J. Vis. Commun. Image R. 37 (2016) 3–13

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Network in network based weakly supervised learning for visual tracking ☆

Yan Chen *, Xiangnan Yang, Bineng Zhong, Huizhen Zhang, Changlong Lin

Department of Computer Science and Technology, Huaqiao University, China

ARTICLE INFO

Article history:
Received 13 December 2014
Accepted 8 October 2015
Available online 22 October 2015

Keywords:
Little supervision
Handcrafted features
Visual tracking
Deep network
Network in network
Object appearance model
Large scale training data
Drifting problem

ABSTRACT

One of the key limitations of the many existing visual tracking method is that they are built upon low-level visual features and have limited predictability power of data semantics. To effectively fill the semantic gap of visual data in visual tracking with little supervision, we propose a tracking method which constructs a robust object appearance model via learning and transferring mid-level image representations using a deep network, i.e., Network in Network (NIN). First, we design a simple yet effective method to transfer the mid-level features learned from NIN on the source tasks with large scale training data to the tracking tasks with limited training data. Then, to address the drifting problem, we simultaneously utilize the samples collected in the initial and most previous frames. Finally, a heuristic schema is used to judge whether updating the object appearance model or not. Extensive experiments show the robustness of our method.

1. Introduction

Visual tracking has gained a great deal of attention in the computer vision community over the past decade. It is one of several fundamental topics in computer vision, e.g., intelligence video surveillance, self-driving vehicles, human computer interaction, and robotics and so on. Given the initialized object states (i.e., locations and sizes), the task of visual tracking is to estimate the incoming object states in subsequent frames. Despite much progress has been made in recent years, designing robust object tracking methods is still a challenging problem due to object appearance variations caused by illumination changes, occlusions, pose changes, cluttered scenes, moving backgrounds, etc. To effectively handling object appearance variations, many top-performing tracking methods have been proposed. According to several recent literature surveys on visual tracking [1–4], the performance of visual tracking can be significantly boosted via using more sophisticated features, more elaborate synergies between tracking and classification, segmentation or detection, and taking into account prior information of the scenes and the tracked objects. Handcrafted low-level visual features (e.g., color [5], subspace [6,7], Haar [8,9], LBP [10], SIFT [11], and HoG [12,13]) have been applied to these tracking methods to construct robust appearance models, and shown promising results for some specific data and tasks. However, they are low-level features and not tuned for the tracked object. Moreover, to capture the object appearance variations during tracking, designing effective features for object tracking usually requires reflecting the time-varying properties of object appearance model. However, most handcrafted features cannot be simply adapted according to the new observed data. Thus, online learning features from the object of interest is considered as a plausible way to remedy the limitation of handcrafted features.

In this paper, inspired by the recent success of deep learning on speech and object recognition problems [14–21], we propose a visual tracking method that relies on a deep neural network structure termed Network in Network (NIN) [21] to address both limitations of handcrafted features and little supervision (limited data) in visual tracking problem. More specifically, as shown in Fig. 1, the discriminative features are first automatically learned via a NIN, which is composed of multiple mlpconvs, in which a multilayer perceptron (MLP) is used to replaces the conventionally linear convolution filter to convolve over the input. With such a "micro network" structure (i.e., MLP), the NIN can effectively handle complexly nonlinear data. Consequently, the proposed tracking method can alleviate the need for handcrafted features, and automatically learn hierarchical and object-specific feature representations by end-to-end training. Then, the proposed tracking method simultaneously exploits both the samples collected in the initial and most previous frames to alleviate the tracker drifting problem.

☆ This paper has been recommended for acceptance by Luming Zhang.

* Corresponding author at: Jimei District, Huaqiao University, Xiamen, Fujian 361021, China.
E-mail address: yannychen@hqu.edu.cn (Y. Chen).

http://dx.doi.org/10.1016/j.jvcir.2015.10.004
1047-3203/

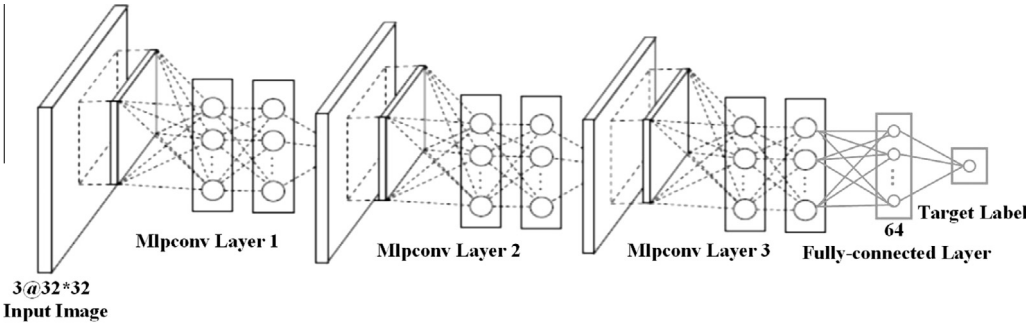

Fig. 1. Illustration of how the proposed tracking method constructs an object appearance model from a NIN [21] which is composed of multiple mlpconvs, in which a multilayer perceptron (MLP) is used to replaces the conventionally linear convolution filter to convolve over the input.

