Liu *et al. EURASIP Journal on Advances in Signal Processing* (2019) 2019:48
https://doi.org/10.1186/s13634-019-0646-0

EURASIP Journal on Advances in Signal Processing

RESEARCH **Open Access**

Robust object tracking via online discriminative appearance modeling

Wei Liu[1*], Xin Sun[2] and Dong Li[3]

Abstract

A robust object tracking algorithm is proposed in this paper based on an online discriminative appearance modeling mechanism. In contrast with traditional trackers whose computations cover the whole target region and may easily be polluted by the similar background pixels, we divided the target into a number of patches and take the most discriminative one as the tracking basis. With the consideration of both the photometric and spatial information, we construct a discriminative target model on it. Then, a likelihood map can be got by comparing the target model with candidate regions, on which the mean shift procedure is employed for mode seeking. Finally, we update the target model to adapt to the appearance variation. Experimental results on a number of challenging video sequences confirm that the proposed method outperforms the related state-of-the-art trackers.

Keywords: Visual tracking, Mean shift, Online learning, Discriminative appearance

1 Introduction

Visual tracking refers to the task of generating the trajectories of the moving objects in a sequence of images. It is a challenging problem in real-time computer vision due to variations of lighting condition, pose, scale, and viewpoint over time. In previous literature, a huge number of tracking methods have been proposed [1–9]. Most object trackers [10–12] search for the target in new frames with several key components: the first is object representation, such as using histogram [13, 14] or sparse representation [15] to model the appearance, using active contours to model the shape [16]; the second is a similarity measure between the reference model and candidate targets [17] and; the third is a local mode-seeking method for finding the most similar location in new frames, such as mean shift [18] or particle filter [19, 20]. Among these three components, appearance modeling of the target is of most importance for robust object tracking. However, it is exceptionally difficult to construct an appearance model with respect to all of those variations in advance.

Many tracking algorithms [16, 18, 20] are based on a fixed target model, and so are unable to track over long time intervals. To increase the robustness, some efforts have been made to employ online update algorithms to adapt to appearance changes of the target objects. Ross in [21] presents an adaptive tracking method which utilizes the incremental principal component analysis and shows robustness to large changes in pose, scale, and illumination. In [22], the authors present a method for evaluating multiple feature spaces while tracking, and a mechanism for online selection of discriminative features to improve tracking performance. The work of Grabner et.al. [23] and Parag [24] shows impressive results of using a classifier as implicit appearance model. They initially learn a binary classifier to distinguish the object of interest from the (neighboring) background and then apply it in each new frame to locate the position of the object. In [25], the authors propose a dynamic weights update mechanism for multiple cues tracking with detection responses as supervision. In [26], the authors consider visual tracking in a weakly supervised learning scenario where multiple imperfect oracles are fused to get a final accurate result. The accuracy of each tracker as well as the most likely object position are simultaneously evaluated by a probabilistic approach. A view-based subspace model is implemented in EigenTracking [27], but it requires intensive off-line learning before tracking.

From a different point of view, parametric density representations also have been used in many tracking algorithms. Han in [28] presents an online appearance

*Correspondence: ldulw@sina.com
[1] Department of Modern Education Technology, Ludong University, 264025 Yantai, China
Full list of author information is available at the end of the article

modeling technique which is based on sequential density approximation and provides accurate and compact representations using Gaussian mixtures. McKenna in [29] suggests Gaussian mixture models created by an EM algorithm for histogram-based trackers, but their method requires knowledge of the number of components, which may not be known in advance. In [30, 31], a pixel-wise target model based on Gaussian distribution is proposed, and it is updated during tracking. However, this method cannot model multi-modal density functions accurately.

Other algorithms based on patches division have also been proposed. The fragment-based tracker [32] divides the target object into several regions and represents them with multiple local histograms. A vote map is used to combine the votes from all the regions in the target frame. In [33], the authors use a patch-based dynamic appearance model in junction with an adaptive Basin Hopping Monte Carlo sampling method to successfully track a non-rigid object. In [34], the authors propose a coupled-layer visual model that combines the target's global and local appearance to address the problem of tracking objects which undergo rapid and significant appearance changes. The local layer in this model, similar as in [33], is a set of local patches that geometrically constrain the changes in the target's appearance. This layer probabilistically adapts to the target's geometric deformation, while its structure is updated by removing and adding the local patches. However, all above works do not consider the discriminative properties of the patches.

In this paper, we propose a robust object tracking algorithm based on an online discriminative appearance modeling mechanism. In contrast with traditional trackers whose computations cover the whole target region and may easily be polluted by the background pixels with similar feature to the foreground model, we divided the target object into a number of patches and take the most discriminative one as the tracking basis. To this patch, we consider both the photometric and spatial information and construct a discriminative target model on it. A likelihood map can be obtained by comparing the target model with candidate regions. Then, the mean shift procedure is employed for mode-seeking. Finally, the target model is updated to adapt to the appearance variation. The preliminary conference version of this work was presented in [35].

The rest of this paper is organized as follows. We review some related works in Section 2. Then, we briefly go over the mean shift framework in Section 3. In Section 4, the proposed discriminative learning-based tracking algorithm is described in detail. Experimental results on challenging video sequences are presented in Section 5, followed by conclusion in Section 6.

2 Related works

Given the dynamic nature of object tracking, having an online learning mechanism to update the target's model is vital to tracking. A large body of work in the literature is dedicated to addressing this issue. In this part, we will review the significant steps that have been taken to achieve a robust target's model.

