Part-based Online Tracking with Geometry Constraint and Attention Selection

Jianwu Fang, Qi Wang, and Yuan Yuan, *Senior Member, IEEE*

***Abstract*—Visual tracking in condition of occlusion, appearance or illumination change has been a challenging task over decades. Recently, some online trackers, based on the detection by classification framework, have achieved good performance. However, problems are still embodied in at least one of the three aspects: 1) tracking the target with a single region has poor adaptability for occlusion, appearance or illumination change; 2) lack of sample weight estimation, which may cause overfitting issue; and 3) inadequate motion model to prevent target from drifting. For tackling the above problems, this paper presents the contributions as follows: 1) a novel part-based structure is utilized in the online AdaBoost tracking; 2) attentional sample weighting and selection is tackled by introducing a weight relaxation factor, instead of treating the samples equally as traditional trackers do; and 3) a two-stage motion model, multiple parts constraint, is proposed and incorporated into the part-based structure to ensure a stable tracking. The effectiveness and efficiency of the proposed tracker is validated upon several complex video sequences, compared with seven popular online trackers. The experimental results show that the proposed tracker can achieve increased accuracy with comparable computational cost.**

***Index Terms*—Attention selection, multiple parts constraint, object tracking, online AdaBoost (OAB), relaxation factor.**

I. Introduction

ROBUST visual tracking has many applications in the field of computer vision, such as face recognition [1], [2], traffic monitoring [3], [4], and human behavior analysis [5], [6]. Though many efforts [7]–[9] have been spent on this topic, it remains a challenge on the accurate localization and tracking for a target of interest. The difficulty is mainly caused by the influence of the dynamic scene, that is, constant occlusion, alternating appearance, and illumination.

Recently, some online learning trackers based on detection by classification, such as online AdaBoost tracking (OAB) [10] and multiple instances learning tracking (MIL) [11], have been established to restrain from the influence of the above factors. However, these online trackers still have some limitations: 1) modeling the object appearance with a single region has poor adaptability; 2) only the current sample in every frame is used to contribute the classification, which is easy to yield an overfitting issue; and 3) without consideration for a reasonable motion model for trajectory constraint.

For tackling the above issues, this paper addresses them from seeking for an adaptive appearance model, a reliable data association strategy and a reasonable motion model for trajectory constraint. Our work is based on the framework of OAB [10], but for the improvement of OAB, contributions are made to the above three aspects.

A. Literature Review

For the appearance model, let us start from the mean shift [12] based trackers. Within this field, many trackers usually track the target with a single region [13]–[15], leading to a poor adaptability for occlusion. To solve that, some literature [5], [16], [17] consider the object as a part-based type, which has been established to demonstrate a superior performance. As for these part-based trackers, the essential point is to seek an adequate part structure. However, most of the part-structures (i.e., point-based [16], [18]–[20] or discriminative part [17], [21]) are not effective for general applications. For example, Frag [17] represents template object by multiple image fragments or patches. However, the fragments or patches are easy to roam away around the ones with similar appearance cues, and prone to cause drift. For a general part structure, Maggio and Cavallaro [22] divide object into four nonoverlapping rectangle parts, but it conducts the object searching by mean shift, which has poor adaptability for appearance and illumination change.

With respect to the online data association strategy, only the current tracked region is used and the importance of the region lacks updating, which may cause localization shift in condition of occlusion. Actually, it is known that human perception can chase the object excellently even with an occlusion. The main reason is that human can omit the other scenes occluding the original object. Based on that, the concept of reweighting for image patches has been presented in some trackers [21], [23]–[27]. Among them, the method of Yang *et al.* [23] selects attentional regions (AR) like salient image regions in the first step, and then fuses the AR clusters to separate the target from background. The weighting strategy for the

Manuscript received February 20, 2012; revised March 6, 2013 and May 16, 2013; accepted September 5, 2013. Date of publication November 14, 2013; date of current version May 2, 2014. This work was supported in part by the National Basic Research Program of China (Youth 973 Program) under Grant 2013CB336500; in part by the National Natural Science Foundation of China under Grants 61172143, 61105012, and 61379094; and in part by the Natural Science Foundation Research Project of Shaanxi Province under Grant 2012JM8024. This paper was recommended by Associate Editor Q. Tian.

J. Fang and Y. Yuan are with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi 710119, China (e-mail: fangjianwu@opt.ac.cn; yuany@opt.ac.cn).

Q. Wang is with the Northwestern Polytechnical University, Xi'an, China (e-mail: crabwq@gmail.com).

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.2013.2283646

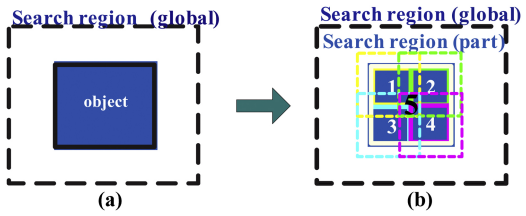


Fig. 1. Part structure utilized in this paper. (a) Original global structure in OAB. (b) Part-based structure presented in this paper.

attentional regions is decided by the distinctiveness of regions. Grabner *et al.* [24] conduct the reweighting of patches (feature points actually) by modeling the motion of points. In [21], the patches' smoothness and steepness are the main aspects for their weights. These reweighting approaches, most of the time, pay more attention to the more discriminative patches. But the discriminative patches (other scene occluding the true object) could gradually replace the original patches in condition of occlusion and background clutter, and the drift occurs if the these strategies of reweighting are adopted.

In terms of object trajectory constraint, these online trackers have no adequate motion model to restrict the object trajectory. For constructing a trajectory constraint, direction consistency [28], [29] and displacement consistency [11] are usually assumed. For example, motion model is constructed using smoothness of direction and speed at adjacent two frames in [28]. Babenko *et al.* [11] predict the new location with a fixed circle at consecutive two frames. But two frame-based motion models are easy to be blurred by an abnormal frame. To avoid that, Kwon and Lee [29] build a motion model based on a cluster strategy k-harmonic means (KHM) [30], which is a Gaussian filter for trajectory and calculates the main direction of the lasted five frames. Although KHM shows a superior performance, it is prone to be influenced by the outliers.

B. Contributions

In view of the above points, this paper proposes a generic tracking algorithm based on geometry constraint and attention selection. The geometry constraint is constituted by a multiple parts constraint (MPC), and the attention selection is implemented with a weight relaxation factor (WRF). Thus the proposed tracker is named as part-based online AdaBoost tracking with geometry constraint MPC and attention selection (WRF) (P-OAB-MW). The detailed contributions are presented as the following threefold.

1) Introduce a generic part-based structure to OAB. This structure is similar to [22]. However, different from [22], this paper conducts the object searching by OAB. The part structure is shown in Fig. 1. Meanwhile, different parts of target can be assigned with different weights within the part-based online AdaBoost tracker (P-OAB).

2) Introduce a new attention selection concept for sample weighting. The more importance the sample gains from previous iterations, the more attention it would be paid in the next iteration. This mechanism is achieved by the consideration that the tracking process can be considered as a combination of different scene conditions (normal conditions and abnormal conditions). Usually, the abnormal conditions are the low frequent[1] sections in the tracking process. Consistently, the samples (tracked regions) collected from these conditions are low frequent ones. In the field of classification, the low frequent samples [31] are unrepresentative and easy to cause an overfitting issue. Based on this *priori* and different from the sample reweighting mechanisms of [21], and [23]–[27], this paper derives a relaxation factor (RF) [32] to lower the initial weight of samples collected from abnormal conditions. In this procedure, a WRF is constructed.

3) Present an MPC for defining the relationship among part-structures. In normal conditions, each part conducts its data association independently. When occlusion, appearance or illumination change occurs, the stable part(s) could recover the otherwise drifting part(s) by the MPC. This principle ensures that the target could not be lost even in complex scenes.

