Journal of Visual Communication and Image Representation 55 (2018) 464–476

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Long-term correlation tracking using multi-layer hybrid features in sparse and dense environments[☆]

Nathanael L. Baisa[a,*], Deepayan Bhowmik[b], Andrew Wallace[a]

[a] School of Engineering and Physical Sciences, Heriot Watt University, Edinburgh, United Kingdom
[b] Department of Computing, Sheffield Hallam University, Sheffield, United Kingdom

ARTICLE INFO

Keywords:
Visual tracking
Correlation filter
CNN features
Hybrid features
Online learning
GM-PHD filter

ABSTRACT

Tracking a target of interest in both sparse and crowded environments is a challenging problem, not yet successfully addressed in the literature. In this paper, we propose a new long-term visual tracking algorithm, learning discriminative correlation filters and using an online classifier, to track a target of interest in both sparse and crowded video sequences. First, we learn a translation correlation filter using a multi-layer hybrid of convolutional neural networks (CNN) and traditional hand-crafted features. Second, we include a re-detection module for overcoming tracking failures due to long-term occlusions using online SVM and Gaussian mixture probability hypothesis density (GM-PHD) filter. Finally, we learn a scale correlation filter for estimating the scale of a target by constructing a target pyramid around the estimated or re-detected position using the HOG features. We carry out extensive experiments on both sparse and dense data sets which show that our method significantly outperforms state-of-the-art methods.

1. Introduction

Visual target tracking is one of the most important and active research areas in computer vision with a wide range of applications like surveillance, robotics and human-computer interaction (HCI). Although it has been studied extensively during past decades as recently surveyed in [1,2], object tracking is still a difficult problem due to many challenges that cause significant appearance changes of targets such as varying illumination, occlusion, pose variations, deformation, abrupt motion, background clutter, and high target densities (in crowded environments). Robust representation of target appearance is important to overcome these challenges.

Recently, convolutional neural network (CNN) features have demonstrated outstanding results on various recognition tasks [3,4]. Motivated by this, a few deep learning based trackers [5,6] have been developed. In addition, discriminative correlation filter-based trackers have achieved state-of-the-art results as surveyed in [7] in terms of both efficiency and robustness due to three reasons. First, efficient correlation operations are performed by replacing exhausted circular convolutions with element-wise multiplications in the frequency domain which can be computed using the fast Fourier transform (FFT) with very high speed. Second, thousands of negative samples around the target's environment can be efficiently incorporated through circular-shifting with the help of a circulant matrix. Third, training samples are regressed to soft labels of a Gaussian function (Gaussian-weighted labels) instead of binary labels alleviating sampling ambiguity. In fact, regression with class labels can be seen as classification. However, correlation filter-based trackers are susceptible to long-term occlusions.

In addition, the Gaussian mixture probability hypothesis density (GM-PHD) filter [8] has an in-built capability of removing clutter while filtering targets with very efficient speed without the need for explicit data association. Though this filter is designed for multi-target filtering, it is even preferable for single target filtering in scenes with challenging background clutter as well as clutter that comes from other targets not of current concern. This filtering approach is flexible, for instance, it has been extended for multiple targets of different types in [9,10].

In this work, we mainly focus on long-term tracking of a target of interest in sparse as well as crowded environments where the unknown target is initialized by a bounding box and then tracked in subsequent frames. Without making any constraint on the video scene, we develop a novel long-term online tracking algorithm that can close the research gap between sparse and crowded scenes tracking problems using the advantages of the correlation filters, a hybrid of multi-layer CNN and hand-crafted features, an incremental (online) support vector machine (SVM) classifier and a Gaussian mixture probability hypothesis density (GM-PHD) filter. To the best of our knowledge, nobody has adopted this

[☆] This paper has been recommended for acceptance by Christian Micheloni.

[*] Corresponding author.

E-mail address: nb30@hw.ac.uk (N.L. Baisa).

https://doi.org/10.1016/j.jvcir.2018.06.027
Received 14 September 2017; Received in revised form 16 March 2018; Accepted 30 June 2018
Available online 07 July 2018

approach.

The main contributions of this paper are as follows:

1. We integrate a hybrid of multi-layer CNN and traditional hand-crafted features for learning a translation correlation filter for estimating the target position in the next frame by extending a ridge regression for multi-layer features.
2. We include a re-detection module to re-initialize the tracker in case of tracking failures due to long-term occlusions by learning an incremental SVM from the most confident frames using hand-crafted features to generate high score detection proposals.
3. We incorporate a GM-PHD filter to temporally filter detection proposals generated from the learned online SVM to find the detection proposal with the maximum weight as the target position estimate by removing the other detection proposals as clutter.
4. We learn a scale correlation filter by constructing a target pyramid at the estimated or re-detected position using HOG features to estimate the scale of the detected target.

We presented a preliminary idea of this work in [11]. In this work, we make more elaborate descriptions of our algorithm. Besides, we include a scale estimation at the estimated target position as well as an extended experiment on a large-scale online object tracking benchmark (OOTB) in addition to the PETS 2009 data sets.