Some early attempts (e.g., [36, 37]) updated the model by combining the old target model and the detected current appearance with a proper weight function. The weight function was foreseen to adjust the effect of the current appearance on the model. In [38], authors proposed a probabilistic target model and devised an updating scheme based on the EM algorithm. The probabilistic model took into account the stable part of the appearance (slowly varying image observations) and the possibility of losing the target due to occlusion, or noise in a unified framework for its decisions.

The concept of feature selection has been widely used for designing robust target models. A pioneer study along this school of thought is the work of Collins et al. where an online and discriminative feature selection scheme was introduced [39]. The idea of online learning for subspaces was developed by Ross et al. [40], where a low-dimensional subspace capturing the target appearance was incrementally updated using the past and current tracking results.

Inspired by the success of sparse coding in computer vision [41, 42], several sparse coding-based trackers have also been proposed [2–4, 43]. For the sparse trackers, the most popular approach for updating the target model is to learn the dictionary in an online fashion. For example, the differences between the current and previous target samples was used in [44] for adapting the dictionary. In [45], an online learning method for creating non-negative dictionaries was proposed. Since solving the optimization problems involving ℓ_1 norms is computationally expensive, authors in [45] suggest to update the dictionary by gradient descent methods. In [46], a similar idea (gradient descent for updating the dictionary) was utilized albeit authors argued that a discriminative dictionary could be attained by utilizing two disjoint dictionaries to model foreground and background of the target.

3 The basic mean shift

The mean shift method iteratively computes the closest mode of a sample distribution starting from a hypothesized mode. In specifically, considering a probability density function $f(\mathbf{x})$, given n sample points $\mathbf{x}_i$, $i = 1, \cdots, n$, in d-dimensional space, the kernel density estimation (also known as Parzen window estimate) of $f(\mathbf{x})$ can be written as

$$\hat{f}(\mathbf{x}) = \frac{\sum\limits_{i=1}^{n} K(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i)}{h^d \sum\limits_{i=1}^{n} w(\mathbf{x}_i)} \tag{1}$$

where $w(\mathbf{x}_i) \geq 0$ is the weight of the sample $\mathbf{x}_i$, and $K(\mathbf{x})$ is a radially symmetric kernel satisfying $\int k(x)dx = 1$. The bandwidth h defines the scale in which the samples are considered for the probability density estimation.

Then, the point with the highest probability density in scale h can be calculated by mean shift method as follows:

$$m_h(\mathbf{x}) = \frac{\sum\limits_{i=1}^{n} G(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i)\mathbf{x}_i}{\sum\limits_{i=1}^{n} G(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i)} \tag{2}$$

where the kernel profile $k(x)$ and $g(x)$ have the relationship of $g(x) = -k'(x)$.

The kernel is recursively moved from the current location $\mathbf{x}$ to the new location $m_h(\mathbf{x})$ according to mean shift vector, and finally converges to the nearest mode.

The mean shift hill climbing method is one of the most common methods that has been popular for years. After its introduction in the literature [47], it has been adopted to solve various computer vision problems, such as segmentation [48] and object tracking [18]. The popularity of the mean shift method is due its ease of implementation, real time response and robust tracking performance.

In the context of tracking, a sample corresponds to a pixel $\mathbf{x}$ and has an associated sample weight $w(\mathbf{x})$, which defines how likely the pixel $\mathbf{x}$ belongs to an object. Given the initial object position, the traditional mean shift tracking method evaluates the new object position by computing the mean shift vector iteratively according to the Eq. (2). The bandwidth h defines the scale of the target candidate, i.e., the number of pixels considered in the localization process. In [18], the original mean shift tracker uses color histograms as an object representation and Bhattacharya coefficient as a similarity measure. An isotropic kernel is used as a spatial mask to smooth a histogram-based appearance similarity function between model and target candidate regions. The mean shift tracker climbs to a local mode of this smooth similarity surface to compute the translational offset of the target blob in each frame.

4 Method

Our goal in this section is to develop an efficient method that continually constructs and updates the discriminative target appearance model for tracking. The assumption we depend on is that the most informative object region for tracking is the same region that best discriminate between object and background classes. Due to the updating mechanism, local discrimination is sufficient for our work. The whole proposed tracking algorithm is described in detail as follows.

4.1 Discriminative target appearance modeling

Given a target region learned from previous views, we divide it into a number of patches from which we select the most discriminative one as the tracking basis. A larger ring of neighboring pixels surrounding the target region is chosen to represent the background. Let $\hat{X}_0$ denote the current location of the object and Y_0^i represent the ith patch, R_0^i indicate the relative position between the patch Y_0^i and the object $\hat{X}_0$, and $I : \mathbf{x} \rightarrow \mathbf{R}^m$ be the image that maps a pixel $\mathbf{x} = [x\, y]^T \in \mathbf{R}^2$ to a value, where the value is a scalar in the case of a grayscale image ($m = 1$) or a three element vector for an RGB image ($m = 3$).

We use the augmented variance ratio (AVR), the ratio of the between class variance to the within class variance, to measure the discriminative power of a patch as in [22]. For each patch, we compute the histogram on it as well as the background. By normalizing their histograms, we can get a discrete probability density $p(j)$ for the patch, and density $q(j)$ for the background, where index j ranges from 1 to b, the number of histogram buckets.