The remainder of this paper is organized as follows. In Section II, the algorithm of P-OAB-MW is introduced. In Section III, online boosting for feature selection is reviewed and attention selection concept by means of WRF is proposed to adjust matching confidence. In Section IV, MPC for localization is presented. Section V presents the experiments. Section VI discusses several issues related to the proposed tracker. Section VII concludes this paper.

II. System Overview

The general framework of the proposed P-OAB-MW is illustrated by the diagram shown in Fig. 2. It is generally divided into three parts: 1) initialization of target part; 2) online boosting for feature selection embedded with WRF; and 3) target localization using MPC.

For the initialization of object parts, denote the part structure as $\mathbf{P}=\{p_j, j = 1, ..., 4\}$. With the part-based structure $\mathbf{P}$, online boosting for feature selection is conducted for every part p_j. The outputs are the matching confidence $conf_{j,t}$ of OAB and the displacement of target $\{\Delta_{x,j,t}, \Delta_{y,j,t}\}$, where $conf_{j,t}$ is the consistency measure of predefined target template and the observation, $\Delta_{x,j,t}$ and $\Delta_{y,j,t}$ represent the shift of target in direction x and y, respectively. By evaluating the priori $conf_{j,t-1}$, the initial sample weight at time t is updated using WRF to restrain from the overfitting issue. The detailed strategy is introduced in Section III-C.

For the localization of target $\mathbf{X}_t$, it is constrained by MPC. MPC can be expressed in two stages.

1) At time t, determine the reliability $r_{j,t}$ of p_j, where $r_{j,t}$ represents the feasibility of estimated target location and the moving trajectory of the j^{th} part, and is modeled by the posterior probability $p(\mathbf{X}_{j,t}^{(l_j)}|\mathbf{Y}_{j,1:t})$. The posterior probability is constructed by the KHM and $conf_{j,t}$, where $\mathbf{X}_{j,t}^{(l_j)}$ and $\mathbf{Y}_{j,1:t}$ respectively represents the target state and observations up to time t, and l_j is the index of candidate searching patches for p_j. Consistently, the estimated state $\hat{\mathbf{X}}_{j,t}$ of p_j is set as $\arg\max_{l_j} p(\mathbf{X}_{j,t}^{(l_j)}|\mathbf{Y}_{j,1:t})$. If $r_{j,t} = p(\hat{\mathbf{X}}_{j,t}|\mathbf{Y}_{j,1:t}) > T$, p_j is reliable,

[1]Low frequent means that the condition, such as occlusion, occurs occasionally and the corresponding tracked region has low frequency in the whole tracking process.

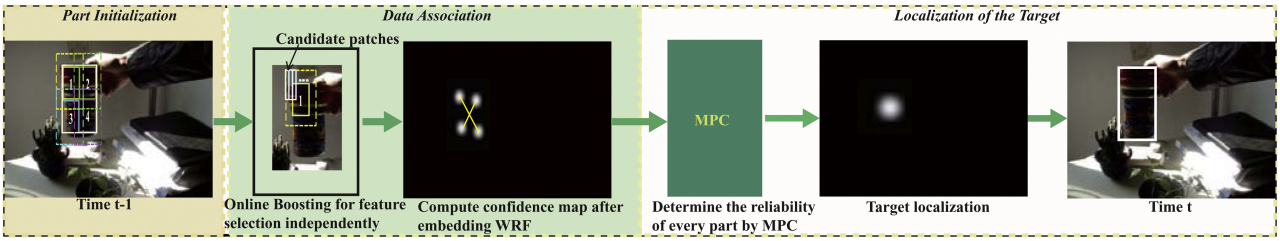


Fig. 2. General framework of P-OAB-MW.

Algorithm 1 P-OAB-MW

Input: $\mathbf{P}=\{p_j, j=1,...,4\}$, time index t, $S=0$.
1: Conduct OAB embedded with WRF.
2: Output $conf_{j,t}$ and $\{\Delta_{x,j,t}, \Delta_{y,j,t}\}$.
3: Localization of p_j: $\hat{\mathbf{X}}_{j,t}=\arg\max_{l_j} p(\mathbf{X}_{j,t}^{(l_j)}|\mathbf{Y}_{j,1:t})$.
4: Determine the number of unreliable parts S.
5: Localize the target state $\mathbf{X}_t$ as the following strategies.
6: **if** $S=0$
7: $\hat{\mathbf{X}}_t=\sum_{j=1}^{4} w_{j,t}\hat{\mathbf{X}}_{j,t}$, $w_{j,t}=\frac{r_{j,t}}{\sum_{j=1}^{4} r_{j,t}}$.
8: **else if** $S=1$
9: Draw back the drifted part using Eq. 17, and then
10: $\hat{\mathbf{X}}_t=\sum_{j=1}^{4} w_{j,t}\hat{\mathbf{X}}_{j,t}$, $w_{j,t}=\frac{r_{j,t}}{\sum_{j=1}^{4} r_{j,t}}$.
11: **else**
12: $\hat{\mathbf{X}}_t=\arg\max_j p(\hat{\mathbf{X}}_{j,t}|\mathbf{Y}_{j,t})$.
13: **end if**
Output: $\hat{\mathbf{X}}_t$.

where T is the threshold for determining the reliability of parts and set as 0.5 empirically.

2) Draw back the drifted part using MPC if necessary. With different number of unreliable parts S, localization is conducted by different strategies (described in Section IV in detail). The framework of P-OAB-MW can be summarized in Algorithm 1.

III. Online Boosting for Feature Selection with Attention Selection

In this section, online boosting is firstly reviewed. Then the WRF is introduced to estimate the sample weight to train a series of classifiers, especially when occlusion, appearance or illumination change occurs. Finally, results are analyzed to evaluate its performance.

A. Online Boosting for Feature Selection

In order to explain the online boosting, offline boosting is introduced firstly. Offline boosting trains all the weak classifiers sequentially using the original training samples $\mathcal{N}$. Suppose the size of $\mathcal{N}$ is d. Then each weak classifier is trained by all the d samples sequentially, which indicates each weak classifier is trained d times. Different from offline boosting, online boosting [33] employs a sequential inputs and discards each one after updating the weak classifiers. Because only the latest sample is utilized to train the weak classifier, in order to obtain a trustworthy weak classifier, online boosting bootstraps the latest sample K times. Based on [33], K is a *Poisson* variable and $K \sim Poisson(\lambda)$, where λ is the weight of latest sample. After N weak classifiers are trained, the final strong classifier $h^{strong}(\mathbf{x})$ is the linear combination of $\{h^{weak_i}(\mathbf{x})\}_{i=1,...,N}$, where $\mathbf{x}$ denotes the object samples.

Inspired by the online boosting framework, Grabner and Bischof [10] propose a feature selection model where it contains M selectors $\{h_m^{sel}\}_{m=1,...,M}$, and each selector is obtained from N weak classifiers $\{h_m^{weak_i}\}_{i=1,...,N}$

$$h_m^{sel}(\mathbf{x})=h_m^{weak}(\mathbf{x}), \tag{1}$$

where h_m^{weak} is the classifier with minimum classification error $e_m = \arg\min_i e_m^i$, and e_m^i is the estimated error of each weak classifier. In summary, the strong classifier is the linear combination of M selectors which select the best features from a subset of global feature pool, and defined as

$$h^{strong}(\mathbf{x})=\text{sign}(\sum_{m}^{M}\alpha_m h_m^{sel}(\mathbf{x})). \tag{2}$$

The matching confidence of the sample is represented as

$$conf=\sum_{m}^{M}\alpha_m h_m^{sel}(\mathbf{x}) \tag{3}$$

where $\alpha_m$is the weight of each selector, and denoted as $\frac{1}{2}\ln(\frac{1-e_m}{e_m})$. Here, the $conf$ is normalized to [0,1]. With the strategy of [10], the features' distinctiveness for object have been strengthened greatly.

