# Object Tracking With Only Background Cues

Annan Li, *Member, IEEE*, and Shuicheng Yan, *Senior Member, IEEE*

***Abstract*— Background cues mainly play a supplementary or accompanying role in most previous approaches for object tracking. If object tracking is treated as a binary target/background classification problem, then the similarity with the target and the difference from the background can be considered equally informative. This leads to an interesting question: is it possible to perform object tracking using background cues only? To answer this question, we propose an object tracking approach that utilizes background cues only. The extensive experimental results positively validate the possibility of performing object tracking using background cues only. The results revealed in this paper also provide a new reference in designing future object tracking methods.**

***Index Terms*— Background cues, object tracking.**

## I. Introduction

OBJECT tracking is a fundamental problem in computer vision [1]. Great progress has been made in constrained settings, such as the scenarios with rigid objects or stationary cameras. Although lots of approaches have been proposed during the past decades, for unconstrained scenarios, robust tracking still remains a challenging and unsolved problem.

One of the key challenges is the appearance variation of the object. In real-world scenarios, variations in the appearance of an object can be caused by many factors, such as shape deformation, pose variation, and occlusion. Different factors usually cause different variations, and the appearances of different kinds of objects may also vary in different ways. Therefore, it is very difficult to handle all kinds of variations with one unique solution due to the great variety of possible changes. In [2], several state-of-the-art methods are evaluated on sequences with different challenging factors. The experimental results show that none of these methods can perform well on all the sequences. How to cover more kinds of variations is a complicated and difficult problem.

In the past decades, many object tracking methods have been proposed [1]. However, the above-mentioned problem still has not been well addressed. In essence, a better tracker requires a higher level of understanding about the visual objects, which is quite complex and challenging. In this paper, we show that this may be avoided in some scenarios. Different from the previous approaches that focus on the objects, we study the object tracking problem only using background information.

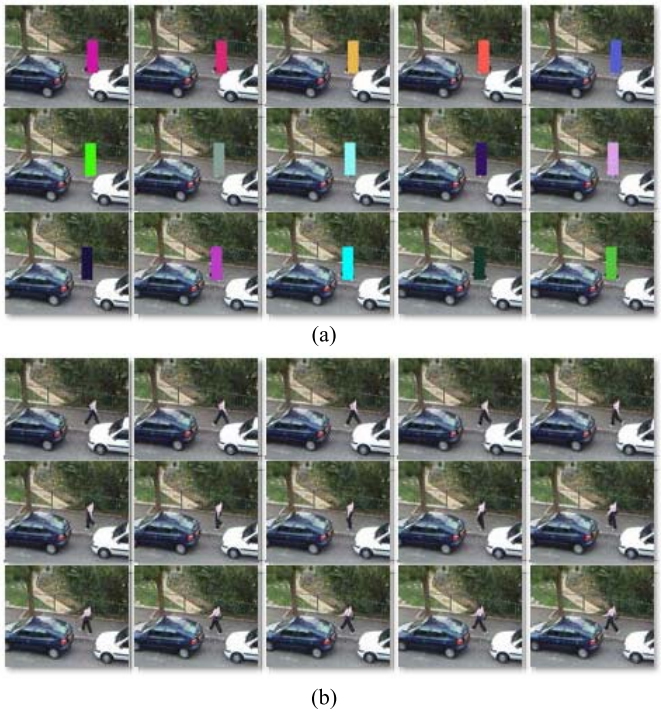


Fig. 1. People can perceive the motion of the target even when its appearance is inconsistent. In the image sequence (a) the moving object is replaced by a patch of random color. Viewing this sequence, people can perceive the movement of the object as well as they view the original sequence (b). For better viewing, please see the original pdf file.

Our study is based on the following observation. In an image sequence shown in Fig. 1(a), a moving object is replaced by patches of random colors. By observing this sequence, we can perceive the movement of the object as well as observe the original sequence shown in Fig. 1(b). The appearance consistency of the object is the basis of most object tracking approaches. Although it does not hold in the sequence (a), we can still perceive the movement. This implies that other cues may help us to perceive the movement. In the sequence (a), the patch color is not consistent, but it is always different from the background color. In other words, we can perceive the movement not only when it is the object but also when it is not the background.

Compared with the being-the-target criterion, the not-being-the-background criterion possesses a natural advantage, i.e., robustness to the appearance variation of an object. When an object changes its appearance, the difference between the object and the background may not necessarily shrink. Thus, the problem of appearance variation can be considerably avoided.

Manuscript received March 1, 2013; revised June 30, 2013, November 21, 2013, and February 24, 2014; accepted March 25, 2014. Date of publication April 17, 2014; date of current version October 29, 2014. This work was supported by the Singapore Ministry of Education under Grant MOE2010-T2-1-087. This paper was recommended by Associate Editor S.-Y. Chien.

A. Li is with Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 119613 (e-mail: lia@i2r.a-star.edu.sg).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: eleyans@nus.edu.sg).

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.2014.2317888

The above observation and analysis lead to an interesting question: can an object be tracked using only background cues? We study this question in this paper and show the feasibility of tracking visual objects using only background cues.

Background cues have been used in the previous studies of object tracking. For the scenarios with stationary cameras, background subtraction is proved to be effective [1]. For the approaches designed for dynamic cameras, background cues are combined with target object cues. However, background cues usually play a supplementary or companying role, which means that background information cannot be used independently in these approaches. Different from the previous approaches, background cues are used independently for object tracking in this paper, which is proved to be feasible according to the extensive experimental results.