Finally, a heuristic schema is used to judge whether updating the object appearance models or not.

The rest of the paper is organized as follows. An overview of the related work is given in Section 2. Section 3 introduces how to learn a deep NIN-based object appearance model. The detailed tracking method is then described in Section 4. Experimental results are given in Section 5. Finally, we conclude this work in Section 6.

2. Related work

This section gives a brief review of related tracking methods using different types of object appearance models, namely, generative and discriminative methods. Moreover, some tracking methods based on deep learning are also briefly reviewed. Please refer to the survey papers [1–3] and a recent benchmark [4] for a comprehensive survey of visual tracking.

Although numerous methods have been proposed, robust visual tracking remains a significant challenge due to object appearance variations caused by non-rigid shapes, occlusions, illumination changes, cluttered scenes, etc. Recently, to build online object appearance models to account for the inevitable appearance changes of objects, numerous object representation strategies have been proposed. Typically, the online tracking methods using different types of object appearance models can be classified into two categories: generative and discriminative methods.

Generative tracking methods use some generative process to describe the object appearance models and make a decision based on the reconstruction errors or the matching scores. Popular generative methods include kernel-based tracker [5], subspace-based tracker [6,7], Gaussian mixture model-based tracker [22], covariance-based tracker [23], low-rank and sparse representation-based trackers [24–29], and visual tracking decomposition [30], and so on. Specifically, Comaniciu et al. [5] propose a kernel-based tracking method using color histogram-based object representations. To handle object appearance changes, Jepson et al. [22] use an online EM algorithm to construct an adaptive Gaussian mixture model-based object representation. In [6], an incremental principal component analysis (PCA)-based tracking method is proposed. In [23], Porikli et al. propose a tracking method using a covariance based object description and a Lie algebra based update mechanism. Recently, a variety of low-rank subspaces and sparse representations based tracking methods have been proposed [24–29] for object tracking due to their robustness to occlusion and image noises. Mei and Ling [24] are among the first to utilize the technique and propose an L1 minimization-based tracking method, in which the target is represented by a sparse linear combination of object templates and trivial templates. However, it is time consuming. To improve the time complexity, many extensions [25,26] have also been proposed. Zhang et al. [27] exploit the relationship between particles via low rank sparse learning for object tracking. In [28], Jia et al. propose a tracking method based on the structural local sparse appearance model. Zhang and Wong [29] propose a tracking method relying on mean shift, sparse coding and spatial pyramids. Kwon and Lee [30] propose an object tracking decomposition method which decomposes an observation model into multiple basic observation models to capture a wide range of pose and illumination changes. While generative tracking methods usually produce more accurate results under less complex environments due to the richer image representations used, they are prone to drift in more complex environments without utilizing the target and background information.

Unlike generative tracking methods, discriminative tracking methods learn to explicitly distinguish the target object from its surrounding background by formulating tracking as a binary classification problem. Based on the variance ratio of two classes, Collins et al. [31] select discriminative color features for constructing the adaptive appearance models. Avidan [32] uses an adaptive ensemble of weak classifiers for object tracking. Grabner and Bischof [8] propose an online boosting based tracking algorithm. Unfortunately, one inherent problem of online learning-based trackers is drift, a gradual adaptation of the tracker to non-targets. To alleviating the drifting problem, Matthews et al. [33] provide a partial solution for template trackers by assuming the current tracker does not stray too far from the initial appearance model. Combining with a prior classifier, Grabner et al. [34] propose a semi-supervised online boosting algorithm for object tracking. However, the method cannot accommodate very large changes in appearance due to using the prior classifier. Babenko et al. [35] propose a tracking method using a multiple instance boosting based appearance model. Hare et al. [9] design a tracking system based on an online kernelized structured output support vector machine. Santner et al. [36] propose a tracking method which combines an online random forest-based tracker, a correlation-based template tracker, and an optical-flow-based mean shift tracker in a cascade-style. Zhang and Maaten [37] propose a multi-object model-free tracker by using an online structured SVM algorithm to learn the spatial constraints along with the object detectors. Duffner and Garcia [38] present a method for tracking non-rigid objects, which combines and adapts a detector and a probabilistic segmentation method in a co-training manner. The co-training based algorithms require the independence among different features which is too strong to be met in practice. Kalal et al. [39] propose a TLD-based tracking system that explicitly decomposes the long-term tracking task into tracking, learning, and detection. Kwon and Lee [40] propose a visual tracker sampler, in which multiple appearance models, motion models, state representation types, and observation types are sampled via Markov Chain Monte Carlo to generate the sampled trackers. However, most of these