The log likelihood of an image value j can be given by

$$L(j) = \log \frac{\max\{p(j), \delta\}}{\max\{q(j), \delta\}} \tag{3}$$

where δ is a small value (set to 0.001) that prevents dividing by zero or taking the log of zero. It is obvious that the log likelihood maps the object and background region into positive values for colors distinctive to the object, and negative for colors associated with the background. Colors that are shared by both object and background tend towards zero.

Then the variance ratio of $L(j)$ can be computed to quantify the separability of the patch and background classes:

$$\mathrm{VR}(L; p, q) = \frac{\mathrm{var}(L; (p+q)/2)}{[\mathrm{var}(L; p) + \mathrm{var}(L; q)]} \tag{4}$$

where

$$\mathrm{var}(L; a) = \sum_j a(j)L^2(j) - [\sum_j a(j)L(j)]^2 \tag{5}$$

defines the variance of $L(j)$ with respect to a discrete probability density function $a(j)$.

Since we would like the log likelihood values of pixels on the object and background to both be tightly clustered while the two clusters should ideally be spread apart as much as possible, the denominator of the variance ratio enforces that the within class variances should be small for both object and background classes, while the numerator rewards cases where values associated with object and background are widely separated.

After we got the most discriminative patch $\hat{Y}_0$ with the largest variance ratio and its respective $\hat{R}_0$, a target appearance model can be construct base on it as follows:

$$T_0 = (\hat{X}_0, \hat{Y}_0, \hat{R}_0) \tag{6}$$

Figure 1 shows an example of the target model on *man* sequence.

4.2 Likelihood map for tracking

For each candidate location X_l, we can get the candidate model $T_l = (X_l, Y_l, \hat{R}_0)$ where Y_l is the respective patch with $\hat{Y}_0$ in X_l, determined according to $\hat{R}_0$. Then, we compute the likelihood of the location X_l. Let $S = \{(x, y, I(x, y)) | (x, y) \in Y\}$, $S \in \mathbf{R}^{m+2}$, denote the super patch of Y, whose pixel is a vector containing the pixel coordinates coupled with their image measurements. This super form of patch enables us to both consider the photometric and the spatial information simultaneously. By warping patch $\hat{Y}_0$ to $\hat{S}_0$ whereas Y_l to S_l, the likelihood of the location X_l can be measured by

$$p(T_l) = \exp(-\lambda \mathrm{DIS}(S_l, \hat{S}_0)) \tag{7}$$

where DIS function returns the normalized sum of squared differences between the patch in the candidate region and that in the target model, and λ denotes the weighting parameter that is set to 25.

Then, a new image composed of likelihood values of all candidate locations becomes the "likelihood map" used for tracking.

4.3 Mode-seeking

For the likelihood map got above, we employ the mean shift procedure for mode-seeking, thereby yielding a new estimate of object location, $\hat{X}_1$. Specifically, given the samples being weighted by $w(X_l)$, we can evaluate the translation of the object centroid by computing the mean shift vector ΔX, such that $\hat{X}_1 = \hat{X}_0 + \Delta X$, using the following:

$$\Delta X = \frac{\sum_{l=1}^{n_h} g\left(\left\| \frac{X_l - \hat{X}_0}{h} \right\|^2\right) w(X_l)(X_l - \hat{X}_0)}{\sum_{l=1}^{n_h} g\left(\left\| \frac{X_l - \hat{X}_0}{h} \right\|^2\right) w(X_l)} \tag{8}$$

where the weight of candidate location X_l is specified by:

$$w(X_l) = p(T_l) \tag{9}$$

4.4 Online update

The algorithm iterates through each subsequent frame of the video, extracting new discriminative patch of object, and constructing new target appearance model. However, adaptively updating target model in this manner may promote the occurrence of model drift. To avoid this problem, we take both the current observation and original reference model into account.

After we build the new target appearance model at location $\hat{X}_1$, as $T_1 = (\hat{X}_1, \hat{Y}_1, \hat{R}_1)$, the super patch of it can be established to be a combination of current observation and the original reference model. And the definition of the super patch for target model (not for candidate model) in

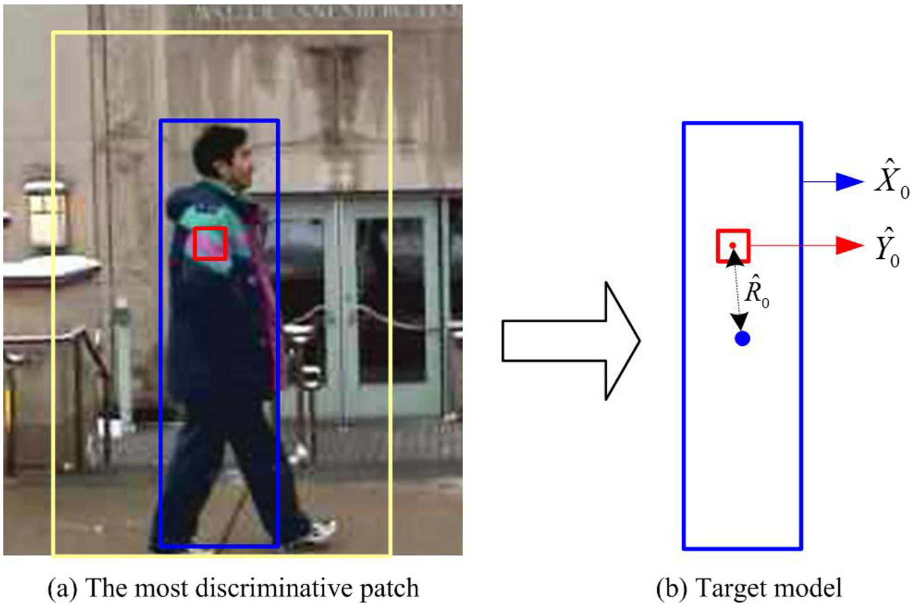


Fig. 1 Example of a target model on *man* sequence. **a** The selection result of the most discriminative patch, where the red square denote the patch and the blue rectangle indicates the target region whereas the yellow indicates local background. **b** The target model constructed base on the selected patch