The remainder of this paper is organized as follows. Section II gives a brief review of the related work. The proposed tracking approach based on background cues only is described in Section III. Section IV shows the experimental results. Conclusion and discussion are given in Section V.

## II. Related Work

Object tracking is an active topic in the computer vision literature. A number of methods have been proposed during the past three decades. From the perspective of background information utilization, the previous object tracking approaches can be divided into two categories. Methods in the first category usually construct trackers by describing the target itself [3], [4], while those in the second category use the information of both the target and the background. According to the ways of utilizing background information, the approaches in the second category can be further divided into three subcategories, i.e., the approaches based on discriminative models, those using context information in the background, and those based on target/background segmentation. Here, we mainly focus on the second category.

Approaches based on discriminative models treat object tracking as a binary classification problem. In such approaches, background elements are used as negative samples to train a target versus background classifier. Meer [5] addressed the problem of distinguishing a target from the background in their mean-shift (MS) algorithm by downweighting the colors of the object according to the color of the background. Collins *et al.* [6] further extended this approach using online feature selection to switch to the most discriminative color space from a set of different color spaces. Avidan [7] proposed a tracking approach that distinguishes between the object and the background by an ensemble of weak classifiers trained in an online way. Grabner and Bischof [8] proposed a tracking approach that updates the target/background classifier incrementally via online AdaBoost (OAB) feature selection. This work was further extended by semisupervised learning in [9]. Besides the above-mentioned approaches, many target/background classifier-based approaches have also been proposed in recent years. For example, Babenko *et al.* [10] used multiple instance learning (MIL) and Kalal *et al.* [11] used positive and negative learning to enhance the target/background classifier.

Besides building the target/background classifier, background cues can also be used as context information to assist the target tracker. Cerman *et al.* [12] used a single companion region close to the target to improve the tracking results. In [13], auxiliary objects in the background were used to improve the performance of tracking. The auxiliary objects should have three properties: 1) persistent co-occurrence with the target; 2) consistent motion correlation to the target; and 3) being easy to track. While tracking the target object, the auxiliary objects are also tracked. Positions of the target object are predicted by fusing auxiliary object trackers and the target tracker. Grabner *et al.* [14] proposed to learn the supporters coupled with the target to predict the positions of invisible target objects. The coupling is obtained via generalized Hough transform. The proposed approach can improve the tracking performance, especially when the object is heavily occluded.

The contours or silhouettes generally lie between the background and the target. In the segmentation-based tracking approaches, background cues are naturally utilized. In [15]–[18], level sets are adopted for target/ground segmentation. Ren and Malik [19] treated object tracking as a repeated figure/ground segmentation problem, and segmented the object from the background using superpixels. Wang *et al.* [20] also represented the target and the background by superpixel-based segmentation. In their approach, superpixels are clustered by MS for constructing a discriminative appearance model. Besides the level sets and superpixels, Fan *et al.* [21] adopted the matting technique to obtain fine segmentation of target objects.

Regardless of the ways of utilizing background cues, the above-mentioned approaches all treat the background cues as a supplement or accompaniment to the target model, which means that the background cues cannot be used independently. The key difference between the above-mentioned approaches and our method is that we use the background cues independently and solely.

## III. Tracking With Only Background Cues

### A. Tracking Framework

The problem of object tracking is mainly composed of two issues: 1) matching and 2) searching. The former defines how to compute the similarity with the target, and the latter is about how to find the position of the target. In our approach, matching is performed using a confidence map that describes the possibility of each pixel belonging to the target. The position of the target is estimated by searching on the confidence map with motion constraints. The details of computing the confidence map and the motion model are given in Sections III-B and III-C, respectively.

The differences between our method and previous tracking approaches lie in the way of computing the target/background confidence map, including two main aspects. The first aspect is the input. For the object tracking problem, the output is the estimation of target position in the new frame, and the input usually includes a new frame and previous frames with esti-

(a) (b)

Fig. 2. Input of our tracking method. Besides (a) new frame, we only use the elements outside the bounding box in (b) previous frame.

mated positions of the target. Usually, the position of the target is given by a rectangle bounding box. Previous approaches build generative tracking models using the elements inside the bounding box or discriminative models using the elements both inside and outside the bounding box. As shown in Fig. 2, besides the new frame, we only use the elements outside the bounding box in the previous frame. It should be pointed out that we only use the rectangle bounding box for separating the target from the background. This means that we do not use any contour information of the target, which reflects the information of both the target and the background. Here, we emphasize that the background we mean includes all the image elements except the target object. Other moving or static objects are considered as part of the background.

The second aspect of difference is the criterion of computing the confidence. Previous approaches obtain the confidence by the similarity with the target. In our method, we compute the confidence according to the difference from the background.

Considering that object tracking is a continuous process, the appearances of the target and the background should be consistent in a short period. Thus, using multiple frames can produce more stable results. In this paper, we represent the background color by principal component analysis (PCA). At the beginning of the tracking process, the PCA model is computed using all the available previous frames. When many frames are available, we only use a few recent frames. The PCA model updates for every new frame in the tracking process. The main steps of the proposed algorithm are summarized in Algorithm 1.