B. Motivation of Weight Relaxation Factor (WRF)

From the work of [33], K is subject to $\frac{e^{-\lambda}}{K!}$. Accordingly, the expectation of K equals λ. In [10], the initial λ is a constant ($\lambda_t=\lambda_{t-1}=1$), which doesn't consider the frequent condition [31] of tracking process. Therefore, when occlusion, appearance and illumination variation occur, the predicted matching confidence $conf$ fluctuates frequently, which makes the following target localization output a poor accuracy. Because the $conf$ is related to the selectors, and the selectors have direct relationship with sample weight λ, to address the fluctuating issue, an adaptive strategy is constructed to determine the weight of samples for training every weak classifier.

Meanwhile, $conf$ to some degree gives an expression indirectly for occlusion, appearance or illumination change. As in these cases, the matching confidence becomes smaller. At the

same time, the cases impact on the sample weight. Therefore, there should be an intrinsic relationship between the sample weight λ and the $conf$. Therefore, this paper reports a concept of WRF to make this bridge.

C. Attentional Selection for Samples

In order ro derive the concept of WRF, the formulation of RF is firstly introduced. Then, the relationship between the matching confidence and sample weight is derived. At last, the formulation of WRF is given.

For a simple introducing, RF [34], [35] is generally used to ensure algorithms' convergence or adjust parameters' changing speed in the field of computer vision. Here, RF is employed to determine the degree-of-attention for the current sample employed to train the weak classifiers. To be more specific, when occlusion, appearance and illumination change occur, the samples extracted in such conditions should be paid less attention, and it is achieved by down-weighting the samples via RF.

The formulation of RF is represented in (4), and utilized for each part p_j independently. So the new sample weight $\lambda_{j,t}$ at time t is obtained by the following iteration

$$\lambda_{j,t} = \lambda_{j,t-1} + RF_{j,t} \cdot \Delta_{j,t-1},\ j \in \{1, ..., 4\},\ \lambda \in [0, 1] \quad (4)$$

where $\Delta_{j,t-1}$ is the step size of $\lambda_{j,t}$ at time $t-1$, and $RF_{j,t}$ is the relaxation factor controlling the changing speed of $\lambda_{j,t-1}$. $RF_{j,t} > 1$ indicates the variation of $\lambda_{j,t-1}$ is accelerated and $RF_{j,t} < 1$ decelerated. After introducing RF, the essential point is to interpret the weighting strategy for samples. Based on the motivation of WRF, there is a relationship between the matching confidence and sample weight. This relationship is explicated in the following.

1) The relationship between the matching confidence and sample weight:

Derived from this formulation of RF, the step size $\Delta_{j,t-1}$ is calculated by $\lambda_{j,t} - \lambda_{j,t-1}$. However, $\lambda_{j,t}$ is unknown and should be estimated. From the motivation of WRF (expressed in Section III-B), if the previous matching confidence $conf_{j,t-1}$ becomes smaller than $conf_{j,t-2}$, it represents that the sample at time $t-1$ is weakened by surrounding scene. Based on this prior, the sample weight $\lambda_{j,t}$ at time t should be reduced accordingly. In other words, the matching confidence $conf_{j,t-1}$ give an another expression for $\lambda_{j,t}$. Based on this expression, $\Delta_{j,t-1}$ can be approximated as $conf_{j,t-1} - \lambda_{j,t-1}$, where $conf_{j,t-1}$ at previous time is less than one.

According to the above description, if $conf_{j,t-1} < conf_{j,t-2}$, $\lambda_{j,t}$ should be reduced. However, the predicted $conf_{j,t-1}$ fluctuates frequently in condition of occlusion [illustrated in Fig. 3 (b)]. Therefore, if $conf_{j,t-1}$ becomes larger than $conf_{j,t-2}$, it can not indicate that the sample at time t is better. In the following explanation, the formulation of WRF is given.

2) WRF formulation:

Denote $\mathbf{C}$ to be a set of real numbers $\{c_{j,t}\}_{j=1:4}$, $c_j \in (0, 1)$. If $conf_{j,t-1} - \lambda_{j,t-1} < 0$, the $conf_{j,t-1}$ can be estimated as $c_j \lambda_{j,t-1}$. The step size $\Delta_{j,t-1}$ can be modified as

$$\Delta_{j,t-1} = conf_{j,t-1} - \lambda_{j,t-1} \approx (c_{j,t} - 1)\lambda_{j,t-1} \quad (5)$$

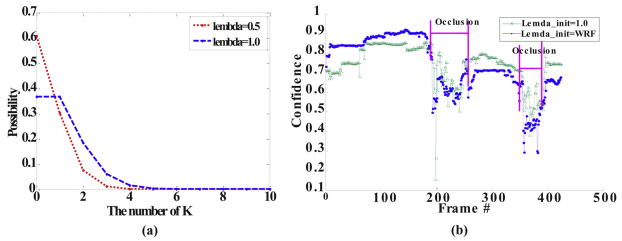


Fig. 3. Illustration of the function of WRF. (a) Poisson distribution with λ=0.5 and 1. (b) Confidence trace before and after the embedding of WRF ($2\sigma^2$=0.6). Green line indicates the new sample having the initial weight 1.0, and blue line indicates the new sample having a relaxed initial weight $\lambda_{t-1}WRF$ (Here, the λ_{t-1} is the weight of global object, and accordingly the index j of patch is neglected).

Algorithm 2 P-OAB with WRF

Input: Time index t. J part samples $\mathbf{x}_{j,t}$. $\lambda_{j,max}$=0.99.

1: Initial the weight importance of $\mathbf{x}_{j,t}$:

2: $\lambda_{j,t} = \lambda_{j,t-1}WRF_{j,t}$.

3: Set $K_{j,t}$ according to $Poisson(\lambda_{j,t})$.

4: Do $K_{j,t}$ times: h^{weak}=$H(h^{weak}, \mathbf{x}_{j,t})$.

5: Boosting for feature selection similar to [10].

Output: h^{strong}=sign$(\sum_{m=1}^{M} \alpha_m h_m^{sel}(\mathbf{x}_{j,t}))$ and matching confidence $conf_{j,t}$=$\sum_{m=1}^{M} \alpha_m h_m^{sel}(\mathbf{x}_{j,t})$.

where $(c_{j,t} - 1) \in (-1, 0)$. Substituting (5) into (4), leads to

$$\lambda_{j,t} = \lambda_{j,t-1}(1 + RF_j(c_{j,t} - 1)). \quad (6)$$

In order to satisfy the relationship between $conf_{j,t-1}$ and $\lambda_{j,t}$, $(1 + RF_{j,t}(c_{j,t} - 1))$ should be within (0,1). In fact, the $RF_{j,t} \in [0, 1]$, and is an under relaxation factor (URF) [32]. Here, there is no need to estimate $RF_{j,t}$ and $c_{j,t}$, but to estimate $(1 + RF_{j,t}(c_{j,t} - 1))$. Utilize an exponent value to approximate $(1 + RF_{j,t}(c_{j,t} - 1))$, leading to the proposed WRF

$$WRF_{j,t} = (1 + RF_{j,t}(c_{j,t} - 1)) = e^{-\frac{(conf_{j,t-1} - \lambda_{j,max})^2}{2\sigma_j^2}} \quad (7)$$

where $\lambda_{j,max}$ is the reference matching confidence, $\lambda_{j,max}$ = 0.990 for a soft estimating, and $\sigma_j^2 \in (0, 1)$ the scale of relaxation. The larger σ_j^2 is, the little relaxed weight $\lambda_{j,t}$ is. Denote the training times $\mathbf{K}_t$ at time t as $\mathbf{K}_t = \{K_{j,t}\}_{j=1:4}$. With (7), the training times $K_{j,t}$ for every weak classifier is chosen by $\lambda_{j,t}$ set as $\lambda_{j,t-1}WRF_j$. P-OAB with WRF is summarized in Algorithm 2. The gray rows represent the weighting for the latest sample.