part 3.2 can be actually rewritten as

$$S = \{(x, y, \omega I(x, y) + (1-\omega) I^{(\mathrm{ref})}(x, y)) | (x, y) \in Y\} \quad (10)$$

where $I(x, y)$ denotes the current image while $I^{(\mathrm{ref})}(x, y)$ represents the reference image of initial frame, and the proportional coefficient ω decide how much one trusts the reference template versus the current observation. Based on the description above, the complete proposed algorithm can be summarized as Algorithm 1. Given an input image and the initial target position in the image, the proposed algorithm first divides the target region into several patches. For each patch, compute the distribution of the target and background and then compute the variance ratio. The patch with the largest variance ration is selected

Algorithm 1 The proposed algorithm.

```
Input:
  Image $I^{(ref)}(x, y)$;
  The initial target/background region and location $\hat{X}_0$;
Output:
  The new region of the target and its corresponding location $\hat{X}_1$ in the subsequent frames;
  % Target modeling
  Divide the target region into patches.
  For each patch $Y_0^i$
    • Compute the density $p(j)$ and $q(j)$ for the patch and background to get the log likelihood of an image value $j$
        $L(j) = \log \frac{\max\{p(j), \delta\}}{\max\{q(j), \delta\}}$
    • Compute the variance ratio of $L(j)$:
        $\mathrm{VR}(L; p, q) = \frac{\mathrm{var}(L; (p+q)/2)}{[\mathrm{var}(L; p) + \mathrm{var}(L; q)]}$
      where
        $\mathrm{var}(L; a) = \sum_j a(j) L^2(j) - [\sum_j a(j) L(j)]^2$
  End For
  Select patch $\hat{Y}_0$ with the largest variance ratio to construct the target appearance by model
    $T_0 = (\hat{X}_0, \hat{Y}_0, \hat{R}_0)$
  % Tracking
  In new arriving frame $I^{(t)}$:
  Initialize the location of the target in the current frame with $\hat{X}_0$.
  Construct the candidate model $T_l = (X_l, Y_l, \hat{R}_0)$ for each candidate location $X_l$.
  Warp $\hat{Y}_0$ to $\hat{S}_0 = \{(x, y, I^{(t)}(x, y)) | (x, y) \in \hat{Y}_0\}$, $Y_l$ to $S_l = \{(x, y, I^{(t)}(x, y)) | (x, y) \in Y_l\}$, and get the likelihood of location $X_l$
    $p(T_l) = \exp\ (-\lambda\ DIS(S_l, \hat{S}_0))$

  Yield a new estimate of object location $\hat{X}_1 = \hat{X}_0 + \Delta X$ according to
    $\Delta X = \frac{\sum_{l=1}^{n_h} g\left(\left\| \frac{X_l - \hat{X}_0}{h} \right\|^2\right) w(X_l)(X_l - \hat{X}_0)}{\sum_{l=1}^{n_h} g\left(\left\| \frac{X_l - \hat{X}_0}{h} \right\|^2\right) w(X_l)}$
  where
    $w(X_l) = p(T_l)$

  Set $\hat{X}_0 \leftarrow \hat{X}_1$.
  Update target model and go to Step 7.
```

to reconstruct the target. At the next frame, we repeat this process and estimate the target position based on a target moving step.

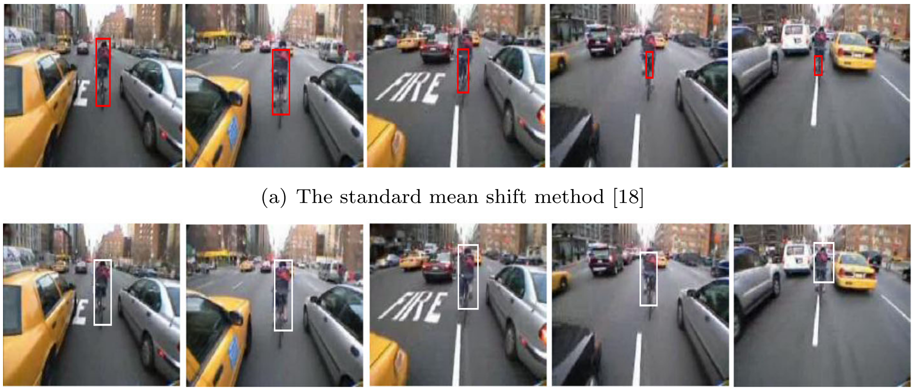
5 Results and discussions

In this section, we use several challenging video sequences taken from moving cameras to illustrate the advantage of our proposed method. We use HSV color space, kernel with Epanechnikov profile and 15 × 15 patch size. We set the values of all parameters of the proposed method by manually adjusting their values to achieve a desired tracking performance.