**Algorithm 1** Tracking Framework

**Input:** $n$ video frames $f_1, \ldots, f_n$ and the target bounding box (rectangle) in the first frame $r_1$
**Output:** The estimated bounding box $r_2, \ldots, r_n$

**for** $i = 2$ to $k^{\text{frm}}$ **do**
  Compute the PCA background color model using frame $f_1, \ldots, f_{i-1}$
  Compute the background confidence map using frame $f_i$, $f_{i-1}$, and the PCA model
  Estimate $r_i$ according to the confidence map and motion model
**end for**
**for** $i = k^{\text{frm}} + 1$ to $n$ **do**
  Compute the PCA background color model using frame $f_{i-k^{\text{frm}}}, \ldots, f_{i-1}$
  Compute the background confidence map using frame $f_i$, $f_{i-1}$, and the PCA model
  Estimate $r_i$ according to the confidence map and motion model
**end for**

### *B. Background Confidence Map*

The color reflects the physical properties of the object, and is robust to shape deformation. Therefore, representing visual objects by color histograms is very popular in the research on object tracking [3], [4]. In this paper, we also use color information to compute the background (target) confidence map.

Although color histogram-based representation is robust to shape deformation, it is sensitive to the sampling regions. Usually, color histograms are calculated using the pixels sampled from rectangle image patches. However, the shapes of visual objects are usually not rectangle. Therefore, noise elements are inevitably involved in the histogram calculation. The noise in the color histograms can be reduced by performing segmentation on the sampling regions. References [19] and [20] show that superpixel segmentation is effective in improving the representation of the target and the background. In this paper, we also adopt superpixel segmentation.

The estimated position of the target is given by a rectangle bounding box. As shown in Fig. 3, we first extract a region of interest (ROI) from the new frame according to the position of the bounding box in the previous frame. Then, simple linear iterative clustering (SLIC) superpixel segmentation [22] is performed on the ROI. As can be seen, the colors are relatively pure in each superpixel. After that, a larger ROI is also sampled in the previous frame. Since we only use the background elements, the elements in the bounding box are excluded in the superpixel segmentation. Based on the segmentation, each ROI can be represented by a set of superpixels. In addition, each superpixel is represented by the position of its center and a color histogram in the hue, saturation, and value color space. Therefore, assuming that there are $m$ superpixels in the ROI of the new frame, the representation of the new frame is denoted as $\text{SP}_{\text{new}} = \{(x_1, y_1, h_1), (x_2, y_2, h_2), \ldots, (x_m, y_m, h_m)\}$. For each $i \in \{1, 2, \ldots, m\}$, $(x_i, y_i)$ is the center position of the $i$th superpixel, and $h_i$ is its corresponding color histogram. Similarly, the previous frame can be represented as $\text{SP}_{\text{previous}} = \{(x_1^*, y_1^*, h_1^*), (x_2^*, y_2^*, h_2^*), \ldots, (x_n^*, y_n^*, h_n^*)\}$. Here, $n$ is the number of superpixels in the previous frame.

For the $i$th superpixel in the new frame and the $j$th superpixel in the previous frame, the similarity $s_{ij}$ between them is defined as

$$s_{ij} = we^{-\gamma_d d_{ij}} + (1 - w)e^{-\gamma_c c_{ij}}. \tag{1}$$

Here, $w \in [0, 1]$ is the weight. $\gamma_d$ and $\gamma_c$ are positive normalization scalars. $d_{ij} = \left(\left(x_i - x_j^*\right)^2 + \left(y_i - y_j^*\right)^2\right)^{1/2}$ denotes the Euclidean distance between the centers of the two superpixels, and $c_{ij}$ is the similarity between histograms $h_i$ and $h_j^*$. Denote the principal component matrix of the background PCA model as $W_{\text{PCA}}$, which is learned from the

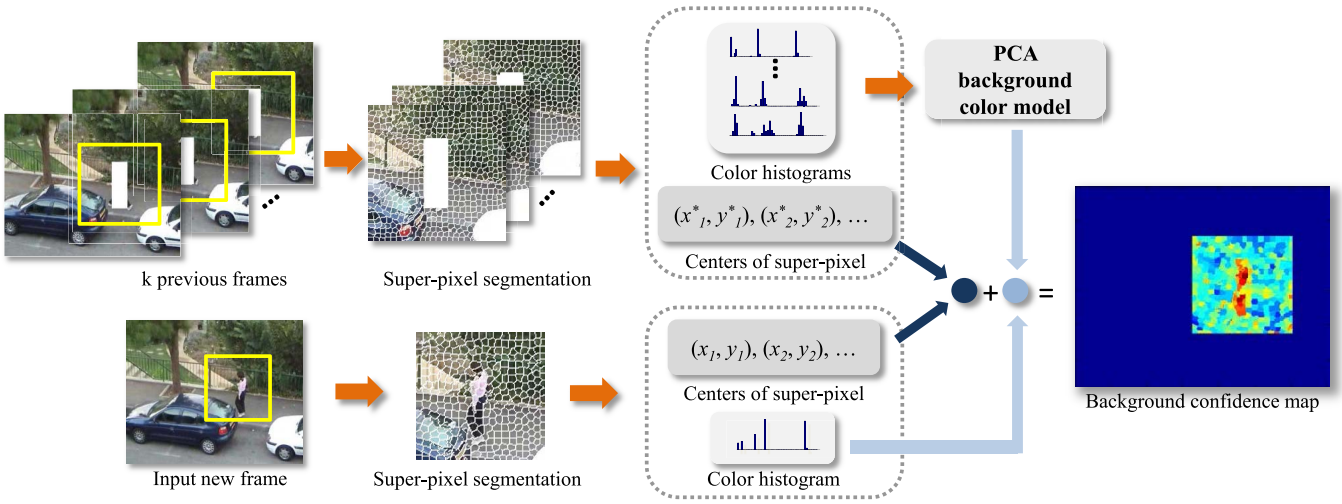



Fig. 3. Flowchart of computing the background confidence map.

