress the disturbance of background. However, the tracker tends to capture target's local part (the hair) which has higher discriminating ability between foreground and background, and this decreases the accuracy of tracking. The proposed method [Fig. 6(c)] can effectively suppress the disturbance of background and capture the whole target accurately.

VI. Experiments

This section shows the performance of the proposed method under different challenging conditions, such as, rotation of target, abrupt changes of lighting, cluttered environment, and severe occlusion. The proposed algorithm is tested on a computer system with the configuration of Quad-Core CPU and 2G RAM. The resolution of the image used for tracking is 320×240. The average processing speed of the proposed algorithm is about 25 frames/s, which meets the real-time requirement.

A. Severe Changes of Viewpoint and Similar Background

We first evaluate performance of the proposed tracker under severe changes of viewpoint and similar background. Fig. 7 shows one of the experimental results, where the first row of each panel shows the tracked objects (enclosed with white rectangles, the green and the red rectangles are the bounding



(a)

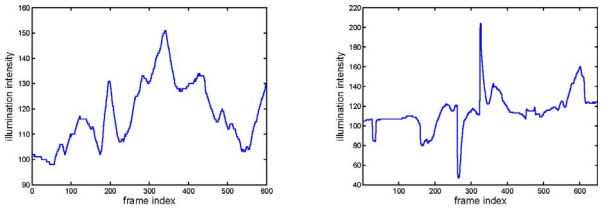
(b)

Fig. 9. (a) Tracking results under large scale and pose changes of target, and slow variations of illumination. (b) Extracted targets. The frames 79, 170, 243, 314; 367, 433, 504, and 560 are shown.

boxes of background R_{bac1} and R_{bac2}, respectively), and the second row shows the current target model, continuous target model, dominant target model, continuous probability image, dominant probability image, and the collaborative probability image. For the convenience of description, only the probability images of the region inside tracking window are shown. Note that the proposed tracker is able to track object undergoing rotation, and the target model obtained online without any off-line learning process is pure and discriminative, although the background door has a similar color as that of skin. The hair has high discriminating ability in the scene, especially when target is in front of the yellow door. In the collaborative probability image, most of the pixels in target region are with high values, and the tracker can accurately capture the whole head. Fig. 8 shows the collaborative probability images of the corresponding regions (inside the red rectangle) of frames 72, 93, 185, and 220, and the door whose color is similar to skin is excluded from the probability images. The proposed method can effectively suppress the disturbance of background, meanwhile it prevents too high probability value in partial target appearance.

B. Large Scale Changes and Slow Variations of Illumination

Fig. 9(a) shows the performance of the proposed method under large scale changes of target and slow variations of illumination. The target moves between the dark and bright areas, in which half of the lamps in the room are turned off and others are turned on. Fig. 10(a) shows the variation of illumination intensity in this experiment, and the illumination intensity is measured using the average intensity of all pixels in each frame. Note that illumination intensity changes with time with the maximum of 150 and minimum of 98. There are also large pose and scale variations in target appearance, as well as fast camera motion. The proposed method is able to capture the target head throughout the sequence without gradual drift. The proposed method can also extract the target from background effectively and accurately, although the target appearance and background change fast with time. Fig. 9(b) shows the extracted targets of experiment shown in Fig. 9(a). In contrast, most tracking algorithms in dynamic scenes are



(a) (b)

Fig. 10. Varying illumination intensities of (a) experiment shown in Fig. 9, (b) experiment shown in Fig. 13. The illumination intensity is measured using the average intensity of all pixels in each frame.

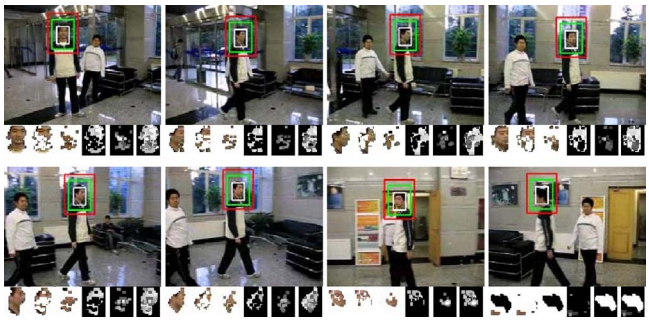


Fig. 11. Tracking results under cluttered background. The frames 37, 72, 113, 150, 187, 276, 486, and 530 are shown.

not expected to extract target well from background, especially under challenging conditions, such as lighting changes and complex environment.