In order to illustrate the function of WRF, firstly we conduct a simulation experiment by setting $\lambda = 1$ and $\lambda = 0.5$ into $Poisson(\lambda)$, respectively. The result is shown in Fig. 3(a). Secondly, a simple analysis with a video sequence including partial occlusion is done to illustrate the effect of embedding WRF, and the result is shown in Fig. 3(b). From the simulation, it can be seen that, with the decreasing of λ in $Poisson(\lambda)$, the distribution of K turns left, which denotes a lower expectation of K. This phenomenon demonstrates the following two points: 1) with the decreasing expectation of K, the $\{h^{weak_i}(\mathbf{x})\}_{i=1,...N}$ [expressed in (1)] trained K times would output higher classification errors, which generates a lower

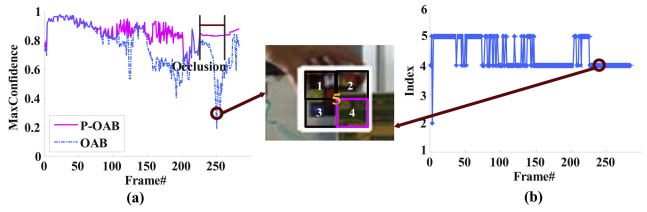


Fig. 4. Failure case of believing the most confident part. (a) Confidence trace. (b) Index of the part with max-confidence. If we choose the most confident part as the guiding one, the localization would be wrong sometimes.

weight of the selector $h^{sel}(\mathbf{x})$, and makes the confidence *conf* [indicated in (3)] reduced. Therefore, the target can recover itself when the occlusion disappeared, not replaced by the occluding scene; and 2) the time for training classifiers is reduced. Meanwhile, from the analysis in Fig. 3(b), before the embedding of WRF, the confidence trace (green line) shows a severe fluctuation in the condition of partial occlusion. However, the confidence trace (blue line) is filtered by the proposed WRF. With the filtering of matching confidence trace, the modeling for localization will be more effective.

IV. Multiple Parts Constraint

The critical problem for part-based methods is how to execute localization with multiple parts. Most existing works choose to believe the most confident part [5]. However, this criterion is not appropriate for online tracking, because the part with the most distinctive features may always have the maximum confidence even when occlusion occurs. For example, in Fig. 4, the max-confidence is assigned to the 4$^{\text{th}}$ part while it is occluded by the green box in the 254$^{\text{th}}$ frame, and the undesirable part may be chosen to support the decision of target location. Therefore, in this paper, a more suitable mechanism for accurate localization is explored.

For tackling the problem mentioned above, the MPC is proposed. MPC is actually a two-stage motion model. Firstly, reliable parts are determined with Bayesian theory. If there are unreliable parts, their location can be recovered by the reliable ones. Secondly, the target location is established according to the number of unreliable parts.

A. Determination of Reliable Parts

In this section, the determination of reliable parts follows the Bayesian framework. For a clear explanation of reliability, we firstly express the localization of each part, then describe the motion model, which is utilized to estimate the localization, and finally give the definition of reliability. The reliability of part p_j at time t is constructed by the posteriori probability $p(\mathbf{X}_{j,t}|\mathbf{Y}_{j,1:t})$ which is defined as

$$p(\mathbf{X}_{j,t}|\mathbf{Y}_{j,1:t}) \propto p(\mathbf{Y}_{j,t}|\mathbf{X}_{j,t}) \cdot \int p(\mathbf{X}_{j,t}|\mathbf{X}_{j,t-1}) p(\mathbf{X}_{j,t-1}|\mathbf{Y}_{j,1:t-1}) d\mathbf{X}_{j,t-1},\ j \in [1, J] \tag{8}$$

where $p(\mathbf{Y}_{j,t}|\mathbf{X}_{j,t})$ is the appearance model measuring the consistency between the target and the observation, $p(\mathbf{X}_{j,t}|\mathbf{X}_{j,t-1})$ is the motion model predicting $\mathbf{X}_{j,t}$ with the previous state $\mathbf{X}_{j,t-1}$.

Then the target state $\hat{\mathbf{X}}_{j,t}$ of p_j can be determined by maximum a posteriori (MAP) estimate over L candidate patches of p_j

$$\hat{\mathbf{X}}_{j,t} = arg \max_{l_j} p(\mathbf{X}_{j,t}^{(l_j)}|\mathbf{Y}_{j,1:t}),\ l_j \in [1 : L]. \tag{9}$$

Here, the appearance model is modeled by Haar features [10], and motion model is constructed by KHM [30], which is contributed by collecting velocity vectors constructed by two neighbor states for the latest five frames. Since online AdaBoost is employed, the appearance model is only related with the current observation $\mathbf{Y}_{j,t}$. But the motion model refers to the latest five frames. Therefore, on the basis of hidden Markov model (HMM) [36], (9) can be changed into

$$\hat{\mathbf{X}}_{j,t} = arg \max_{l_j} p(\mathbf{Y}_{j,t}|\mathbf{X}_{j,t}^{(l_j)}) p(\mathbf{X}_{j,t}|\mathbf{X}_{j,t-1}). \tag{10}$$

For the appearance model, it has been illustrated in Section III in detail. With respect to the motion model, it is presented in the following description.

1) Construction of motion model:

The motion of p_j is constructed by KHM, which is determined by the variation $\{\Delta_{x,j,t}, \Delta_{y,j,t}\}$, and expressed as

$$p(\mathbf{X}_{j,t}|\mathbf{X}_{j,t-1}) = p(\mathbf{X}_{j,t}|\mathbf{X}_{j,t-1})^x p(\mathbf{X}_{j,t}|\mathbf{X}_{j,t-1})^y \tag{11}$$

where

$$p(\mathbf{X}_{j,t}|\mathbf{X}_{j,t-1})^x = \mathcal{N}(\mu_{j,t}, \sigma_{j,t}^2)^x, \tag{12}$$

$$p(\mathbf{X}_{j,t}|\mathbf{X}_{j,t-1})^y = \mathcal{N}(\mu_{j,t}, \sigma_{j,t}^2)^y. \tag{13}$$

However, KHM is easy to be disturbed by an outlier. Therefore, the variation $\{\Delta_{x,j,t}, \Delta_{y,j,t}\}$ at time t is weighted according to [37] as

$$\begin{aligned} \Delta_{x,j,t} &= \Delta_{x,j,t}\eta(\Delta_{x,j,t}|\mu_{j,t-1}, \sigma_{j,t-1}^2), \\ \Delta_{y,j,t} &= \Delta_{y,j,t}\eta(\Delta_{y,j,t}|\mu_{j,t-1}, \sigma_{j,t-1}^2), \end{aligned} \tag{14}$$

where η satisfies Gaussian distribution.

2) Definition of reliability: By integrating the above appearance and motion models, the reliability $r_{j,t}$ of p_j can be defined as

$$r_{j,t} = p(\mathbf{Y}_{j,t}|\hat{\mathbf{X}}_{j,t}) p(\hat{\mathbf{X}}_{j,t}|\mathbf{X}_{j,t-1}) > T. \tag{15}$$

The threshold T is set as 0.5 empirically in our work.

B. Recovery by MPC

In line with the different number of unreliable parts, the process of recovery by MPC is executed by the following different strategies.

If all the parts are reliable, then the target location $\hat{\mathbf{X}}_t$ at time t can be determined by the linear combination of $\hat{\mathbf{X}}_{j,t}$ as

$$\begin{aligned} \hat{\mathbf{X}}_t &= \sum_{j=1}^{4} w_{j,t}\hat{\mathbf{X}}_{j,t}, \\ w_{j,t} &= r_{j,t} / \sum_{j=1}^{4} r_{j,t} \end{aligned} \tag{16}$$

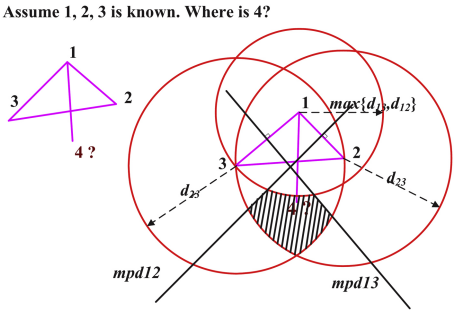


Fig. 5. MPC. For the $d_{14} > d_{12}$ and $d_{14} > d_{13}$, point 4 should be out of the circle centered at point 1 and with a radius of $\max\{d_{13}, d_{12}\}$. For $d_{23} > d_{12}$ and $d_{23} > d_{24}$, point 4 should fall into the circle centered at point 2 and with a radius of d_{23}. For $d_{23} > d_{13}$ and $d_{23} > d_{34}$, point 4 should fall into the circle whose radius is d_{23} and center is point 3. For $d_{14} > d_{24}$ and $d_{23} > d_{34}$, point 4 should be in the mid-perpendiculars between (*mpd13*) and (*mpd12*).

where $r_{j,t}$ is represented in (15). However, because of the occlusion, alternating appearance or illumination change, it is difficult to avoid the occurrence of unreliable parts. How to recover these unreliable part(s) is the key issue for localization.