The rest of this paper is organized as follows. In Section 2, related work is discussed. An overview of our algorithm and the proposed algorithm in detail are described in Sections 3 and 4, respectively. In Section 5, the implementation details with parameter settings is briefly discussed. The experimental results are analyzed and compared in Section 6. The main conclusions and suggestions for future work are summarized in Section 7.

2. Related work

Various visual tracking algorithms have been proposed over the past decades to cope with tracking challenges, and they can be categorized into two types depending on the learning strategies: *generative* and *discriminative* methods. *Generative* methods describe the target appearances using generative models and search for target regions that best-fit the models i.e. search for the best-matching windows (patches). Various generative target appearance modelling algorithms have been proposed such as online density estimation [12], sparse representation [13,14], and incremental subspace learning [15]. On the other hand, *discriminative* methods build a model that distinguishes the target from the background. These algorithms typically learn classifiers based on online boosting [16], multiple instance learning [17], P-N learning [18], transfer learning [19], structured output SVMs [20] and combining multiple classifiers with different learning rates [21]. Background information is important for effective tracking as explored in [22,23] which means that more competing approaches are discriminative methods [24] though hybrid generative and discriminative models can also be used [25,26]. However, sampling ambiguity is one of the big problems in discriminative tracking methods which results in drifting. Recently, correlation filters [27–29] have been introduced for online target tracking that can alleviate this sampling ambiguity. Previously, the large training data required to train correlation filters prevented them from application to online visual tracking though correlation filters are effective for localization tasks. However, recently all the circular-shifted versions of input features have been considered with the help of a circulant matrix producing a large number training samples [27,28].

There are many strong sides of correlation methods such as inherent parallelism, shift (translation) invariance, noise robustness, and high discrimination ability [30]. Both digital and optical correlators are discussed in detail in [31] though more emphasis is given to optical correlators. Performance optimization of the correlation filters by preprocessing the input target image was introduced in [32]. Recent research trends of correlation filters for various applications with more emphasis on face recognition (and object tracking) is given in [30]. Due to the effectiveness of the correlation methods, they have been successfully applied to many domains such as swimmer tracking [33], pose invariant face recognition [34], road sign identification for advanced driver assistance [35], etc. Some types of correlation filters are sensitive to challenges such as rotation, illumination changes, occlusion, etc. For instance, the Phase-Only Filter (POF) is sensitive to changes in rotation, illumination changes, occlusion, scale and noise contained in targets of interest [32] though it can give very narrow correlation peaks (good localization); a pre-processing step was used to make it invariant to illumination in [34]. Recent correlation filters such as KCF [28] are more suitable for online tracking by generating a large number of training samples from input features using a circulant matrix and are more robust to the tracking challenges such as rotation, illumination changes, partial occlusion, deformation, fast motion, etc (as shown on its results section in [28]) than its previous counterparts [30]. Using CNN features has even improved the online tracking results as shown in [36] against these tracking challenges, however, log-term occlusion is still a problem in correlation filter-based tracking approaches.

There are three tracking scenarios that are important to consider: short-term tracking, long-term tracking, and tracking in a crowded scene. If objects are visible over the whole course of the sequences, short-term model-free tracking algorithms are sufficient to track a single object without applying a pre-trained model of target appearance. There are many short-term tracking algorithms developed in the literature [1,7] such as online density estimation [12], context-learning [37], scale estimation [29], and using features from multiple CNN layers [36,38]. However, these short-term tracking algorithms can not re-initialize the trackers once they fail due to long-term occlusions and confusions from background clutter.

Long-term tracking algorithms are important in video streams that are of indefinite length and have long-term occlusions. A Tracking-Learning-Detection (TLD) algorithm has been developed in [18] which explicitly decomposes the long-term tracking task into tracking, learning and detection. In this case, the tracker tracks the target from frame to frame and provides training data for the detector which re-initializes the tracker when it fails. The learning component estimates the detector's errors and then updates it for correction in the future. This algorithm works well in very sparse videos (video sequences with few targets) but is sensitive to background clutter. Long-term correlation tracking (LCT), developed in [39], learns three different discriminative correlation filters: translation, appearance and scale correlation filters using hand-crafted features. Even though it includes a re-detection module by learning the random ferns classifier online for re-initializing a tracker in case of tracking failures, it is not robust to long-term occlusions and background clutter. Multi-domain network (MDNet) [40] pre-trains a CNN network composed of shared layers and multiple domain-specific layers using a large set of videos to get generic target representations in the shared layers. This proposed network has separate branches of domain-specific layers for binary classification to identify the target in each domain. However, when applied to fundamentally different videos other than the related videos on which it was trained, it gives poorer results.

Tracking of a target in a crowded scene is very challenging due to long-term occlusion, many targets with appearance variation and high clutter. Person detection and tracking in crowds is formulated as a joint energy minimization problem by combining crowd density estimation and localization of individual person in [41]. Although this approach does not require manual initialization, it has low performance for tracking a generic target of interest as it was mainly developed for tracking human heads. The method developed in [42] trained Hidden Markov Models (HMMs) on motion patterns within a scene to capture spatial and temporal variations of motion in the crowd which is used for tracking individuals. However, this approach is limited to a crowd with