Firstly, we compare the proposed discriminative learning algorithm with the standard mean shift, which builds the target appearance model by considering the whole object region, on a ridding sequence. To give a convincing comparison, experiments of these two algorithms are carried out under the same conditions. This sequence contains a man ridding bicycle on a busy road. Most of the target region has similar color feature to the background, the gray road surface, expect a small piece of red on his back. The camera keeps moving fast to follow the riding man, with the background made up of cars changing dramatically. Figure 2 shows the tracking results of these two algorithms. As we can see, it is a challenge for traditional mean shift to accurately follow the target since it takes the whole target region into account and is easily confused by the similar background. Draft occurs when the background pollution passes down to the following frames. In comparison, our proposed method can perform well by discriminative appearance modeling for the target.

Further, we compare the proposed method with other prevalent tracker on two video sequences which correspond to different challenges for visual tracking. The first video sequence describes a man in black clothes, which is not outstanding or discriminative, walking outside with cluttered background behind. The branches and the littered stuff result in a lot of interference as well as occurrence of occlusion. The three algorithms we compare are (a) standard mean shift, which builds the target appearance model by considering the whole object region and using RGB histogram; (b) the DF tracker [17] where a distribution field is proposed as the image descriptor. A DF is an array of probability distributions that defines the probability of a pixel of taking each feature value; (c) the proposed discriminative learning method. Figure 3 gives the tracking results of these three algorithms. As we can see, it is difficult for the traditional trackers to follow the target over long time because of the indiscriminate color feature on the most part of the target as well as the similar background pixel pollution. In contrast, the proposed method can seize the most discriminative region of target as the tracking basis, and can achieve pleased performance by both considering its photometric and spatial information. The second video sequence describes a man riding a motorcycle on the hill. The target looks small and shows indiscriminate color appearance on most of the region, expect a small piece of white on his back. Figure 4 shows the tracking results of the compared algorithms on this sequence. We can see the proposed method outperforms the others due to its strong discrimination power between object and its local background.

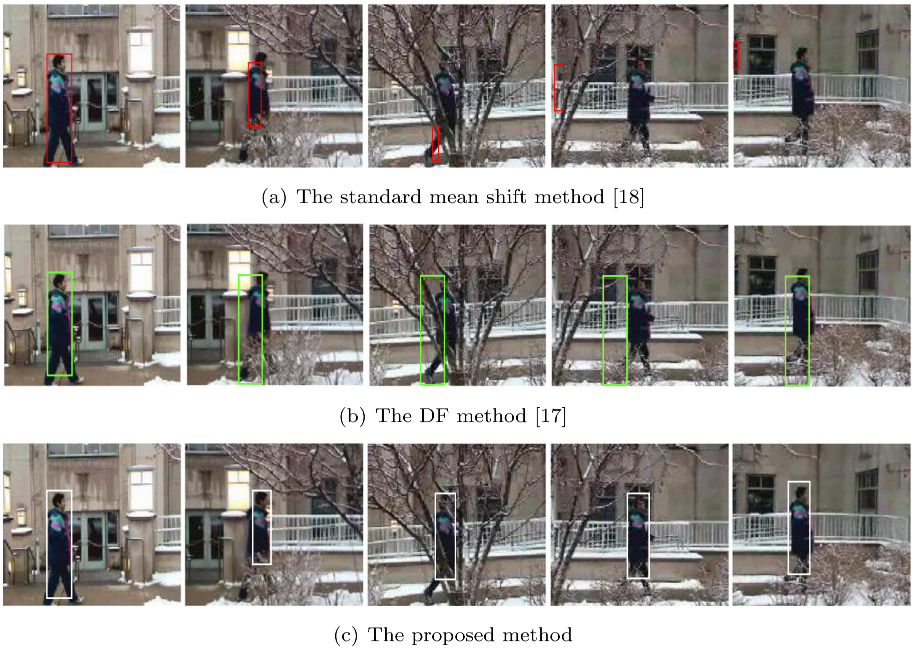
Next, we use another three video sequences with different environment and tracking challenges to further evaluate the performance of the proposed algorithm. In the first video sequence, a tiger table lamp is held and swayed behind a bunch of plant. As the table lamp moves and opens its mouth to reveal the bulb, dramatic appearance change as well as severe occlusion occurs. The second video sequence describes a woman walking in the street, with many other stuff and sheltering cases in the background. The third video sequence is a high jump match, which contains a player with fast and drastic motion.



(a) The standard mean shift method [18]

