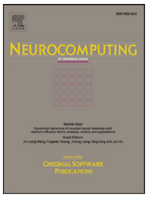
Contents lists available at ScienceDirect

Neurocomputing

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



Visual tracking using Siamese convolutional neural network with region proposal and domain specific updating

Han Zhang*, Weiping Ni, Weidong Yan, Junzheng Wu, Hui Bian, Deliang Xiang

Northwest Institute of Nuclear Technology, Baqiao, Xi'an, 710024, China

ARTICLE INFO

Article history:
Received 22 March 2017
Revised 5 November 2017
Accepted 17 November 2017
Available online xxx

Communicated by Zhu Jianke

Keywords:
Visual tracking
Convolutional neural network
Siamese network
Region proposal

ABSTRACT

This paper deals with the problem of arbitrary object tracking using Siamese convolutional neural network (CNN), which is trained to match the initial patch of the target in the first frame with candidates in a new frame. The network returns the most similar candidate with the smallest margin contrastive loss. For candidate proposals in each frame, a Siamese region proposal network is applied to identify potential targets from across the whole frame. It is also able to mine hard negative examples to make the network more discriminative for the specific sequence. The Siamese tracking network and the Siamese region proposal network share weights which are trained end-to-end. Taking advantage of the fast implementation of fully convolutional architecture, the Siamese region proposal network does not cost much spare time during online tracking. Although the network is trained to be a generic tracker that can be applied to any video sequence, we find that domain specific network updating with a short- and long-term strategy can significantly improve the tracking performance. After combining generic Siamese network training, Siamese region proposal, and domain specific updating, the proposed tracker obtains state-of-the-art tracking performance.

1. Introduction

Taking advantage of the deep convolutional neural networks (CNNs) as well as huge training datasets, computer vision technique has been improved forward with a great step [1]. Especially in the field of image classification [2] and target detection [3,4], the computer has gained human-level performance. However, generic visual object tracking is still challenging, even with the help of deep CNNs and large training datasets. It is caused by the fact that the object is unknown before tracking and also shows arbitrary size and appearance during the whole tracking interval. Most existing trackers, with either a generative model or a discriminative model, such as SCM [5], MILTrack [6], Struck [7], TLD [8] and KCF [9] are based on the object-specific approach. The model of the object's appearance is learned in an online fashion using examples extracted from the video itself. However, the labels of these online training examples are assigned by the online tracking results, hence they are not guaranteed to be the ground truth. The reliability of those labels heavily depends on the validity of the tracking model.

Although the trackers using deep CNNs are not satisfactory, they still outperform the traditional trackers with shallow architectures. To fully take advantage of the representation power of CNNs in visual tracking, several deep architectures have been developed. It is a tricky work to train a proper CNN in an object specific approach. Approaches have been proposed by transferring the CNN weights pre-trained from a large scale image dataset such as the ImageNet [10–12]. Due to the fact that there is only one reference image (ground truth) for tracking, it makes sense to compare the candidate image patches with the reference image patch, and then choose the most similar one. A Siamese CNN satisfies the demand exactly. However, it is difficult for a Siamese network to learn features that are representative enough to adapt to various appearance changes of a same target, such as changes in geometry/photometry, camera viewpoint, illumination, or partial occlusion. There is still a long journey to build a reliable generic visual tracker for arbitrary object tracking.

In this paper, we propose a Siamese CNN tracker based on VGG-M network [13]. Similar CNN architecture has also been employed by trackers in [12,14], which represent the best performance on either accuracy or efficiency up to now. Our proposed network is initialized by the network weights pre-trained on ImageNet classification dataset, and further fine-tuned using three video datasets [15–17]. The entire network is trained end-to-end by SGD with a back-propagation algorithm using the common libraries Caffe [18].

* Corresponding author.
E-mail address: zhang.han@aliyun.com (H. Zhang).

https://doi.org/10.1016/j.neucom.2017.11.050
0925-2312/© 2017 Elsevier B.V. All rights reserved.

The Siamese CNN is not only used for tracking, but also adopted for candidate object proposals by embedding an additional convolutional layer. The proposed Siamese CNN tracker can be used as a generic tracker that can be directly applied to any video sequence. Inspired by the outstanding performance of MDNet [12], we find that domain specific updating can improve the tracking performance significantly. Experiment results show that our proposed network gives best performance among the proposed Siamese networks [10,11,19].