background superpixels in several previous frames. The color similarity $c_{ij}$ is given by

$$c_{ij} = \left\| W_{\text{pca}}^T h_i - W_{\text{pca}}^T h_j^* \right\|. \tag{2}$$

Denote the background confidence of the $i$th superpixel in the new frame as $C_i^{\text{bkg}}$. It can be given by its maximum similarity to the superpixels in the previous frame

$$C_i^{\text{bkg}} = \max s_{ij}, \quad \text{s.t. } j \in \{1, 2, \ldots, n\}. \tag{3}$$

In this paper, we replace the similarity with the target by the difference from the background. As aforementioned, if object tracking is formulated as a binary classification problem, they can be considered equally informative. Therefore, the target confidence of the $i$th superpixel can be computed as $C_i^{\text{target}} = 1 - C_i^{\text{bkg}}$.

### C. Motion Model

Our proposed tracking approach contains two main steps. In the first step, a confidence map is computed. The second step, as described in this section, is to estimate the target position based on the confidence map. In this paper, we use rectangle bounding boxes to represent the positions of the targets. A bounding box is described by the position of its center $P = (x, y)^T$ and its scale $S = (w, h)^T$, where $w$ is the width of the bounding box and $h$ is its height.

The estimation of the target position also includes two steps: 1) sampling and 2) weighting. In the first step, $N$ candidate bounding boxes are sampled randomly on the confidence map. In the second step, the sampled candidates are weighted by the motion models that are assumed to follow the Gaussian distribution. We use two Gaussian distributions to describe the movement of the target. One is for position, and the other is for scale. Let $P^t = \{P_1^t, P_2^t, \ldots, P_N^t\}$ be the centers of the candidates, and $S^t = \{S_1^t, S_2^t, \ldots, S_N^t\}$ be the scales. For any $k \in [1, 2, \ldots, N]$, the probability of belonging to the target for the $k$th candidate is denoted as $p(P_k^t | P^{t-1})$. Here, $P^{t-1}$ is the center of the bounding box in the previous frame. Similarly, the probability given by the scale distribution is denoted as $p(S_k^t | S^{t-1})$, and $S^{t-1}$ is the scale of the bounding box in the last frame. $p(P_k^t | P^{t-1})$ and $p(S_k^t | S^{t-1})$ can be calculated according to

$$\begin{aligned} p\left(P_k^t | P^{t-1}\right) &= \frac{1}{2\pi |\Sigma_p|^{\frac{1}{2}}} e^{\left(P_k^t - P^{t-1}\right)^T \Sigma_p^{-1} \left(P_k^t - P^{t-1}\right)} \\ p\left(S_k^t | S^{t-1}\right) &= \frac{1}{2\pi |\Sigma_s|^{\frac{1}{2}}} e^{\left(S_k^t - S^{t-1}\right)^T \Sigma_s^{-1} \left(S_k^t - S^{t-1}\right)}. \end{aligned} \tag{4}$$

In the above equations, $\Sigma_p = \begin{bmatrix} \sigma_p & 0 \\ 0 & \sigma_p \end{bmatrix}$ and $\Sigma_s = \begin{bmatrix} \sigma_s & 0 \\ 0 & \sigma_s \end{bmatrix}$ are diagonal covariance matrices whose diagonal elements are the standard deviations $\sigma_p$ and $\sigma_s$. Lacking prior knowledge of the target objects, we set $\sigma_p$ and $\sigma_s$ empirically in the experiments.

The target confidence of the $k$th candidate $\text{Conf}_k$ is obtained by its mean confidence value in the sampled region. Its weighted confidence is given by

$$\text{Conf}_k^{\text{weighted}} = \text{Conf}_k \cdot p\left(P_k^t | P^{t-1}\right) \cdot p\left(S_k^t | S^{t-1}\right). \tag{5}$$

We choose the candidate with the maximal weighted confidence value as the estimation of the target bounding box.

## IV. Experimental Results

### A. Experimental Setups

The proposed approach is evaluated on 12 sequences for quantitative comparison. The *coke* and *dollar* sequences from [10], the *faceocc* and *woman* sequences from [23], the *basketball* and *singer* sequences from [24], the *bolt* sequence from [20], and our own sequence *garnett* are used in the experiments. We compare our background (BKG)-based tracking approach with nine popular and representative target-based tracking methods. These methods include the fragments (Frag)-based tracking approach [23], the incremental learning-based visual tracking (IVT) [25], the MIL [10], MS [4], color-based particle filters (PFs) [3], OAB [26], semisupervised online boosting (SemiBoost) [9], superpixel tracking (SPT) [20], and visual tracking decomposition (VTD) [24].


Fig. 4. Exemplar tracking results. It is evident that our tracker is robust to partial occlusion in both static (*coke*, *dollar*, and *faceocc* sequence) and dynamic background (*woman* sequence). It can also handle severe illumination (*singer* sequence) and shape variations (*basketball*, *bolt*, and *garnett* sequence). For better viewing, please see the original color pdf file.