C. Cluttered Background

We also evaluate the proposed tracker under cluttered environment. Fig. 11 shows the performance of the proposed method in a hall where there are plenty of sets of furniture which have colors similar to target head, and the lighting direction and intensity are varying with time. The challenges also include the existence of cluttered woods outside the window, the potted plants on the wall sill, and the interference from other person. These challenges may cause decrease of tracking accuracy, due to the poor separability between target and the surrounding background. Nevertheless, experiments demonstrate that the proposed tracker can effectively handle these challenges, and capture the target with high accuracy. The target model has powerful distinguishing ability at the cost of completeness. In this experiment, skin color features whose distributions can well discriminate between the tracked head and surrounding background are absorbed in target model automatically in most frames. However, hair color features rarely appear in target model, due to the existence of objects in surrounding background which have colors similar to hair.

D. Other Challenging Conditions and Comparisons

To evaluate the empirical performance of the proposed tracker, we also test the algorithm under other challenging conditions, such as complex motion of target, camera vibration, abrupt illumination changes, and severe occlusion. The tracking results are compared with other methods, such as the elliptical head tracker [39], the fragments-based tracker [40], the method proposed in [22], and mean-shift [6] with target

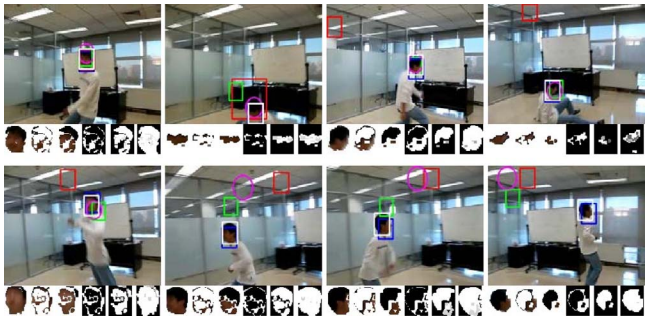


Fig. 12. Comparison of tracking results under complex motion of target (such as jump, abrupt acceleration), illumination and viewpoint changes, and similar background, between different trackers: the proposed tracker (indicated with a *white box*), the elliptical head tracker [39] (indicated with a *pink ellipse*), the fragments-based tracker [40] (indicated with a *green box*), Collins' method [22] (indicated with a *blue box*), and the mean-shift tracker with target model updating (indicated with a *red box*). The frames 61, 80, 253, 276, 399, 419, 508, and 529 are shown.

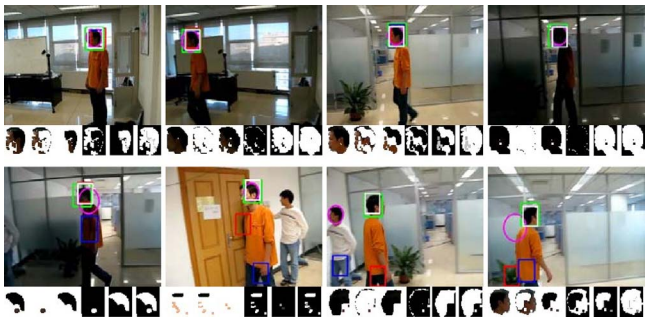


Fig. 13. Comparison of tracking results under abrupt illumination changes between different trackers. The frames 61, 165, 258, 264, 277, 323, 603, and 646 are shown.

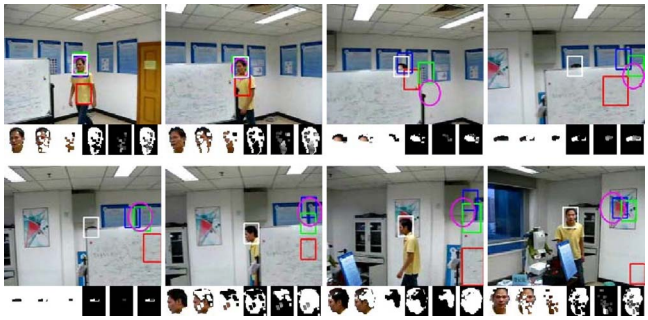


Fig. 14. Comparison of tracking results under severe occlusion between different trackers. The frames 81, 120, 139, 178, 199, 209, 243, and 311 are shown.

model updating in which target model is updated through the weighted sum of the initial target model and the current target observation.