The key contribution of our work lies in two aspects. First, a Siamese CNN tracker is presented and trained using a combined dataset containing 1719 sequences. Without any model updating, the tracker is able to obtain comparable performance with MEEM [20]. We also try to train our Siamese net with a small dataset containing only 58 sequences. Comparable experiment on OTB-100 dataset shows that training data with a large size and good variety is important to obtain a good generic CNN tracker. However, the performance of CNN based tracker is still not as good as expected. We assume it is because that the deep feature is not representative enough for all object variations such as deformation, occlusion, et al. It means that domain specific model updating is essential to further improve the performance of any tracker. By fine-tuning the last three layers of our network during online tracking procedure, significant performance improvement is observed. During online updating, a short- and long-term memory strategy is adopted. The second contribution is that a Siamese region proposal network is constructed based on the proposed Siamese CNN tracker. The region proposal network identifies potential object candidates across the whole incoming frame instead of a small search radius around the previous target location. The improvement brought by the region proposal network mainly lies in three aspects. The first one is to reduce the number of candidate image patches in each frame. It helps to improve the efficiency of the tracking procedure. The second is to re-detect the tracked object after we lost it or the object is blocked for a while. Third, the region proposal procedure helps to mine hard negative examples that are used for model updating. During the tracking process, we update the object model concentrating on hard false-positives that are supplied by the region proposal network. Hard false-positive samples help to suppress distractors caused by complex background clutters, and learn how to re-rank proposals according to the object model. The proposed Siamese region proposal network is designed by taking advantage of the fast implementation of the fully convolutional network proposed in [14], where an additional correlation layer is appended at the end. By sharing parameters between the Siamese tracker network and the Siamese region proposal network, only a little spare time consuming is needed for candidate target region proposal in a new frame. The Siamese CNN tracker and the Siamese region proposal network are combined in a way similar to the combination of target objectness region proposal network and the target detection network in the faster-RCNN target detector [21].

The paper is organized as follows. Section 2 discusses related works in tracking and convolutional neural networks. Our tracking framework is described in Section 3. Section 4 presents the experimental evaluations and results. Finally, conclusions are given in Section 5.

2. Related work

2.1. Siamese tracker

Object representation is one of the major components in any visual tracking algorithm. Wang [22] concludes that the feature extractor is the most important part of a tracker and the observation model is not significantly important if the features are good enough. Fortunately, deep CNN is a powerful tool to learn good visual features. Given the initial state (e.g., position and extent) of a target object in the first image, the goal of tracking is to estimate the states of the target in the subsequent frames, which means to find the new position and extend for the target in the new frame. It is straightforward to build a Siamese CNN architecture to find the potential target state that best resembles the initial state. Siamese CNNs have been successfully applied to face and pedestrian verification. Several recent works have also suggested using Siamese CNNs in the context of tracking [10,11,19].

A Siamese tracking architecture is designed in [10], which is based on the deep VGGNet incorporating with ROI pooling features from multiple convolution layers. The network is trained using the ALOV dataset [17] from which 60,000 pairs of frames and each pair of frames containing 128 pairs of boxes are extracted. However the proposed tracker only performs slightly better than MEEN [20] and MUSTer [23], both of which are shallow feature based approaches. Due to the large amount of parameters, the tracker takes about a second to process one frame on GPU, even without any model updating during online tracking.

A novel fully-convolutional Siamese network is proposed in [14] which is trained end-to-end based on the ILSVRC15 video object detection dataset. The training dataset contains more than 4000 sequences. Taking advantage of the fast implementation of the fully-convolutional architecture, this approach achieves competitive performance in modern tracking benchmarks. The fully-convolutional Siamese network is able to perform at speeds that exceed the real-time requirement. The fast implementation also relies on the fact that no model update is incorporated during tracking.

2.2. Region proposal

Most of the current visual trackers search the potential target position in a search radius centered on the previous predicted object location with the radius no more than two times of the previous target extent [7]. One important reason that the existing trackers avoid employing a larger search radius is the potential distractions from the background. It is not a trivial task to update a discriminative classifier when the negative sample space grows greatly with the samples coming from the extended search radius. Besides, the radius search approach requires dense region proposal with great redundancy which slows down the tracking procedure. Also it may not be valid always, in particular for fast and irregularly moving objects.

Proper region proposals are able to improve the performance of any tracker significantly. In [10], the performance of the proposed SINT tracker is improved by a considerable margin with an adaptive sampling and a simple use of optical flow [24]. In [25] spatial-temporal saliency guided sampling approach is adopted to guide a more accurate target localization for qualified sampling within an inter-frame motion flow map. In [26], an object tracker termed as EBT is presented that is not limited to a local search window and has an ability to probe efficiently the entire frame. The EBT tracker searches for instance-specific candidate target proposal across the whole frame by training a SVM classifier based on Edge Box feature [27]. Experiment results show that the edge box based proposals outperforms traditional target candidate proposal approaches such as nearby window search or particle search.

2.3. Online updating

Online updating is essential no matter for shallow feature based tracker, or deep feature based one. However, it is time consuming and difficult to control. The Discriminative Correlation Filter (DCF) based trackers apply a more tractable online optimization problem by changing to the Fourier basis [28]. Benefiting from