In the proposed method, the target/background confidence map is obtained by performing SLIC superpixel segmentation on normalized ROI (Fig. 3). Denote the width and height of the estimated bounding box in the previous frame as $w$ and $h$. The width of ROI window is set to $1.5 \cdot \max(w, h)$ in the current frame and $2 \cdot \max(w, h)$ in the previous frames. The ROI regions in the current and the previous frame are normalized to $128 \times 128$ and $171 \times 171$, respectively. We perform SLIC superpixel segmentation using the VLFeat [27]. The region size and regularizer in SLIC are set to 7 and 0.1, respectively. In the experiments, parameters $w$, $\gamma_d$, and $\gamma_c$ in (1) are set to 0.1, 0.001, and 1 respectively, and the dimension of the PCA model is set to 40. Since the image sequences used in the experiments are obtained from [10] and [20], we adopt their motion parameters setups.

### B. Results on Static Background Sequences

We first evaluate the proposed method on image sequences with the static background. Three sequences (*coke*, *dollar*, and *faceocc*) in [10] are used in the experiments. It should be pointed out that we do not perform background subtraction.

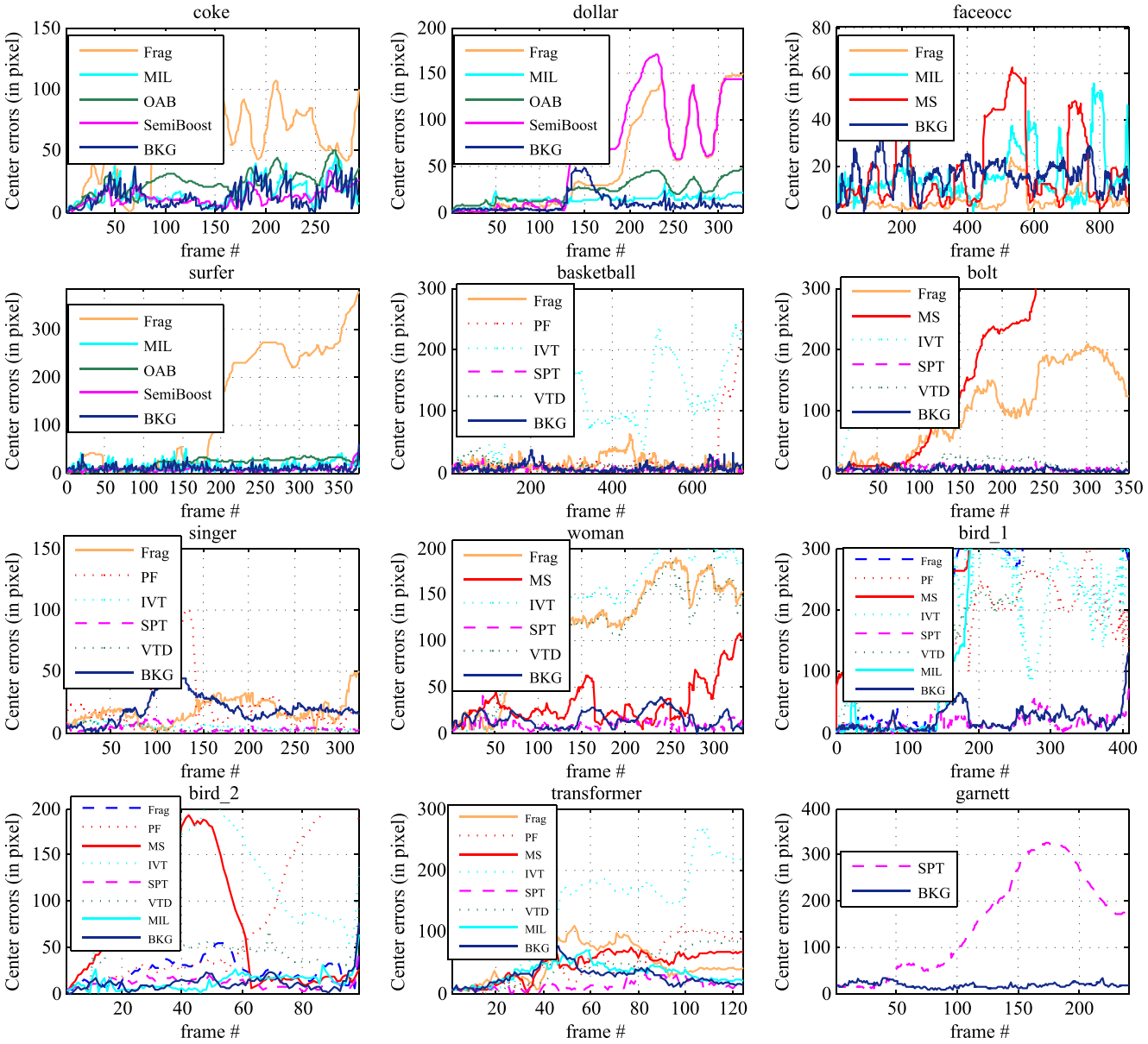



Fig. 5. Performance comparison based on center location errors. For better viewing, please see the original color pdf file.

We compare the results of our method with those of MIL, Frag, MS, OAB, and SemiBoost methods reported in [10].