1) *Complex Motion of Target and Existence of Distractor:* Fig. 12 shows the tracking results of different tracking methods under complex motion of tracked person, changes of illumination and viewpoint, and similar background. The tracked person runs in the room with sudden stop, abrupt acceleration and jump, and the mobile camera also reacts quickly to follow the moving person. These result in degraded image quality due to motion blur, abrupt changes in position, and scale of target head between two adjacent frames. Furthermore, the pose and lighting variations combined with complex background

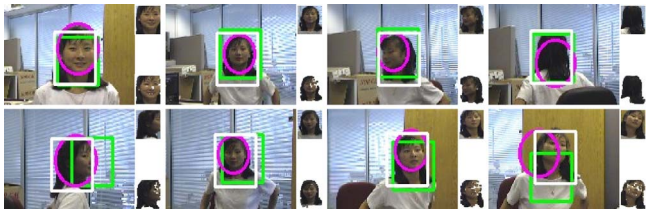


Fig. 15. Comparison of the proposed tracker (indicated with a *white box*) with the elliptical head tracker in [39] (indicated with a *pink ellipse*) and the fragments-based tracker [40] (indicated with a *green box*) using the public sequence.

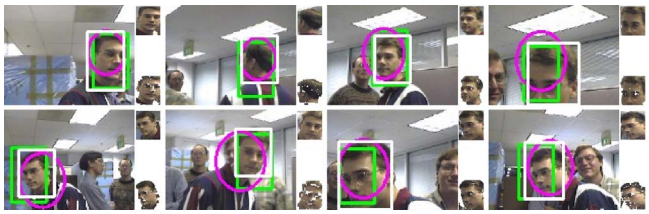
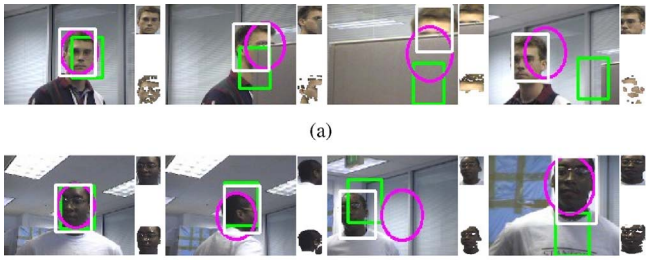


Fig. 16. Comparison of the proposed tracker (indicated with a *white box*) with the elliptical head tracker in [39] (indicated with a *pink ellipse*) and the fragments-based tracker [40] (indicated with a *green box*) using the public sequence.



(a)

(b)

Fig. 17. Comparison of the proposed tracker (indicated with a *white box*) with the elliptical head tracker in [39] (indicated with a *pink ellipse*) and the fragments-based tracker [40] (indicated with a *green box*) using the public sequences. (a) Target is severely occluded. (b) Scale and pose of target are changing quickly with time.

make the tracking task rather challenging. The mean-shift tracker with target model updating loses the target and wrongly captures the desk whose color is similar to hair (frame 80). The elliptical head tracker [39] and fragments-based tracker [40] lose the target when the target jumps. The tracker in [22] and the proposed tracker capture the target well, and the proposed method obtains a higher accuracy than method in [22]. The target models obtained using the proposed method are pure (without interfusion of background pixels) and complete (all parts of the tracked head are involved in target model) most of the times, except when the person falls down on the floor (e.g., frames 80 and 276, Fig. 12). In frame 80 and frame 276, it is difficult to discriminate between the tracked head and the surrounding desk, and the discriminating ability of target model is achieved at the cost of completeness.

2) *Abrupt Illumination Changes:* Fig. 13 demonstrates the tracking results under abrupt illumination changes. During tracking, some or all of the lamps in the room are suddenly turned on or turned off. Fig. 10(b) shows the abrupt illumination changes in this experiment. In the beginning, half of the eight lamps in the room are turned on, and the tracker

Fig. 18. Comparison results of Human body tracking using the proposed tracker (indicated with a *red box*) and the fragments-based tracker [40] (indicated with a *green box*).

is initialized. In frame 159, two of the remaining lamps are turned off, and in frame 205 these two lamps are turned on again. In frame 262, all lamps in the room are turned off suddenly, and it is difficult to distinguish the tracked head from background, even for human eyes. All lamps are turned on suddenly in frame 325, and after about 20 frames half of the lamps are turned off again. Furthermore, both the shirt of the tracked person and the door have colors similar to skin, especially under strong lighting condition. The elliptical head tracker [39], the mean-shift tracker with target model updating and the tracker in [22] both lose the target when illumination changes abruptly. The proposed method and the fragments-based method [40] are able to track the head throughout the sequence, and the proposed method obtains a higher accuracy. Note that in the proposed method the target model is well extracted even when all lamps are turned off from frame 262 to frame 280, and the continuous model of target only contains the profile of the tracked head in this situation (frame 264, Fig. 13), due to the abrupt variation of target appearance.