1) Recovery of one part:

For this case, one part needs to be recovered by the other three reliable ones. In order to clarify the MPC, Fig. 5 is employed to show the main principle. Suppose the center of the three reliable parts are denoted as points 1, 2, and 3. The premise is that the proper alternative space of point 4 connected with other three points is not in the global image. To make the part-based structure having a reasonable geometric shape, we propose the following proposition.

Points 1, 2, 3, and 4 form a quadrangle. The lengths of the edges of the quadrangle cannot be larger than the two diagonals of the quadrangle. This assumption is proven to be effective in our experiments. To be more mathematically specific, let $\mathbf{D}=\{d_{mn}, m, n = 1, 2, 3, 4\}$ be the distance matrix with its element d_{mn} being the distance between point m and n. The point 4 conforming with the constraint should satisfy

$$\begin{array}{llll} d_{14} > d_{12} & \& & d_{14} > d_{13} \\ d_{23} > d_{12} & \& & d_{23} > d_{24} \\ d_{23} > d_{13} & \& & d_{23} > d_{34} \\ d_{14} > d_{24} & \& & d_{14} > d_{34}. \end{array} \tag{17}$$

Equation (17) can be illustrated as Fig. 5. The solution space is the shadow area. From the solution space, we can see that the part structure defined in this paper can be reconstructed after recovered the unreliable part. After the recovery, the location of target is estimated using (16).

2) Recovery of more than one part:

In this situation, there are at least two unreliable parts and recovering the drift parts using (17) is infeasible, as its solution space may be the global image. In this situation, the constraint is constructed using a greedy strategy similar to [5] (choosing the part with the maximal posterior probability).

The most reliable part p_{best} can be estimated as $\arg\max_j p(\hat{\mathbf{X}}_{j,t}|\mathbf{Y}_{j,1:t})$. The other parts should follow p_{best}. Here, different from [5], this paper straightforwardly localizes the other parts with the geometric relationship of the four parts as illustrated in Fig. 1.

V. Experiments

A. Data Sets

In order to evaluate the proposed P-OAB-MW algorithm, experiments on eight video sequences are conducted. There are five publicly available videos as illustrated in Fig. 6 (namely, *Occluded-Face* and *Woman* in [17] and *David-Indoor*, *Twinning-Box*, and *Dollar* in [11]) and three recorded by ourselves (*ColorCup*, *WhiteBoy*, and *WhiteCar*). These video sequences can be classified into two categories: partial occlusion (*Occluded-Face*, *Woman*) and appearance change (*David-Indoor*, *Twinning-Box*, *Dollar*, *ColorCup*, *WhiteBoy*, and *WhiteCar*). Among them, the ground truths of the five public video sequences from [11] and [17] are provided on the authors' homepages, and those of the other three recorded in this paper are labeled by ourselves in every frame.

B. Implementation Details

1) *Experimental Setups:* Performance evaluation is conducted using seven popular trackers: Fragment tracking (Frag) [17], OAB [10], semi-supervised online AdaBoost tracking (SemiBoost) [38], MIL [11], tracker by sparsity-based collaborative model (SCM) [39], block histogram tracking (BHT) [40], and tracker by patch-based dynamic appearance modeling and adaptive basin hopping Monte Carlo sampling (BHMC) [21]. Among them, BHT and BHMC are originally designed for tracking non-rigid objects. As for choosing them, the purpose is to prove the efficiency of the part-based model in this paper. In order to prove the efficiency of the proposed tracker, each tracker is executed ten times, and the best result is selected. It is known that the initialization of target has direct impact on the tracking result. In this paper, the initial bounding box of target abides by the following principle: Contain the target pixels as many as possible and background pixels as few as possible. Meanwhile, for a fair comparison, the initialized bounding box in every tracker is the same.

2) *Parameter Configuration:* For the parameters in this paper, the best result is generated when the numbers of candidate weak classifiers N and selectors M of OAB are set as 200 and 40, respectively. Therefore, the number of weak classifiers and the number of selectors are set as the same as OAB. The scale of each part's searching region is twice the size of the original part. For each tracker, this paper not only considers the default parameters provided by the authors, but also changes them tentatively in a reasonable range according to the initial size of bounding box. The best performance for each tracker is obtained from ten runs.

C. Localization analysis

Firstly, the quantitative performance of the trackers is evaluated. In order to quantitatively evaluate the accuracy of the tracking algorithms, center location error (CLE) is employed to measure the deviation (in pixels) between the detected position and the ground truth. Precision with accepted bias (PAB) is also utilized to represent the ratio of frames below a certain accepted CLE threshold. Besides, average center location error (ACLE) is employed to specify the average center location error of all the frames in every sequence. Figs. 7 and 8

Fig. 6. Eight video sequences. (a)–(f) Severe appearance or illumination change. (g)–(h) Partial occlusion.

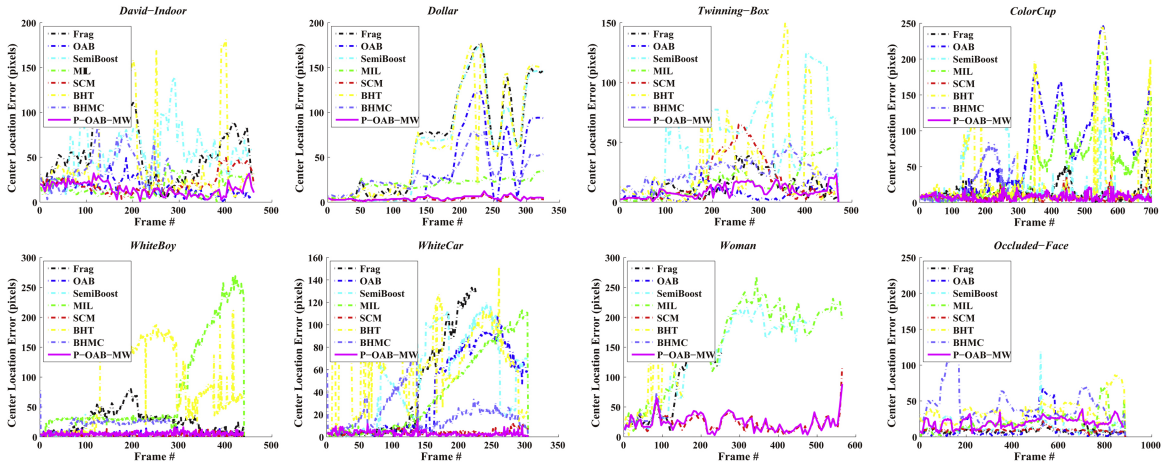


Fig. 7. Center location error. Comparison is done between eight trackers (Frag [17], OAB [10], SemiBoost [38], MIL [11], SCM [39], BHT [40], BHMC [21] and our P-OAB-MW).