The target objects in the three sequences are all rigid objects. The main challenge of these sequences is partial occlusion. Some exemplar results of our method and the MIL method are shown in Fig. 4. As can be seen, our method can successfully track the objects with and without partial occlusion as the MIL method does. We also show quantitative comparison with MIL, Frag, MS, OAB, and SemiBoost based on the center location errors in Figs. 5 and 6.[1] We can see that on the *coke* and *dollar* sequences, our method outperforms Frag, and achieves comparable results with MIL. On the *faceocc* sequence, our approach is comparable with MIL, Frag, and MS.

Because of the rigidity, the appearances of target objects in these sequences are stable if there is no occlusion. However, if partial occlusion exists, the target appearance may be polluted. When the pollution accumulates, the estimated position may drift from the target. It is interesting that the proposed method is robust to partial occlusion since the background model is not polluted. In the *coke* sequence, the target is occluded by the plants that are part of the static background. In the *dollar* and *faceocc* sequences, targets are occluded by moving hands and a book. They are modeled as the background before they occlude the target. As a result, partial occlusion only causes temporary decline in accuracy of our approach. It does not lead to drifting. This explains why the proposed method can achieve comparable performance with MIL on these sequences.

### C. Results on Dynamic Background Sequences

Besides the sequences with the static background, we also conduct experiments on sequences with dynamic background. Eight publicly available sequences (*surfer*, *basketball*, *bird1*,

[1]The results of MIL, Frag, MS, OAB, and SemiBoost methods on *coke*, *dollar*, *faceocc*, and *surfer* sequences shown in Figs. 4–6 are cited from [10].

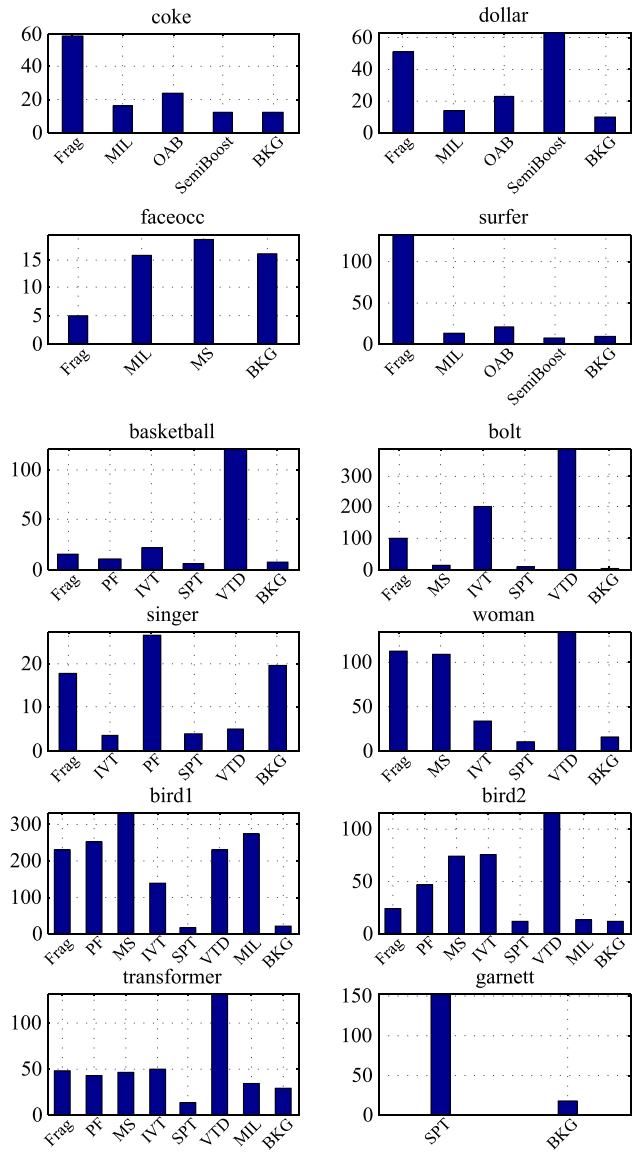



Fig. 6. Average center location errors.

*bird2*, *bolt*, *singer*, *transformer*, and *woman*) [10], [20] and one sequence of our own (*garnett*) are used in the experiments. These sequences have different challenging factors besides the dynamic background. The main challenges in the *singer* sequence are large illumination and scale variation, while the key challenge in *surfer* is pose change. In the *woman*, *bird1*, and *bird2* sequences, the main challenge is partial occlusion. The main difficulty in the *transformer* sequence is shape deformation. For *basketball*, *bolt*, and *garnett* sequences, the large shape variation and confusion caused by similar moving objects make the sequences very challenging. Besides the above-mentioned challenges, the *garnett* sequence also contains severe motion blur.

We compare the proposed method with Frag, MS, OAB, and SemiBoost on the *surfer* [10] sequence and with MS, PF, IVT, Frag, MIL, VTD, and SPT on the *basketball*, *bolt*, *singer*, *woman*, *bird1*, *bird2*, and *transformer* sequences [20]. On the *garnett* sequence, the proposed method is compared with SPT. Some exemplar results and center location error comparisons are shown in Figs. 4–6,[2] respectively.