3) *Severe Occlusion:* Fig. 14 shows the comparison results under severe occlusion. During tracking, the target head is severely occluded, and only a small part of the target is visible. When severe occlusion happens, there is little information of target for tracking, which makes tracking more challenging. The other four trackers all lose the target when severe occlusion happens. Only the proposed method tracks the target steadily and accurately throughout the sequence, even when there is only a very small part of target appearance is visible. Note that the proposed method can adaptively extract the nonoccluded part of target (frames 139, 178, and 199), and the immediate surrounding background are completely suppressed in the final probability map, which makes the tracker more robust, although little information of target can be used.

E. Evaluation Using Public Sequences

The proposed algorithm is also evaluated using the publicly available video sequences and compared with other tracking algorithms.[1]

Figs. 15–17 show the tracking results compared with the elliptical head tracker [39] and the fragments-based tracker [40]. The tracked target and the obtained target model of the proposed method are shown at the upper right corner and lower

[1]Data and code are downloaded from the websites at http://www.ces.clemson.edu/stb/research/headtracker/ and http://www.cs.technion.ac.il/amita/fragtrack/fragtrack.htm.

right corner, respectively. In Fig. 15, the scale and pose of target are varying with time, and the door has a color similar to face. In Fig. 16, there are other persons in the background, and the target scale and pose are also varying with time. In Fig. 17(a) the target is severely occluded, and in Fig. 17(b) the scale and pose of target are changing quickly with time. The proposed method can effectively cope with these challenging conditions, and obtain a higher tracking accuracy.

We also test the proposed method in outdoor environment to track the human body. Fig. 18 shows the comparison results with the fragments-based tracker [40]. The proposed method can track the human body under partial occlusion and cluttered background with a higher accuracy.

F. Discussion

The proposed target model learning and generation method can effectively adapt to the varying appearance of target and resist gradual drift, and the experiments have demonstrated its robustness under various challenging conditions. Nevertheless, some issues still need to be further discussed.

1) *The Incompleteness of Target Model Under the Condition of Similar Background:* The proposed method can effectively cope with the situation that there are similar objects in the surrounding background (Figs. 6–8). However, when the background is very similar to some parts of target, the classifier in (24) will classify these parts into background, and the features in these parts will not be absorbed in the new target model. This is helpful for improving the discriminating ability of the target model. However, the rejection of the features that exist in both the surrounding background and the tracked object will result in an incomplete target model, i.e., there are some holes in the obtained target models (Fig. 7, frame 93) or the target model only contains some parts of the real target appearance (Fig. 12, frames 80 and 276). Therefore, it is a tradeoff between the discriminating ability and the completeness of target model.

Single cue is usually not enough for discrimination under all circumstances, and there is no cue which is always more effective than others under all conditions, especially under the condition of dynamic scene and similar background. The proper integration of multiple cues, such as color, texture and shape, can effectively help us to obtain a more complete target model and meanwhile keep its discriminating ability.

2) *Failure Situation:* There are two kinds of failure cases in our tests. The first situation is when the surrounding background R_{bac1} cannot well represent the background patterns interfused in the target observation, e.g., there are other nontarget objects in the background which are initially occluded by the target [it does not meet the assumption in (28)]. In this case, the immediate surrounding background R_{bac1} has different feature distributions with the interfused background in target observation, and features from these new appearing objects will be absorbed into the new target model. Fig. 19 shows the model drift under this condition. A blue block which is a little smaller than the target head is initially occluded by the head. With the slow appearing of the occluded block, the target model gradually absorbs the features of this block by mistake. Although the tracker finally escapes from the trap, it is still a problem needing further treatment. To cope with this situation, the solution includes the integration of multiple cues, such as texture and shape, and the addition of spatial information in target representation.

Fig. 19. Possible failure situation: there are other nontarget objects in the background which are initially occluded by the target. The frames 8, 17, 19, and 41 are shown.

The other failure situation is when the target is completely occluded, although the proposed method is able to deal well with severe occlusion (Fig. 14). In this case, there is no target observation information for target model learning, and the tracker will capture other objects. To cope with this situation, future work will concentrate on using of motion information of target and the combination of detection methods to reacquire the target.