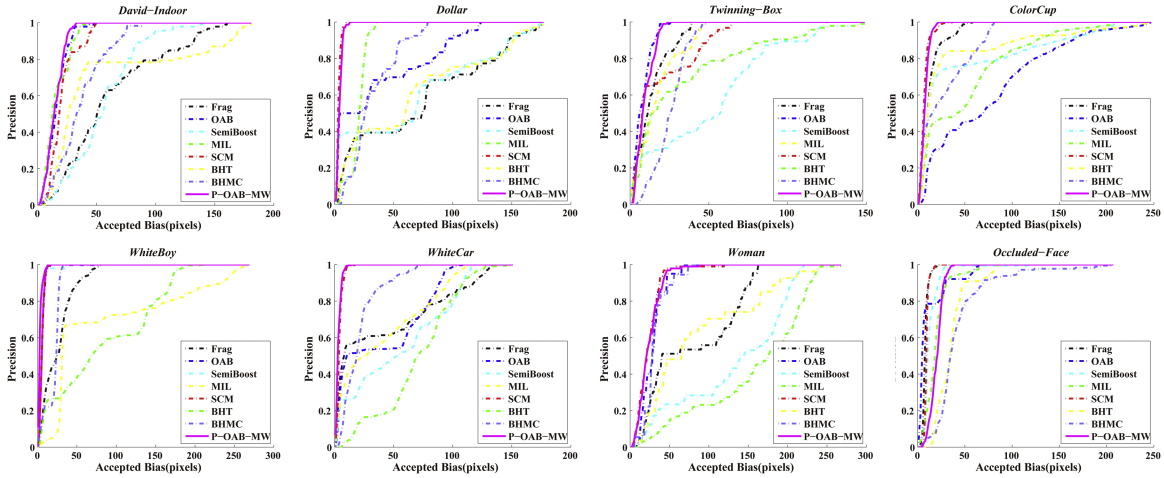


Fig. 8. Precision with accepted bias. Horizontal axis represents an accepted threshold bias (in pixels) between the target center and the ground truth center. Vertical axis is the ratio of the number of frames whose location error is below the threshold.

show CLE (in pixels) and PAB results, respectively. It can be seen from the illustration that the influence of appearance or illumination change can be restrained effectively by the proposed tracker. Beside quantitative evaluation, the qualitative experimental results of eight video sequences are explained in detail as follows.

1) Severe appearance or illumination change:

David-Indoor—In this sequence, illumination variation and appearance change are the main challenge. In this situation, mean shift-like data association model (Frag) is weaker than online boosting framework. But for the OAB, because of the appearance change, the *face* features have been replaced by the *shirt* and generates drift. For the SCM tracker, it constructs a sparsity-based generative model (SGM) module and assigns quite small weights to the background candidates. However, the initial bounding box contains more and more background clutter in the tracking process. Therefore, the tracking result demonstrates drift from the 380th to the 460th frame. As for the patch-based BHMC tracker, it models the patch's cues by the smoothness and steepness characteristics in RGB channels. However, the patches with similar cues in the background may lead the tracking result to a mistake, such as the wall under the shadow which has similar cues to David's hair. Therefore, BHMC generates a mistaken location in the 173th frame. Besides, BHT constructs the appearance model using block histogram, and tries the best to approximate the whole target

TABLE I
AVERAGE CENTER LOCATION ERROR (IN PIXELS). Bold IN EACH ROW IS THE BEST CHOICE, AND *Italic* IS THE SECOND

	Frag	OAB	SemiBoost	MIL	SCM	BHT	BHMC	P-OAB-MW
David-Indoor	61	*16*	55	*16*	20	52	36	**14**
Dollar	69	35	64	*20*	**4**	65	33	**4**
Twinning-Box	*14*	**9**	50	17	18	32	25	**9**
ColorCup	*14*	73	41	51	**7**	32	27	**7**
WhiteBoy	25	5	*4*	78	5	82	20	**3**
WhiteCar	39	38	53	39	*4*	70	21	**3**
Woman	80	28	130	156	*24*	85	31	**23**
Occluded-Face	**8**	12	13	20	*9*	36	43	20

with several blocks. Nevertheless, this tracker has a problem that if the block with the maximum confidence has drift, the whole target will be led to a false location. For example, in the 173th frame, BHT drifts to the bookcase. In contrast, P-OAB-MW has a more stable and accurate performance.

Dollar—In *Dollar* sequence, similar object is its key challenge. Frag and SemiBoost (first frame-based) show a poor performance because of the similar *dollar*. SemiBoost only uses label information from the first frame, and updates the appearance model with online semi-supervised learning in the following frames. Therefore, it is robust for the situation where the target leaves the scene completely. However, this method relies strongly on the first frame, which is prone to cause drift when the target is surrounded by similar objects. MIL has tiny drift because of its background clutter. As for the patch-based BHMC tracker, it cannot distinguish the target with the similar object, which is caused by the fact that the *dollars* have the patches with similar appearance cues. In addition, because the block covering the top-half target has the same appearance to the similar object, BHT generates drift in the 235th frame. Actually, because the left-top part of target maintains constant, P-OAB-MW can draw back the lost part(s) and show more robustness.

Twinning-Box—This sequence has drastic scale and appearance variation. For this video sequence, our tracker and OAB have a similar performance and outperform the other trackers. Because SemiBoost initialize the target appearance with the first frame and bootstrap the samples using semi-supervised classification model, which is easy to be influenced by outliers, such as the background clutter, it generates a poor localization. MIL also has an unfavorable performance, which is mainly caused by the influence of the positive samples bag. If the samples in positive bag are all with insufficient confidence, it will cause a location bias. The precision of accepted bias (PAB) curve of *Twinning-Box* in Fig. 8 also proved this observation. For the SCM tracker, it assigns the lower weight to the background clutters. However, the other side of the box is also seen as the background. Therefore, it shows severe drift in the 318th frame. Similarly, BHT and BHMC generate false target locations because the patches and blocks have similar appearance cues with target.

ColorCup and *WhiteBoy*—The main challenges for these sequences are severe illumination change. From the experiment, our P-OAB-MW and SCM tracker outperform the other trackers. However, in the 649th frame of *ColorCup*, SCM generates a drift. The reason is that the appearance of the whole *ColorCup* is influenced by the severe illumination change. Similarly, in the sequence *WhiteBoy*, because of the whole appearance change, Frag shows drift. MIL demonstrates rather poor performance in the sequence *WhiteBoy*. The main reason is that the positive bag may be replaced by the samples of road, whose appearance cues are similar to the *jeans*. As for the BHMC tracker, the sampling of patches pay more attention to the *white-coat*. Therefore, the patches within the coat guide the tracking process. Taking the 123th frame for example, the bounding box of target shrinks into a box which only contains the coat. For the BHT, the boy wearing the black jacket has the similar block histogram owning to the severe reflection of light, which makes BHT generate a mistaken location in the 123th frame. When the *white boy* walks from the bright region to the shaded region, BHT tends to the region having higher illumination. Therefore, BHT drifts away at last.

WhiteCar—In this sequence, the scale alternating and similar object is the main problem. From the CLE analysis, our tracker and SCM demonstrate a superior performance. However, from the demonstrated video shots in Fig. 9, P-OAB-MW has a better adaptability for scale. OAB, Frag and BHT show poor localization result because of the similar car. As for the BHMC tracker, the patches with similar appearance make the tracking bounding box contain cluttered background and show a poor performance.

2) Partial occlusion:

Woman—In this sequence, partial occlusions, illumination variations and sudden scale change are the main problems. It is obviously that the performance of P-OAB-MW and SCM show a superior performance than the other trackers. For Frag, illumination variation is the main challenge for data association. Because of the *leg* feature is replaced by *car*, OAB outputs an early drift. The representative shots are shown in Fig. 9. Particularly, the 560th frame with sudden scale change is accurately localized by P-OAB-MW. Similarly, BHT is caused by the mistaken block when the woman is occluded. In this situation, the block having the cues of car play more important role in tracking. For the BHMC tracker, the tracking result similarly is influenced by the patch sampling and generates a contracted bounding box in the 96th frame.