TABLE I
EXPERIMENTAL COMPARISONS BASED ON THE OVERLAP RATIOS DEFINED IN PASCAL VOC [28]

| Sequence | MS | PF | IVT | Frag | MIL | VTD | SPT | BKG |
|---|---|---|---|---|---|---|---|---|
| *bolt* | 15 | 172 | 4 | 32 | 12 | 195 | 224 | **335** |
| *woman* | 35 | 31 | 49 | 44 | 38 | 27 | **310** | 234 |
| *singer* | 64 | 96 | 332 | 87 | 87 | **350** | 297 | 66 |
| *basketball* | 78 | 455 | 80 | 512 | 204 | 601 | **707** | 685 |
| *bird1* | 1 | 6 | 4 | 47 | 114 | 7 | **139** | 120 |
| *bird2* | 36 | 19 | 9 | 42 | 86 | 9 | **94** | 89 |
| *transformer* | 28 | 32 | 29 | 38 | 30 | 47 | **124** | 63 |

Our method achieves comparable results with MIL, OAB, and SemiBoost on the *surfer* sequence. On *basketball*, *bolt*, *bird1*, and *bird2* sequences, our method achieves comparable performance with the SPT method, which gets the best performance. On *woman* and *transformer* sequences, our method performs better than MS, Frag IVT, and VTD, but not as accurately as SPT. Generally speaking, on these sequences, the proposed method can achieve comparable performance with those tracking approaches using target information.

On the *singer* sequence, our method performs better than PF but not as accurately as other methods. The essential reason why our method can track the target without using its appearance in previous frames is that the contrast between the target and the background is usually big enough to distinguish them. In the *singer* sequence, the color of the singer's clothes and the stage lights are both white. When the stage lighting becomes strong, the absolute target/background contrast declines, which leads to accuracy decrease of our method. In this scenario, using both target and background cues can obtain relative contrast between the target and the background, which makes better results than only using background cues.

The *singer* sequence can be viewed as a scenario where solely using background cues is vulnerable, while the *garnett* sequence is a special case where only using background cues is better than using both target and background information. The main challenging factor in the *garnett* sequence is the severe motion blur. As shown in Fig. 7, in some frames, the target is even transparent. The SPT tracker drifts away from the target, while our tracker successfully tracks the object. Both based on superpixel segmentation, the SPT tracker is similar to our tracker except that it uses target information. In the special case of the *garnett* sequence, its target color model is severely polluted by the transparency. As a result, it drifts away from the target. It shows that in some scenarios, discarding target cues may reduce drifting.

Besides the center location errors, we also evaluate the proposed method by the number of successful frames according to the criterion used in the PASCAL VOC [28] challenge. This criterion is designed for evaluating the results of object

[2]In Figs. 4–6 and Table I, the results of MS, PF, IVT, Frag, MIL, VTD, and SPT on *basketball*, *bolt*, *singer*, *woman*, *bird1*, *bird2*, and *transformer* sequences are cited from [20]. The results of SPT on the *garnett* sequence are obtained by our implementation.


Fig. 7. Our tracker (solid green rectangles) can handle severe motion blur, which makes SPT tracker (dotted magenta rectangles) drift away from target.

TABLE II
SPEED OF THE PROPOSED TRACKER IN FRAMES PER SECOND (FPS)

| Sequence | FPS | Sequence | FPS |
|---|---|---|---|
| *bolt* | 0.89 | *dollar* | 0.44 |
| *basketball* | 0.67 | *garnett* | 0.76 |
| *woman* | 0.96 | *bird1* | 0.76 |
| *faceocc* | 0.91 | *bird2* | 0.72 |
| *coke* | 0.83 | *transformer* | 0.75 |
| *singer* | 0.81 | *surfer* | 0.77 |

detection. Recently, it is adopted to evaluate the results of object tracking [10], [23]. The tracking results are evaluated by the overlap ratio between the predicted bounding box $B^p$ and the ground truth $B^{\text{gt}}$. The overlap ratio $R^{\text{overlap}}$ is defined by

$$R^{\text{overlap}} = \frac{\text{area}(B^p \cap B^{\text{gt}})}{\text{area}(B^p \cup B^{\text{gt}})}. \tag{6}$$

The object is considered to be successfully tracked only if $R^{\text{overlap}} > 0.5$. The results of our method and those of MS, PF, IVT, Frag, MIL, VTD, and SPT reported in [20] are shown in Table I. We can see that the proposed method performs the best on the *bolt* sequence, and achieves the second place on *basketball*, *woman*, *bird1*, *bird2*, and *transformer* sequences. On the *singer* sequence, it is not as accurate as the compared methods except MS.

It should be pointed out that although some results are better, we do not claim that our method can outperform the compared methods. Due to the differences in challenging factors and implementation details, the experimental results are not consistent in all sequences, but they are quite encouraging to validate that it is feasible to track objects using only background cues. The proposed BKG tracker is evaluated on a laptop with an Intel(R) Core(TM) i3-3110M CPU, and its average speed on each sequence is shown in Table II.

## V. CONCLUSION

In the previous approaches for object tracking, background cues are usually used as supplements or accompaniments to the target cues. They cannot be used independently. In this paper, we for the first time study object tracking using only background cues. A novel approach for object tracking is proposed by replacing the similarity with the target criterion used in the previous approaches with the difference from the background criterion. Encouraging experimental results demonstrate the effectiveness of our approach.

Note that the key contribution of this paper is not the proposed tracking method but the phenomenon it reveals. Tracking an object without using its appearance seems incredible, but our extensive experimental results clearly show that it is indeed feasible for both static and dynamic background scenarios. It proves that background cues can be used independently and solely in object tracking.

