Neurocomputing 000 (2018) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-layer convolutional network-based visual tracking via important region selection[☆]

Xiao Yun[*], Yanjing Sun, Sainan Wang, Yunkai Shi, Nannan Lu

School of Information and Control Engineering, China University of Mining and Technology, 1 Daxue Road, Jiangsu, Xuzhou 221116, China

ARTICLE INFO

Article history:
Received 28 September 2017
Revised 3 July 2018
Accepted 5 July 2018
Available online xxx

Communicated by Shen Jianbing Shen

Keywords:
Visual tracking
Convolutional neural network
Convolutional network-based tracker
Gabor filter
Important region selection

ABSTRACT

The convolutional network-based tracking (CNT) algorithm provides a training network with warped target regions in the first frame instead of large auxiliary datasets, which solves the problem of convolutional neural network (CNN)-based tracking requiring very long training time and a large number of auxiliary training samples. However, the two-layer CNT uses only gray feature that causes sensitivity to appearance variations. Besides, some samples with useless information should be removed to avoid drifting problems. For these reasons, a multi-layer convolutional network-based visual tracking algorithm via important region selection (IRST) is proposed in this paper. The proposed important region selection model is built via high entropy selection and background discrimination, which enables the training samples to be informative in order to provide enough stable information and also be discriminative so as to resist distractors. The feature maps are also obtained by weighting the template filters with cluster weights. Instead of simple gray features, IRST adds the Gabor layer to explore the texture feature of the target that is effective on coping with illumination and rotation variations. Extensive experiments show that the proposed algorithm achieves superior performances in many challenging visual tracking tasks.

1. Introduction

Visual tracking has received widespread attention for its extensive applications in intelligent video surveillance, human-machine interfaces, robotics, motion analysis, etc [1–3] Depending on the appearance model, existing tracking algorithms can be categorized into two categories: generative and discriminative tracking. Generative tracking algorithms build a target model and search for the candidate image patch with maximal similarity [4–8]. These methods are likely to cause drift problems, and do not take background information into account that improves tracking performance. Therefore, another common solution is to cast tracking as a classification task which separates the target foreground from the background [9–13]. Discriminative and generative methods have complementary advantages in appearance modeling, and the success of a visual tracker depends on both its representation ability against appearance variations and its discriminability between the target and its background, thus leading to a number of tracking methods that are both generative and discriminative [14–18]. Despite these efforts, tracking still faces challenges caused by complicated environments, such as the presence of noise, occlusion, background clutter, and illumination changes.

Convolutional neural networks (CNNs) have shown significant performances on image classification and object detection tasks [19–21]. Many researchers adopted CNNs to the problem of visual tracking due to their success on alleviating the need for hand-crafted features and learning hierarchical and object-specific feature representations automatically [22–26]. CNN shows superior performance on visual tracking, but requires long-time training and large numbers of training samples. Consequently, some methods are proposed to simplify the system [27–32]. Although much effort has been made, CNNs are still not very applicable to visual tracking because the application relies heavily on large auxiliary dataset such as ImageNet [33] to pretrain model [30], thus are not capable to fully take into account the information explored from consequent frames which is effective at discriminating the target from background for visual tracking [34]. For this reason, Zhang et al. [34] presented a convolutional network-based tracking (CNT) algorithm to exploit the local structure and inner geometric layout information of the target. Different from the traditional CNNs, CNT is the first one to use the warped target regions in the first frame as the training samples instead of training with large auxiliary datasets. This state-of-the-art tracker has a simple architecture, whereas constructs a robust tracking effectively. However, only the gray feature of the image patches is used in the

[☆] Fully documented templates are available in the elsarticle package on CTAN.
[*] Corresponding. author.
E-mail address: xyun@cumt.edu.cn (X. Yun).

https://doi.org/10.1016/j.neucom.2018.07.010
0925-2312/© 2018 Elsevier B.V. All rights reserved.

convolution process of the two-layer CNT, thus resulting in sensitivity to illumination change, target rotation, and other appearance changes. Besides, some training samples contain useless information, e.g. information without edge, corner, texture, etc, which not only doesn't improve tracking but also increases computational complexity.

The above disadvantages of CNT lead us to consider how to select the appropriate features and extract the training samples: the features are expected to be illumination and rotation invariant; moreover, the template filters are supposed to be informative in order to provide enough stable information and also be discriminative so as to resist distractors and drifting problems. By integrating the above demands, the multi-layer convolutional network-based visual tracking algorithm via important region selection (IRST) is proposed in this paper. The main contributions of our work are summarized as follows:

- Instead of simple gray features, IRST adds the Gabor layer to explore the texture feature of the target that is effective on coping with illumination and rotation variations.
- The proposed important region selection (IRS) model extracts informative and discriminative subregions with high entropy selection and background discrimination, which enables the training samples to be informative in order to provide enough stable information and also be discriminative so as to resist distractors.
- Moreover, the feature maps are obtained by weighting the template filters with cluster weights to make the more informative clusters influence more on tracking.