Occluded-Face—Partial occlusion is the main challenge. P-OAB-MW seems with a weak performance for

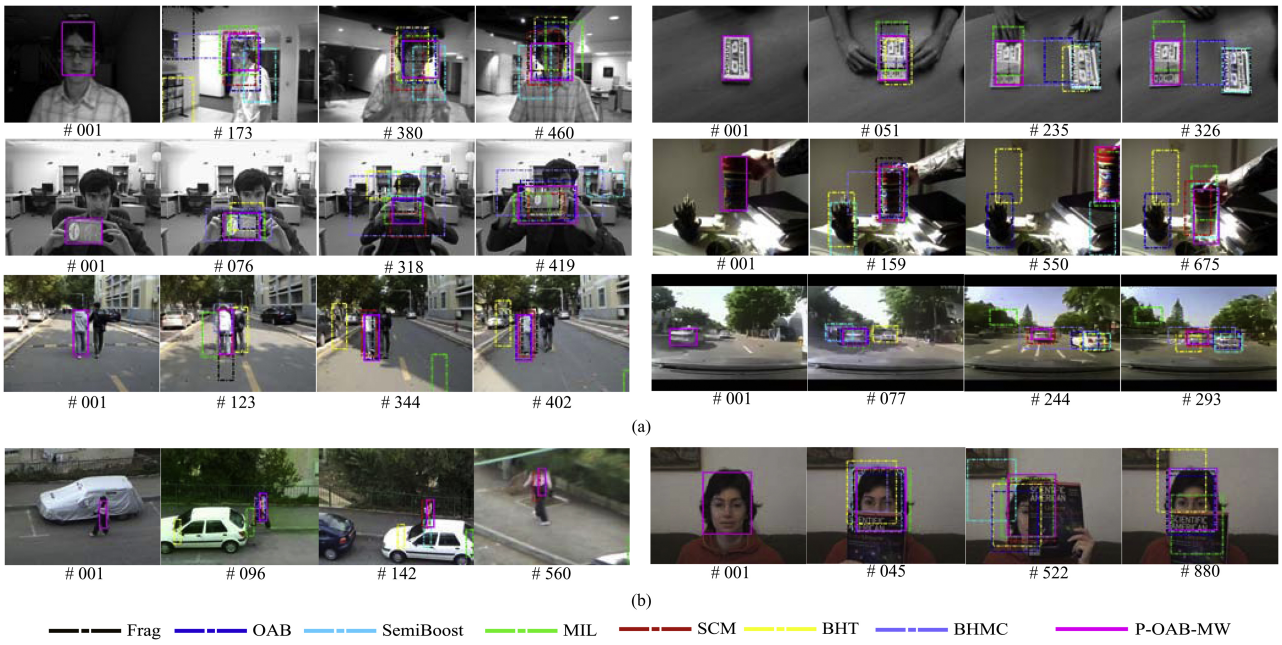


Fig. 9. Video clip shots. (a) Tracking with severe appearance or illumination change. (b) Tracking with partial occlusion.

Occluded-Face because the occluded part is against the assumption (occlusion is a low frequent situation), but its location error remains stable. However, Frag demonstrates an excellent performance. Because the histogram structure of Frag is a superior structure for partial occlusion with constant appearance. It votes the target region as multiple annular bins. Every bin votes different weight for the target. When partial occlusion occurs, the internal bins put more weights for the target and occluded part of external bins take a small portion. But this structure is inadequate to model data association for P-OAB-MW. For the SCM, it can estimate the occluded patches and compare the histograms only formed by the non-occluded ones. Therefore, it shows a beautiful localization. From Table I, OAB, SemiBoost and MIL demonstrate a better average error than P-OAB-MW, but there is large location error in the 522$^{\text{th}}$ frame for OAB and SemiBoost and in the 880$^{\text{th}}$ frame for MIL. Meanwhile, from Fig. 8, the accepted bias for *Occluded-Face* is less than 40 in our tracker (it is an acceptable situation). As for BHMC and BHT, they generate a poor performance. Especially for the BHMC, the patches within the face region make the BHMC tend to shrinkage.

From the above qualitative analysis, our P-OAB-MW demonstrates a superior performance in summary.

TABLE II

PARAMETER OF THREE REPRESENTATIVE FRAMES

	Without MPC			Embedded MPC		
Frame Index	001	096	190	1	96	190
S	0	1	2	0	1	1
r_1	0.808	0.754	0.637	0.835	0.709	0.865
r_2	0.635	0.780	**0.478**	0.709	0.832	0.578
r_3	0.598	**0.125**	0.095	0.726	**0.245**	**0.233**
r_4	0.927	0.867	0.854	0.914	0.947	0.745

S is the number unreliable parts.
r_i represents the reliability of different parts.

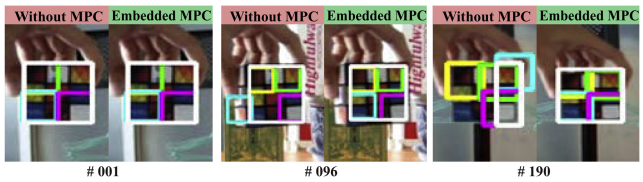


Fig. 10. Experiment for MPC. Top row: object model embedded MPC. Bottom row: object model without MPC. The white box represents the stability of the part structure after the drift part is recovered by MPC, and is set as $\{\min(p_1^{left}, p_3^{left}), \max(p_1^{top}, p_2^{top}), \max(p_2^{left}, p_4^{left}) + width - \min(p_1^{left}, p_3^{left}), \max(p_3^{top}, p_4^{top}) + height - \max(p_1^{top}, p_2^{top})\}$. p_j^{left} and p_j^{top} represent the top-left coordinate of part p_j, *height* and *width* represent the height and width of each part, and are set as the same size.

VI. DISCUSSION

A. Adaptability of MPC

In this discussion, a video sequence *Magic* is employed to validate the feasibility of MPC explicitly. T in (15) is set as 0.5. Some representative frames are shown in Fig. 10. Taking a closer look at Fig. 10, the unreliable part p_3 is always with a hopping. Take the 96$^{\text{th}}$ frame as an example. Without the MPC, the global object localization presents a rough result. In contrast, by embedding MPC, p_3 is recovered by the other reliable parts. The detailed parameters are expressed in Table II. In addition, because the MPC is constituted with latest five frames, the confidence of p_3 could not regain immediately and remains low after recovering. In addition, without the MPC, the number of unreliable parts became two in the 190$^{\text{th}}$ frame. However, after initializing the MPC, this situation does not occur. From the experiment, the MPC can recover the hopping part p_3 robustly, and yield a superior performance for localization.

TABLE III
AVERAGE FRAME TIME CONSUMING (AFTC) FOR EVERY TRACKER

Tracker	Frag	OAB	SemiBoost	MIL	SCM	BHT	BHMC	P-OAB-MW
Compiler	C++	C++	C++	C++	Matlab	Matlab	C++	C++
AFTC	164 ms	284 ms	502 ms	560 ms	1012 ms	120 ms	1504 ms	307 ms

Naturally, the structure in MPC may not be the best choice for localization of all kinds of object, but if the structure of moving object satisfies the form shown in Fig. 1, the MPC could work well regardless of the rigid or non-rigid objects. To further interpret the efficiency of MPC, it can be summarized as follows. MPC provides a strategy to draw back the drifted part of target. Taking the first part of target for example, the other three parts implicitly represent its context. To be specific, the localization of each part not only depends on itself, but also the results generated by the other ones. The construction of MPC provides not only the motion cues for the localization, but also a kind of geometrical context implicitly.

B. Computational Complexity Analysis

In the framework of OAB, the feature pool is extracted in integral image. The computational complexity of integral image is related to the scale of original image. Suppose the size of image $\boldsymbol{I}$ is $w * h$, where w and h represent the width and height of $\boldsymbol{I}$. The computational complexity of integral image is $\mathcal{O}(wh)$. Therefore, the computational complexity for a new frame does not only depends on $\mathcal{O}(MN)$ described in OAB, but $\mathcal{O}(wh)\mathcal{O}(MN)$, where N is the number of candidate weak classifiers, and M is the number of selectors. However, from the structure shown in Fig. 1, the searching region of P-OAB-MW approximates OAB. Therefore, the size of integral image utilized in this paper approximates OABs. Meanwhile, we simulated the average frames per second for OAB, P-OAB-MW (all video sequences). They can reach 3.52 f/s, 3.26 f/s on a PC with a 3.0GHz Intel(R) Core(TM)2 Duo CPU and C++ implementation, respectively.