This phenomenon will provide a new reference in designing tracking methods. In current discriminative tracking approaches, the interactions between target information and background information lead to better prediction. However, combining them together is not always better than using a single one of them if the other is severely polluted. Our results imply that detecting the pollution and temporarily removing the polluted one may be a better choice than always using both of them. The removal can include not only the background information but also the target cues.

It should be pointed out that solely using background cues is not necessarily always better than only using target cues or both target and background cues. It depends on the difficulties of modeling the background. When the background variations are severe, for example, in the scenario that the environment illumination changes suddenly, solely using background cues may be more vulnerable.

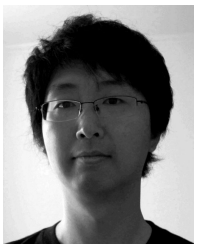
## ACKNOWLEDGMENT

The authors would like to thank Q. Fu for her help in English writing.

## REFERENCES

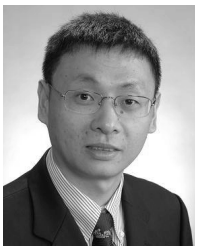
[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, Dec. 2006.
[2] Q. Wang, F. Chen, W. Xu, and M. Yang, "An experimental comparison of online object-tracking algorithms," *Proc. SPIE*, vol. 8138, pp. 81381A-1–81381A-11, Sep. 2011.
[3] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.
[4] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, vol. 2. Jun. 2000, pp. 142–149.
[5] P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.
[6] 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.
[7] S. Avidan, "Ensemble tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 261–271, Feb. 2007.
[8] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, vol. 1. Jun. 2006, pp. 260–267.
[9] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. 10th ECCV*, 2008, pp. 234–247.
[10] 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.
[11] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in *Proc. IEEE Computer Society Conf. CVPR*, Jun. 2010, pp. 49–56.
[12] L. Cerman, J. Matas, and V. Hlaváč, "Sputnik tracker: Having a companion improves robustness of the tracker," in *Proc. Scandinavian Conf. Image Analysis*, 2009, pp. 291–300.
[13] M. Yang, Y. Wu, and G. Hua, "Context-aware visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1195–1209, Jul. 2009.
[14] H. Grabner, J. Matas, L. Van Gool, and P. Cattin, "Tracking the invisible: Learning where the object might be," in *Proc. IEEE Computer Society Conf. CVPR*, Jun. 2010, pp. 1285–1292.

[15] D. Cremers, "Dynamical statistical shape priors for level set-based tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1262–1273, Aug. 2006.
[16] C. Bibby and I. Reid, "Robust real-time visual tracking using pixel-wise posteriors," in *Proc. ECCV*, Oct. 2008, pp. 831–844.
[17] D. Mitzel, E. Horbert, A. Ess, and B. Leibe, "Multi-person tracking with sparse detection and continuous segmentation," in *Proc. 11th ECCV*, 2010, pp. 397–410.
[18] E. Horbert, K. Rematas, and B. Leibe, "Level-set person segmentation and tracking with multi-region appearance models and top-down shape information," in *Proc. IEEE ICCV*, Jun. 2011, pp. 1871–1878.
[19] X. Ren and J. Malik, "Tracking as repeated figure/ground segmentation," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.
[20] S. Wang, H. Lu, F. Yang, and M. Yang, "Superpixel tracking," in *Proc. IEEE 11th ICCV*, Nov. 2011, pp. 1323–1330.
[21] J. Fan, X. Shen, and Y. Wu, "Closed-loop adaptation for robust tracking," in *Proc. ECCV*, 2010, pp. 411–424.
[22] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[23] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Computer Society Conf. CVPR*, Jun. 2006, pp. 798–805.
[24] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *Proc. IEEE Computer Society Conf. CVPR*, Jun. 2010, pp. 1269–1276.
[25] D. 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, 2008.
[26] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. British Mach. Vision Conf.*, vol. 1. 2006, pp. 47–56.
[27] A. Vedaldi and B. Fulkerson. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms* [Online]. Available: http://www.vlfeat.org/
[28] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.



**Annan Li** (M'13) received the B.S. and M.S. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 2003 and 2006, respectively, and the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011.

He was a Post-Doctoral Research Fellow with National University of Singapore, Singapore, from 2011 to 2013. He is currently a Research Scientist with Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His research interests include computer vision, pattern recognition, and statistical learning.



**Shuicheng Yan** (SM'13) is an Associate Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, where he is the Founding Lead of the Learning and Vision Research Group. He has authored or co-authored over 300 technical papers over a wide range of research topics, with more than 1 000 Google Scholar citations and an H-index of 44. His research interests include computer vision, multimedia, and machine learning.

Dr. Yan is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) and *ACM Transactions on Intelligent Systems and Technology*. He received the Best Paper Awards from ACM MM'13, ACM MM'12 (demo), PCM'11, ACM MM'10, ICME'10, and ICIMCS'09; the winner prizes of the classification task in PASCAL VOC from 2010 to 2012; the winner prize of the segmentation task in PASCAL VOC in 2012; the honorable mention prize of the detection task in PASCAL VOC in 2010; the TCSVT Best Associate Editor Award in 2010; the Young Faculty Research Award in 2010; the Singapore Young Scientist Award in 2011; and the NUS Young Researcher Award in 2012.