(b) The proposed method

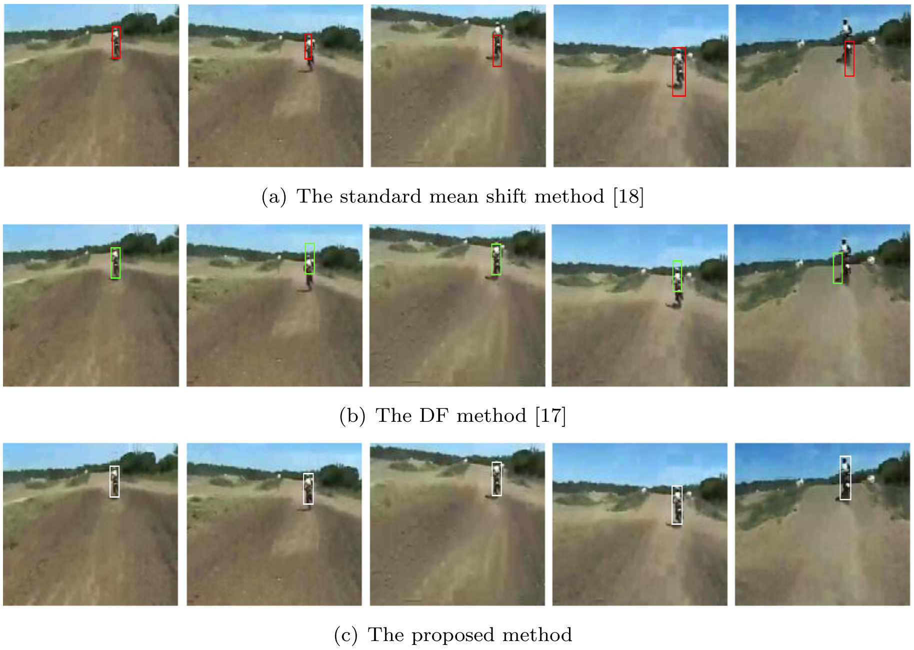
Fig. 2 Tracking results on *riding* video sequence for frames of 0, 99, 213, 251, 319. **a** The standard mean shift method [18]. **b** The proposed method


(a) The standard mean shift method [18]

(b) The DF method [17]

(c) The proposed method

Fig. 3 Tracking results on *man* video sequence for frames of 2, 85, 145, 197, and 240. **a** The standard mean shift method [18]. **b** The DF method [17]. **c** The proposed method

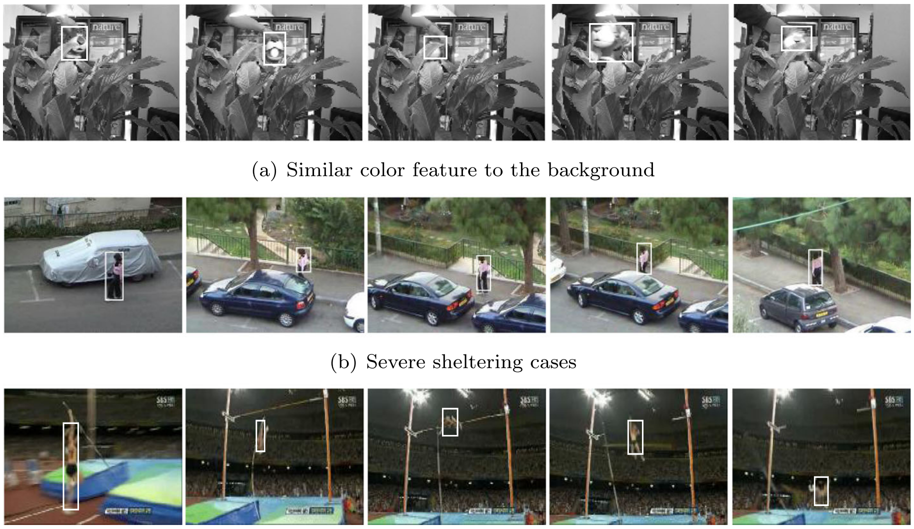

(a) The standard mean shift method [18]

(b) The DF method [17]

(c) The proposed method

Fig. 4 Tracking results on *motocross* video sequence for frames of 2, 7, 10, 17, and 27. **a** The standard mean shift method [18]. **b** The DF method [17]. **c** The proposed method

Tracking results of the first test are shown in Fig. 5a, where we can see that the tiger has been successfully tracked despite of its similar color feature to the background plants. Figure 5b shows the tracking results of the second video sequence, demonstrating that our method provides an effective solution to capture the occluded woman by the updating scheme of the target model. Tracking result of the high jump sequence shown in Fig. 5c indicates the validation of our method in tracking target with dramatic appearance changes. As seen from the



(a) Similar color feature to the background

(b) Severe sheltering cases

(c) Fast motion and dramitic appearance changes

Fig. 5 Further evaluation. **a** Similar color feature to the background. **b** Severe sheltering cases. **c** Fast motion and dramitic appearance changes

tracking results, the key reason why our proposed method outperforms is the online discriminative appearance modeling mechanism. In contrast with other compared trackers whose computations cover the whole target region and may easily be polluted by the background pixels with similar feature to the foreground model, our method divided the target object into a number of patches and take the most discriminative one as the tracking basis. To this patch, we consider both the photometric and spatial information, and construct a discriminative target model on it. A likelihood map can be obtained by comparing the target model with candidate regions. Then, the mean shift procedure is employed for mode-seeking. Finally, the target model is updated to adapt to the appearance variation.

6 Conclusion

A robust object-tracking algorithm is proposed in this paper based on an online discriminative appearance modeling mechanism. By dividing the target into a number of patches, we extract the most discriminative piece of the target as the tracking basis. With the consideration of both the photometric and spatial information, a discriminative target model is constructed base on it. Then, a likelihood map can be obtained by comparing the target model with candidate regions, on which the mean shift procedure is employed for mode seeking. Experiment results have confirmed the effectiveness and robustness of our method.

Abbreviations

Not applicable.

Acknowledgements

This work acknowledged the Editor and anonymous Reviewers.

Authors' contributions

WL conceived the method and developed the algorithm. XS conceived the method, oversaw the project, and wrote the first draft. DL assembled formulations, and drafted the manuscript. All authors read and approved the final manuscript.