For a more detailed computational complexity analysis, the average frame time consuming of all the sequences is computed for every tracker and denoted as the average frame time consuming (AFTC), which is listed in Table III. For a fair comparison, the compiler of each tracker also is listed in the table. From the simulation result and localization analysis mentioned above, P-OAB-MW has comparable computational cost with OAB, but with an increased accuracy. BHT actually is a fastest one in the comparison list. However, the tracking performance is a little weaker, as shown from the above localization analysis.

VII. CONCLUSION AND FUTURE WORK

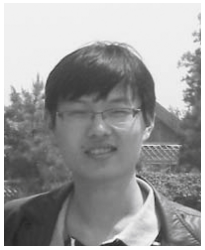
This paper proposes a part-based online AdaBoost tracking with geometry constraint (MPC) and attention selection (WRF)(P-OAB-MW). Because of the generic part-based structure, more reliable tracking results are achieved for different objects, such as face, vehicle, and pedestrian. In situation of occlusion or appearance change, with a strategy of attentional sample selection, the distinctiveness of features extracted from current sample is reduced for tackling the overfitting issue. A two-stage motion model MPC can support the stable part-based tracking by pulling back the drifted part(s) and achieve an accurate localization. The proposed tracker demonstrates superior performance to seven recently developed popular trackers under the condition of appearance or illumination change.

Our future research will firstly concentrate on the adaptive selection strategy of features for part-based model. For each part, through determining the number of features selected by online AdaBoost, the tracking method will improve the adaptability of restraining the influence of occlusion, appearance or illumination change. Secondly, embedding other cues to the tracking method, such as depth information, might avoid the above scene clutter efficiently. At last, a better energy minimization method like graph cut [41] for motion model construction is also our concentration.

REFERENCES

[1] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. Eur. Conf. Comput. Vision*, pp. 1–8, 2008.
[2] F. Dornaika and F. Davoine, "On appearance based face and facial action tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 9, pp. 1107–1124, Sep. 2006.
[3] R. 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.
[4] X. Cao, J. Lan, P. Yan, and X. Li, "Vehicle detection and tracking in airborne videos by multi-motion layer analysis," *Mach. Vision Appl.*, vol. 23, no. 5, pp. 921–935, Sep. 2012.
[5] J. Guo, Y. Liu, C. Chang, and H. Nguyen, "Improved hand tracking system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 5, pp. 693–701, May 2012.
[6] H. Zhou, Y. Yuan, Y. Zhang, and C. Shi, "Non-rigid object tracking in complex scenes," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 98–102, 2009.
[7] J. Wen, X. Gao, Y. Yuan, D. Tao, and J. Li, "Incremental tensor biased discriminant analysis: A new color-based visual tracking method," *Neurocomput.*, vol. 73, pp. 827–839, 2010.
[8] J. Zhu, Y. Lao, and Y. Zheng, "Object tracking in structured environments for video surveillance applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 2, pp. 223–235, Feb. 2010.
[9] C. Shen, J. Kim, and H. Wang, "Generalized kernel-based visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 1, pp. 119–130, Jan. 2010.
[10] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 260–267, 2006.
[11] B. Babenko, M. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.
[12] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 142–149, 2000.
[13] J. Wang, X. Chen, and W. Gao, "Online selecting discriminative tracking features using particle filter," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1037–1042, 2005.

[14] 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.
[15] Z. Han, Q. Ye, and J. Jiao, "Online feature evaluation for object tracking using Kalman filter," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1–4, 2008.
[16] L. Sun and G. Liu, "Visual object tracking based on combination of local description and global representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 408–420, Apr. 2011.
[17] A. Adam and E. Rivlin, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 798–805, 2006.
[18] W. He, T. Yamashita, and H. Lu, "SURF tracking," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1586–1592, 2009.
[19] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers, "Combined region and motion-based 3D tracking of rigid and articulated objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 402–415, Mar. 2010.
[20] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using SIFT features and mean shift," *Comp. Vis. Image Underst.*, vol. 113, no. 1, pp. 345–352, 2009.
[21] J. Kwon and K. Lee, "Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping Monte-Carlo sampling," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1208–1215, 2009.
[22] E. Maggio and A. Cavallaro, "Multi-part target representation for color tracking," in *Proc. Int. Conf. Image Process.*, pp. 729–732, 2005.
[23] M. Yang, J. Yuan, and Y. Wu, "Spatial selection for attentional visual tracking," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, pp. 1–8, 2007.
[24] H. Grabner, J. Matas, L. V. Gool, and P. Cattin, "Tracking the invisible: Learning where the object might be," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1285–1292, 2010.
[25] J. Fan, Y. Wu, and S. Dai, "Discriminative spatial attention for robust tracking," in *Proc. Eur. Conf. Comput. Vision*, pp. 480–493, 2010.
[26] W. Zhou, L. Zhuang, and N. Yu, "A robust part-based tracker," in *Proc. IEEE Int. Conf. Multimedia Expo*, pp. 766–771, 2010.
[27] F. Yang, H. Lu, and Y. Chen, "Bag of features tracking," in *Proc. Int. Conf. Pattern Recognit.*, pp. 153–156, 2010.
[28] V. Takala and M. Pietikainen, "Multi-object tracking using color, texture and motion," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 1–7, 2007.
[29] J. Kwon and K. Lee, "Visual tracking decomposition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1269–1276, 2010.
[30] B. Zhang, M. Hsu, and U. Dayal, "K-harmonic means—A data clustering algorithm," HP, Palo Alto, CA, USA, Tech. Rep. HPL-1999-124, 1999.
[31] H. Cheng, X. Yan, J. Han, and C. Hsu, "Discriminative frequent pattern analysis for effective classification," in *Proc. Int. Conf. Data Eng.*, pp. 716–725, 2007.
[32] G. Batchelor, *An Introduction to Fluid Dynamics*. Cambridge, U.K.: Cambridge University Press, 1967.
[33] N. Oza, "Online ensemble learning," Ph.D. dissertation, Univ. California, Berkeley, CA, USA, 2001.
[34] G. Wollny and F. Kruggel, "Computional cost of non-rigid registration algorithms based on fluid dynamics," *IEEE Trans. Med. Imag.*, vol. 21, no. 8, pp. 946–952, Aug. 2002.
[35] Y. Qu, T. Wong, and P. Heng, "Manga colorization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1214–1220, 2006.
[36] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. of IEEE*, no. 77, no. 2, pp. 257–286, Feb. 1989.
[37] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Realtime tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul. 1997.
[38] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. Eur. Conf. Comput. Vision*, pp. 234–247, 2008.
[39] W. Zhong, H. Lu, and M. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1838–1845, 2012.
[40] S. M. N. Shahed, J. Ho, and M.-H. Yang, "Visual tracking with histograms and articulating blocks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 1–8, 2008.
[41] N. Papadakis and A. Bugeau, "Tracking with occlusions via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 144–157, Jan. 2011.



Jianwu Fang received the B.E. degree in automation and the M.E. degree in traffic information engineering and control from Chang'an University, Xi'an, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

His research interests include computer vision and pattern recognition.

Qi Wang received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent system from University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is an associate professor with Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Yuan Yuan (M'05–SM'09) is a Full Professor with Chinese Academy of Sciences (CAS), Xi'an, China. She has published over 100 papers, including over 70 in journals such as the IEEE transactions and *Pattern Recognition*, as well as conferences papers in CVPR, BMVC, ICIP, and ICASSP. Her research interests include visual information processing and image/video content analysis.