VII. Conclusion

This paper proposed a novel adaptive visual tracking algorithm by learning and generating pure, adaptive, and time-continuous target models online. First, an effective classifier was constructed online to classify the patches sampled from target observation into foreground and background. Then two novel processes the absorption process and the rejection process were introduced to extract the new target model. Based on the extracted target model, a combination of dominant model and continuous model was proposed to minimize the influence of background and prevent sudden variation of target model. The collaborative target models will be used to localize target in the next frame and the proposed model learning and generation mechanisms are finally embedded in an adaptive tracking system. Performance evaluated on sequences captured by a mobile camera and also publicly available sequences demonstrates that the proposed method can effectively cope with the drastic variation of target appearance and resist gradual drift under different challenging conditions.

The future work includes the integration of multiple cues and spatial information in target representation to cope with the situation that there are other nontarget objects in the background which are initially occluded by the target. Future work also includes the combination of motion information of target and detection methods to reacquire the target when complete occlusion happens.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions for improving the quality of this paper.

References

[1] A. Yilmaz, X. Li, and M. Shan, "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1531–1536, Nov. 2004.

[2] Y. Rathi, N. Vaswani, and A. Tannenbaum, "A generic framework for tracking using particle filter with dynamic shape prior," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1370–1382, May 2007.
[3] A. Mansouri, "Region tracking via level set PDEs without motion computation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 947–961, Jul. 2002.
[4] H. T. Nguyen and A. W. M. Smeulders. "Fast occluded object tracking by a robust appearance filter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 1099–1104, Aug. 2004.
[5] C. Gentile, O. Camps, and M. Sznaier. "Segmentation for robust tracking in the presence of severe occlusion," *IEEE Trans. Image Process.*, vol. 13, no. 2, pp. 166–178, Feb. 2004.
[6] D. Comaniciu, V. Ramesh, and P. Meer. "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 25, no. 5, pp. 564–577, May 2003.
[7] C. Shen, M. J. Brooks, and A. Hengel, "Fast global kernel density mode seeking: Applications to localization and tracking," *IEEE Trans. Image Process.,* vol. 16, no. 5, pp. 1457–1469, May 2007.
[8] B. Han, D. Comaniciu, Y. Zhu, and L. S. Davis, "Sequential kernel density approximation and its application to real-time visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 30, no. 7, pp. 1186–1197, Jul. 2008.
[9] Z. Fan, M. Yang, and Y. Wu, "Multiple collaborative kernel tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 29, no. 7, pp. 1268–1273, Jul. 2007.
[10] M. Isard and A. Blake, "CONDENSATION: Conditional density propagation for visual tracking," *Int. J. Comput. Vis.,* vol. 29, no. 1, pp. 5–28, 1998.
[11] H. Wang, D. Suter, K. Schindler, and C. Shen, "Adaptive object tracking based on an effective appearance filter," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 29, no. 9, pp. 1661–1667, Sep. 2007.
[12] Y. Lao, J. Zhu, and Y. F. Zheng, "Sequential particle generation for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 19, no. 9, pp. 1365–1378, Sep. 2009.
[13] P. Wang and H. Qiao, "Adaptive probabilistic tracking with reliable particle selection," *Electron. Lett.,* vol. 45, no. 23, pp. 1160–1161, Nov. 2009.
[14] M. J. Black and A. D. Jepson, "Eigen tracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comput. Vis.,* vol. 26, no. 1, pp. 63–84, 1998.
[15] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Comput. Vis. Image Underst.*, vol. 99, no. 3, pp. 303–331, 2005.
[16] D. A. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, no. 1, pp. 125–141, May 2008.
[17] J. Ho, K. C. Lee, M. H. Yang, and D. Kriegman, "Visual tracking using learned linear subspace," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun.–Jul. 2004, pp. 782–789.
[18] A. Elgammal and C. S. Lee, "Tracking people on a torus," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 31, no. 3, pp. 520–538, Mar. 2009.
[19] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Int. J. Comput. Vis.,* vol. 73, no. 1, pp. 41–59, 2007.
[20] F. M. Noguer, A. Sanfeliu, and D. Samaras, "Dependent multiple cue integration for robust tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 30, no. 4, pp. 670–685, Apr. 2008.
[21] K. Shearer, K. D. Wong, and S. Venkatesh, "Combining multiple tracking algorithms for improved general performance," *Pattern Recognit.,* vol. 34, no. 6, pp. 1257–1269, Jun. 2001.
[22] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection Of discriminative tracking features," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 27, no. 10, pp. 1631–1643, Oct. 2005.
[23] J. Wang and Y. Yagi, "Integrating color and shape-texture features for adaptive real-time object tracking," *IEEE Trans. Image Process.,* vol. 17, no. 2, pp. 235–240, Feb. 2008.
[24] B. Han and L. Davis, "Object tracking by adaptive feature extraction," in *Proc. Int. Conf. Image Process.,* vol. 3. 2004, pp. 1501–1504.
[25] A. P. Leung and S. Gong, "Online feature selection using mutual information for real-time multi-view object tracking," in *Proc. ICCV Workshop AMFG,* 2005, pp. 184–197.
[26] H. Stern and B. Efros. "Adaptive color space switching for face tracking in multi-colored lighting environments," in *Proc. 5th IEEE Int. Conf. Auto. Face Gesture Recognit.*, May 2002, pp. 249–254.
[27] H. T. Chen, T. L. Liu, and C. S. Fuh, "Probabilistic tracking with adaptive feature selection," in *Proc. 17th Int. Conf. Pattern Recog.,* vol. 2. Aug. 2004, pp. 736–739.
[28] S. Avidan, "Support vector tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 26, no. 8, pp. 1064–1072, Aug. 2004.
[29] H. T. Nguyen and A. W. M. Smeulders, "Robust tracking using foreground-background texture discrimination," *Int. J. Comput. Vis.,* vol. 69, no. 3, pp. 277–293, 2006.
[30] S. Avidan, "Ensemble tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 29, no. 2, pp. 261–271, Feb. 2007.
[31] Y. Lei, X. Ding, and S. Wang, "Visual tracker using sequential Bayesian learning: Discriminative, generative, and hybrid," *IEEE Trans. Syst., Man, Cybern. Part B-Cybern.,* vol. 38, no. 6, pp. 1578–1591, Dec. 2008.
[32] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.,* Aug. 2009, pp. 983–990.
[33] L. Lu and G. D. Hager, "A nonparametric treatment for location/segmentation based visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.,* Jun. 2007, pp. 1–8.
[34] I. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 26, no. 6, pp. 810–815, Jun. 2004.
[35] D. Schreiber, "Robust template tracking with drift correction," *Pattern Recognit. Lett.*, vol. 28, no. 12, pp. 1483–1491, 2007.
[36] J. Tu, H. Tao, and T. Huang, "Online updating appearance generative mixture model for mean-shift tracking," *Mach. Vis. Appl.,* vol. 20, no. 3, pp. 163–173, 2009.
[37] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 25, no. 10, pp. 1296–1311, Oct. 2003.
[38] P. Wang and Q. Ji, "Robust face tracking via collaboration of generic and specific models," *IEEE Trans. Image Process.,* vol. 17, no. 7, pp. 1189–1199, Jul. 2008.
[39] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.,* Jun. 1998, pp. 232–237.
[40] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.,* Jun. 2006, pp. 798–805.