Funding

This research is supported in part by the National Natural Science Foundation (grant nos. 61602128 and 61702138) and the Shandong Province Natural Science Foundation (grant no. ZR2016FQ13) of China, and in part by the China Postdoctoral Science Foundation (grant no. 2017M621275 and 2018T110301), in part by the Young Scholars Program of Shandong University, Weihai (no.1050501318006).

Availability of data and materials

Data and source code are available from the corresponding author upon request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All the data (including individual details, images or videos) in the paper are all from public datasets.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Modern Education Technology, Ludong University, 264025 Yantai, China. [2] School of Computer Science and Technology, Harbin Institute of Technology, 264209 Weihai, China. [3] School of Electrical and Information Engineering, Shandong Univeristy, 264209 Weihai, China.

Received: 15 April 2019 Accepted: 20 September 2019
Published online: 29 October 2019

References

1. L. Zhang, W. Wu, T. Chen, N. Strobel, D. Comaniciu, Robust object tracking using semi-supervised appearance dictionary learning. Pattern Recogn. Lett. **62**, 17–23 (2015)
2. S. Zhang, H. Zhou, F. Jiang, X. Li, Robust visual tracking using structurally random projection and weighted least squares. IEEE Trans. Circ. Syst. Video Technol. **25**(11), 1749–1760 (2015)

3. S. Zhang, X. Lan, H. Yao, H. Zhou, D. Tao, X. Li, A biologically inspired appearance model for robust visual tracking. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2357–2370 (2017)
4. Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, M.-H. Yang, Structure-aware local sparse coding for visual tracking. IEEE Trans. Image Process. **27**(8), 3857–3869 (2018)
5. S. Zhang, Y. Qi, F. Jiang, X. Lan, P. C. Yuen, H. Zhou, Point-to-set distance metric learning on deep representations for visual tracking. IEEE Trans. Intell. Transp. Syst. **19**(1), 187–198 (2018)
6. X. Lan, S. Zhang, P. C. Yuen, R. Chellappa, Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker. IEEE Trans. Image Process. **27**(4), 2022–2037 (2018). https://doi.org/10.1109/TIP.2017.2777183
7. Y. Yao, X. Wu, L. Zhang, S. Shan, W. Zuo, in *Proceedings of the European Conference on Computer Vision (ECCV)*. Joint representation and truncated inference learning for correlation filter based tracking, (2018), pp. 552–567. https://doi.org/10.1007/978-3-030-01240-3_34
8. Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, M.-H. Yang, Hedging deep features for visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **41**(5), 1116–1130 (2019)
9. X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, H. Zhou, Learning modality-consistency feature templates: A robust rgb-infrared tracking system. IEEE Trans. Ind. Electron. **66**(12), 9887–9897 (2019). https://doi.org/10.1109/TIE.2019.2898618
10. J. Kwon, K. M. Lee, in *IEEE Conf. Computer Vision and Pattern Recognition*. Visual tracking decomposition, (2010), pp. 1269–1276. https://doi.org/10.1109/cvpr.2010.5539821
11. C. Bao, Y. Wu, H. Ling, H. Ji, in *IEEE Conf. Computer Vision and Pattern Recognition*. Real time robust l1 tracker using accelerated proximal gradient approach, (2012), pp. 1830–1837. https://doi.org/10.1109/cvpr.2012.6247881
12. X. Sun, J. Zhang, Z. Xie, et al., Active-matting-based object tracking with color cues. SIViP (2014). https://doi.org/10.1007/s11760-014-0637-4
13. S. He, Q. Yang, R. W. H. Lau, J. Wang, M. H. Yang, in *IEEE Conf. Computer Vision and Pattern Recognition*. Visual tracking via locality sensitive histograms, (2013), pp. 2427–2434. https://doi.org/10.1109/cvpr.2013.314
14. S. M. N. Shahed, J. Ho, M. H. Yang, in *IEEE Conf. Computer Vision and Pattern Recognition*. Visual tracking with histograms and articulating blocks, (2008). https://doi.org/10.1109/cvpr.2008.4587575
15. X. Wang, Y. Wang, W. Wan, J. Hwang, Object tracking with sparse representation and annealed particle filter. SIViP (2014). https://doi.org/10.1109/icig.2013.81
16. A. Elgammal, R. Duraiswami, L. Davis, in *IEEE Conference on Computer Vision and Pattern Recognition*. Probability tracking in joint feature-spatial spaces, (2003), pp. 781–788. https://doi.org/10.1109/cvpr.2003.1211432
17. L. Sevilla-Lara, E. Learned-Miller, in *IEEE Conf. Computer Vision and Pattern Recognition*. Distribution fields for tracking, (2012), pp. 1910–1917. https://doi.org/10.1109/cvpr.2012.6247891
18. D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5), 234–240 (2003)
19. K. Nummiaro, E. Koller-Meier, L. V. Gool, An adaptive color-based particle filter. Image Vis. Comput. **21**(1), 99–110 (2003)
20. P. Perez, C. Hue, J. Vermaak, M. Gangnet, in *IEEE Conference on European Conference on Computer Vision*. Color-based probabilistic tracking, (2002), pp. 661–675. https://doi.org/10.1007/3-540-47969-4_44
21. D. Ross, J. Lim, R. Lin, M. Yang, Incremental learning for robust visual tracking. IEEE Trans. Int. J. Comput. Vis. **77**(1-3), 125–141 (2008)
22. R. T. Collins, L. Yanxi, M. Leordeanu, Online selection of discriminative tracking features. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1631–1643 (2005)
23. H. Grabner, H. Bischof, *On-line boosting and vision* (IEEE, 2006), pp. 260–267. https://doi.org/10.1109/cvpr.2006.215
24. T. Parag, F. Porikli, A. Elgammal, *Boosting adaptive linear weak classifiers for online learning and tracking*, (2008). https://doi.org/10.1109/cvpr.2008.4587556
25. G. Jia, Y. Tian, Y. Wang, T. Huang, M. Wang, in *ACM Conference on Multimedia*. Dynamic multi-cue tracking with detection responses association, (2010), pp. 1171–1174. https://doi.org/10.1145/1873951.1874179
26. B. Zhong, H. Yao, S. Chen, R. Ji, X. Yuan, S. Liu, W. Gao, in *IEEE Conf. Computer Vision and Pattern Recognition*. Visual tracking via weakly supervised learning from multiple imperfect oracls, (2010), pp. 1323–1330. https://doi.org/10.1109/cvpr.2010.5539816
27. M. J. Black, A. Jepson, in *IEEE Conference on European Conference on Computer Vision*. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation, (1996), pp. 610–619. https://doi.org/10.1007/bfb0015548
28. B. Han, L. Davis, in *IEEE Conference on International Conference on Computer Vision*. On-line density-based appearance modeling for object tracking, (2005), pp. 1492–1499. https://doi.org/10.1109/iccv.2005.181
29. S. McKenna, Y. Raja, S. Gong, Tracking colour objects using adaptive mixture models. Image Vis. Comput. J. **17**, 223–229 (1999)
30. B. Frey, in *IEEE Conference on Computer Vision and Pattern Recognition*. Filling in scenes by propagating probabilities through layers into appearance models, (2000), pp. 185–192. https://doi.org/10.1109/cvpr.2000.855818
31. H. T. Nguyen, M. Worring, R. van den Boomgaard, in *IEEE Conference on International Conference on Computer Vision*. Occlusion robust adaptive template tracking, (2001), pp. 678–683. https://doi.org/10.1109/iccv.2001.937587
32. A. Adam, E. Rivlin, I. Shimshoni, in *IEEE Conf. Computer Vision and Pattern Recognition*. Robust fragments-based tracking using the integral histogram, (2006), pp. 798–805. https://doi.org/10.1109/cvpr.2006.256
33. J. Kwon, K. M. Lee, in *IEEE Conf. Computer Vision and Pattern Recognition*. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling, (2009), pp. 1208–1215. https://doi.org/10.1109/cvprw.2009.5206502
34. L. Cehovin, M. Kristan, A. Leonardis, in *IEEE International Conference on Computer Vision*. An adaptive coupled-layer visual model for robust visual tracking, (2011). https://doi.org/10.1109/iccv.2011.6126390
35. X. Sun, H. Yao, S. Zhang, B. Zhong, in *IEEE Conference on Pervasive Computing Signal Processing and Applications*. On-line discriminative appearance modeling for robust object tracking, (2010), pp. 78–81. https://doi.org/10.1109/pcspa.2010.28
36. P. Pérez, C. Hue, J. Vermaak, M. Gangnet, *Color-based probabilistic tracking, ECCV'*, (2002), pp. 661–675. https://doi.org/10.1007/3-540-47969-4_44
37. A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, 798–805 (2006). https://doi.org/10.1109/cvpr.2006.256
38. A. Jepson, D. Fleet, T. El-Maraghi, Robust online appearance models for visual tracking. TPAMI. **25**(10), 1296–1311 (2003)
39. R. Collins, Y. Liu, M. Leordeanu, On-line selection of discriminative tracking features. TPAMI. **27**(10), 1631–1643 (2005)
40. D. Ross, J. Lim, R. Lin, M. Yang, Incremental learning for robust visual tracking. IJCV. **77**, 125–141 (2008)
41. P. Wilf, S. Zhang, S. Chikkerur, S. A. Little, S. L. Wing, T. Serre, Computer vision cracks the leaf code. Proc. Natl. Acad. Sci. U.S.A. **113**(12), 3305–3310 (2016). https://doi.org/10.1073/pnas.1524473113
42. H. Zhu, X. Huang, S. Zhang, P. C. Yuen, Plant identification via multipath sparse coding. Multimed. Tools Appl. **76**(3), 4599–4615 (2017). https://doi.org/10.1007/s11042-016-3538-4
43. S. Zhang, X. Lan, Y. Qi, P. C. Yuen, Robust visual tracking via basis matching. IEEE Trans. Circ. Syst. Video Technol. **27**(3), 421–430 (2017)
44. S. Zhang, H. Yao, H. Zhou, X. Sun, S. Liu, Robust visual tracking based on online learning sparse representation. Neurocomputing. **100**(1), 31–40 (2013)
45. N. Wang, J. Wang, D.-Y. Yeung, Online robust non-negative dictionary learning for visual tracking. ICCV (2013). https://doi.org/10.1109/iccv.2013.87
46. C. Gong, K. Fu, A. Loza, Q. Wu, J. Liu, J. Yang, Discriminative object tracking via sparse representation and online dictionary learning. IEEE Trans. Cybern. **44**(4), 539–553 (2014)
47. K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inf. Theory. **21**(1), 32–40 (1975)
48. D. Comaniciu, P. Meer, in *IEEE Conference on International Conference on Computer Vision*. Mean shift analysis and applications, (1999), pp. 1197–1203. https://doi.org/10.1109/iccv.1999.790416

Publisher's Note