The remainder of this paper is organized as follows. The next section gives an overview of the prior work most relevant to this work. After that, the tracking framework of the proposed IRST algorithm is introduced in Section 3. In Section 4, we describe our work in detail. The experimental results are presented in Section 5. Conclusions and future work are finally given in Section 6.

2. Related work

Visual object tracking is a fundamental computer vision problem, and many researches have been focused on solving it. In this section, we briefly review studies that are mostly related to our work.

Generally speaking, visual tracking methods work by constructing a target appearance model from the observed image information using either generative or discriminative approaches. Generative tracking algorithms aim at describing the target appearance using e.g. statistical models or templates [35]. For example, Lan et al. [4,5] proposed a joint sparse representation model for robust feature-level fusion tracking by dynamically removing unreliable features for feature matching. Zhang et al. [6] proposed a robust tracking based on basis matching which learns the target model using a set of Gabor basis functions, so as to have large responses on the corresponding spatial positions after a max pooling. The work in [7] incorporates the temporal and spatial information to boost the tracking performance. The trajectory of the target is encoded through feature representations, and the temporal correspondences are learned directly to estimate the object state from a global perspective based on it, and then the object tracking state is further refined using local spatial object information. Shen et al. [8] introduced the minimum output sum of squared error filter to adapt tracking method for refining the tracking targets via correcting the detection mistakes, and proposed an alternative targets hypotheses to reduce the dependence on detections and adaptively refine the object detection boxes.

On the contrary, discriminative algorithms employ classification methods to differentiate between the target appearance and the surrounding background. Ma et al. [9] learned sparse codes and classifiers jointly under the linearization to nonlinear learning theory. A more generalized feature pooling method for visual tracking is proposed [10] by utilizing the probabilistic function to model the statistical distribution of sparse codes. Ma et al. [11] used tensor-pooled features which are obtained from local sparse codes to model the target whose appearance model not only satisfactorily distinguishes target form background discriminatively, but also alleviates dimensionality. Supancic and Ramanan [12] formulated online tracking as a partially observable decision-making process which allowing trackers to learn action policies appropriate for short-term and long-term tracking. They also demonstrate that reinforcement learning can be used to leverage massive training datasets, which will likely be needed for further progress in data-driven tracking. The work in [13] proposes an adaptive tracker where easy frames are processed with cheap features, while challenging frames are processed with expensive deep features to improve the tracking speed without losing accuracy.

Besides, several trackers are developed in which generative and discriminative models are combined. For instance, in the compressive tracking (CT) algorithm [14], the object is represented by features extracted in the compressive domain, and these features are used to distinguish the fore-ground from the background. Motivated by CT, Song [15] took into account both appearance and spatial layout information in the projections and further proposed an online informative feature selection approach via maximizing entropy energy, which can select the most informative features from the pool. Fan and Ling [16] proposed a novel parallel tracking and verifying framework enjoying both the high efficiency provided by the tracker and the strong discriminative power by the verifier, by taking advantage of the ubiquity of multi-thread techniques and borrowing from the success of parallel tracking and mapping in visual SLAM. Guo et al. [17] proposed a dynamic Siamese network via a fast transformation learning model that enables effective online learning of target appearance variation and background suppression from previous frames, and presented the element-wise multi-layer fusion to adaptively integrate the network outputs using multi-level deep features. Galoogahi et al. [18] proposed a computationally efficient Background-aware correlation filter-based on hand-crafted features that can efficiently model how both the foreground and background of the object varies over time.

Recently, convolutional neural networks (CNNs) have successfully been applied to visual tracking, and have shown to provide excellent results on benchmark tracking datasets. For example, Ma et al. [22] utilized the correlation filters to encode the holistic templates while the convolution filters to encode the part-based templates so as to maintain the long-term memory of target appearance. To cope with long-existing difficulties such as heavy occlusion, Li et al. [23] enhanced the stochastic gradient descent approach in CNN training and used truncated structural loss function to maintain as many training samples as possible. Ning et al. [24] studied the regression capability of long short-term memory in temporal domain for direct prediction of tracking location by exploiting the history of locations and the feature learned by CNN. The two-flow CNN [25] is a generic approach that can be applied to track all kinds of object because it is pretrained with single image in ImageNet [33]. The state-of-the-art MDNet [26], a multi-domain learning framework, is composed of multiple branches of domain-specific layers and shared layers that are combined with a new binary classification layer. It pretrains a CNN using a large set of videos with tracking ground-truth.

All the above trackers require very long training time and a mass of training samples, giving birth to methods that simplify the algorithms. For instance, Fan et al. [27] presented a human tracking algorithm that learns a specific feature extractor with CNNs from

Please cite this article as: X. Yun et al., Multi-layer convolutional network-based visual tracking via important region selection, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.07.010