Peng Wang (M'10) received the B.Eng. degree in electrical engineering and automation from Harbin Engineering University, Harbin, China, in 2004, the M.Eng. degree in automation science and engineering from the Harbin Institute of Technology, Harbin, in 2007, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010.

He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. He has worked on computer vision, image processing, and intelligent robots. His current research interests include computer vision, human–robot interaction, high precision sensing and control, and visual perception.

Hong Qiao (SM'06) received the B.Eng. degree in hydraulics and control and the M.Eng. degree in robotics from Xian Jiaotong University, Xian, China, the M.Phil. degree in robotics control from the Industrial Control Center, University of Strathclyde, Strathclyde, U.K., and the Ph.D. degree in robotics and artificial intelligence from De Montfort University, Leicester, U.K., in 1995.

She was a University Research Fellow with De Montfort University from 1995 to 1997. She was a Research Assistant Professor from 1997 to 2000 and an Assistant Professor from 2000 to 2002 with the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong. Since January 2002, she has been a Lecturer with the School of Informatics, University of Manchester, Manchester, U.K. Currently, she is a Professor with the Laboratory of Complex Systems and Intelligent Science, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her current research interests include information-based strategy investigation, robotics and intelligent agents, animation, machine learning (neural networks and support vector machines), and pattern recognition.

Dr. Qiao was a member of the Program Committee of the IEEE International Conference on Robotics and Automation from 2001 to 2004. She is currently the Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B and the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING.