Neurocomputing 123 (2014) 344–353

Contents lists available at ScienceDirect

Neurocomputing

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

Robust tracking via patch-based appearance model and local background estimation

Bineng Zhong [a,*], Yan Chen [a], Yingju Shen [a], Yewang Chen [a], Zhen Cui [a], Rongrong Ji [b], Xiaotong Yuan [c], Duansheng Chen [a], Weibin Chen [a]

[a] Department of Computer Science and Engineering, Huaqiao University, Jimei District, Fujian 361021, China
[b] School of Information Science and Technology, Xiamen University, China
[c] School of Information and Control, Nanjing University of Information Science and Technology, China

ARTICLE INFO

Article history:
Received 27 January 2013
Received in revised form
28 April 2013
Accepted 23 June 2013
Communicated by Ran He
Available online 20 August 2013

Keywords:
Visual tracking
Background estimation
Context-awareness

ABSTRACT

In this paper, to simultaneously address the tracker drift and occlusion problem, we propose a robust visual tracking algorithm via a patch-based adaptive appearance model driven by local background estimation. Inspired by human visual mechanisms (i.e., context-awareness and attentional selection), an object is represented with a patch-based appearance model, in which each patch outputs a confidence map during the tracking. Then, these confidence maps are combined via a robust estimator to finally get more robust and accurate tracking results. Moreover, we present a local spatial co-occurrence based background modeling approach to automatically estimate the local context background model of an interested object captured from a single camera, which may be stationary or moving. Finally, we utilize local background estimation to provide supervision to an analysis of possible occlusions and the adaption of patch-based appearance model of an object. Qualitative and quantitative experimental results on challenging videos demonstrate the robustness of the proposed method.

1. Introduction

Visual tracking is a critical task in various applications such as intelligent video surveillance, human-computer interfaces, vision-based control, and so on. The two principle challenges in the tracking problem are: (1) appearance changes of the objects due to background clutters, pose and illumination variations, and (2) partial occlusion of objects of interest. To achieve desirable tracking results even in these challenging conditions, a large number of tracking methods have been proposed over the years.

To capture appearance changes of the objects, a number of authors [4–8] have formulated the problem of object tracking as an online learning problem, in which the object appearance models are adaptively updated. Unfortunately, one inherent problem of online learning-based tracking methods is drift, a gradual adaptation of the trackers to non-targets. Moreover, most of the tracking methods use the bounding box-based representations, which not only incorporate a large amount of noise or background during online learning as the objects typically do not cover the whole boxes, but also cannot explicitly handle the occlusion problem.

In this paper, to simultaneously handle the tracker drift and occlusion problem, we propose a robust visual tracking algorithm via a patch-based adaptive appearance model driven by local background estimation. The inspiration for this work comes from human visual mechanisms, i.e., the attentional selection mechanism among local regions according to context-awareness in a tracking process. To simulate the human context-aware mechanism, we firstly use local background estimation to capture local context information around an object. Then, an object is represented with a patch-based appearance model, in which each foreground patch is adaptively updated via an analysis of possible occlusions provided by local background estimation. In such a way, our tracker can simulate the human attentional selection mechanism. The key idea of our method is to utilize local background estimation to provide supervision to an analysis of possible occlusions and the adaption of patch-based appearance model of an object. Thus, our tracker can gradually adapt to appearance changes while effectively address the drifting and occlusion problem. Experimental results on challenging videos demonstrate the robustness of the proposed method.

The rest of the paper is organized as follows: Section 2 reviews the related work and Section 3 introduces the overview of the proposed tracking algorithm. Then, we describe the patched-based appearance model in Section 4. In Section 5, we briefly introduce the local spatial co-occurrence based background modeling

* Corresponding author. Tel.: +86 592 6162556.
E-mail addresses: bnzhong@gmail.com, zhongbineng@163.com (B. Zhong).

0925-2312/$ - see front matter
http://dx.doi.org/10.1016/j.neucom.2013.06.044

algorithm. The detailed algorithm of tracking and model updating is described in Section 6. Experimental results are given in Section 7, and we conclude in Section 8.

2. Related work

This section gives a brief overview of related tracking methods. The comprehensive review is beyond the scope of this paper. Please refer to [32–34] for more complete reviews on visual tracking and recent online learning-based tracking methods.

In the tracking literature, one popular technique is to track object using fixed appearance models [1–3]. These methods assume that object will look nearly identical in each new frame. Thus, an appearance model of the object from the first frame can be always used to describe object appearance. In [2], Comaniciu et al. use the space-weighted color histograms to represent the targets. Adam et al. [3] use the patch-based appearance models to handle the occlusion problems. However, these methods cannot achieve long-term persistent tracking in ever-changing environments, which often requires addressing difficult target appearance update problem. To handle this problem, a number of authors have formulated the problem of visual tracking as an online learning problem, in which the target appearance is updated adaptively using the images tracked from the previous frames. Collins et al. [4] present a method to adaptively select one color feature from several different color spaces to construct adaptive appearance models, which can best discriminate the object from the current background. In [5], Avidan proposes a method using an adaptive ensemble of classifiers for object appearance model maintenance and tracking. An adaptive tracking method which utilizes the incremental principal component analysis is presented in [6]. Han et al. [7] track an object by approximately estimating the pixel-wise color density in a sequential manner. In [8], an online boosting-based tracking method is proposed. Unfortunately, one inherent problem of online learning-based trackers is drift, a gradual adaptation of the tracker to non-targets.

Matthews et al. [9] have noticed the problem of tracker drift and provided a partial solution for template trackers. In [10], discriminative attentional regions are chosen on-the-fly as those best discriminate current object motion from background motion. In this way, tracker drift is unlikely since no on-line updates of attentional regions, and no new features are chosen after initialization in the first frame. Lu and Hager [11] propose model adaptation driven by feature matching and feature distinctiveness that may be robust to drift. He et al. [12] use a tracker based on online learning for key-point matching. They perform tracker update only when motion consensus of local SURF descriptors is verified. This method can alleviate the drift problem in some extent but it only works well for the tracking of texture-rich objects. Grabner et al. [13] apply an online semi-supervised boosting method to address the drift problem in visual tracking. Despite its success, the approach is limited by the fact that it cannot accommodate very large changes in appearance. In [14], Babenko et al. propose to use a multiple instance learning based appearance model for object tracking. Instead of using a single positive image patch to update a traditional discriminative classifier, they use one positive bag consisting of several image patches to update a multiple instance learning classifier. However, the method uses the bounding box-based representation, which incorporates a large amount of noise or background during online learning as the objects typically do not cover the whole box. In [15] and [16], co-training technique is applied to online multiple trackers learning with different features. The trackers collaboratively classify the new unlabeled samples and use these newly labeled samples with high confidence to update each other. However, occlusion is not explicitly handled in both approaches.

Recently, to explicitly handle the occlusion problem, a number of authors have proposed the Hough-based tracking methods. In [17], Gall et al. firstly propose the Hough forests for object detection, tracking, and action recognition. Then, Schulter et al. [18] extend the Hough forest to a on-line Hough Forest. Furthermore, Godec et al. [19] propose an online Hough-based method for tracking non-rigid objects, in which a back-projection scheme is used to roughly segment the tracked objects. However, this method may heavily rely on the prior information of the foreground and background. The false seeds of foreground or background may cause serious segmentation errors. In [44], Li et al. propose a tracking method via incremental Log-Euclidean Riemannian subspace learning, in which the covariance matrices of image features in the five modes are used to represent object appearance. However, the five modes can only capture the global and semi-global information of an object. Moreover, the computation of Log-Euclidean Riemannian metric is time-consuming.

Since shape is a powerful tool in image processing, a number of attempts have been made to combine tracking and segmentation for alleviating the drift problem. The idea behind these methods is that accurate foreground/background segmentation provides useful object contour constraints that can be helpful in alleviating model drift during online-learning of the tracker. Yin and Collins [20] propose a novel method to embed global shape information into local graph links in a Conditional Random Field (CRF) framework. Fan et al. [21] propose a matting-based tracking method, which relies on trackable points on both sides of the object boundary. Wang et al. [22] propose a superpixel tracking method by using mid-level cues that capture spatial information to some extent. However, in general, segmentation based methods only benefit from the situation when the foreground is in high contrast to the background, which is not always the case in natural scenes. In [42], Aeschliman et al. propose a probabilistic framework for joint segmentation and tracking.

A number of attempts have been made to utilize multiple observation models to improve the performance of a tracker. Santner et al. [23] propose a method which combines three trackers (i.e., an online random forest-based tracker, a correlation-based template tracker and an optical-flow-based mean shift tracker) in a cascade-style. However, how to set the overlapping and confidence thresholds that trigger the cascading process is crucial. In [24], a observation model is decomposed into multiple basic observation models to capture a wide range of pose and illumination changes. Kwon et al. [25] 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. Zhong et al. [26] propose a tracking method via weakly supervised learning from multiple imperfect oracles, in which several tracking algorithms are combined in a Bayesian framework.

With the popularity of sparse representations in image processing and machine learning, a variety of sparse representations based tracking methods have been recently proposed [27–29] for robust object tracking. Most sparse representation based trackers [27,28] only consider the holistic representation and hence may not handle partial occlusion or distracters. Differently, Jia et al. [29] propose a tracking method based on the structural local sparse appearance model, in which the patches sampled from all target images are used as the dictionary to sparsely represent the patches sampled from the candidates. The tracking accuracy is further improved by using the block-division spatial pooling schemes. However, the dictionary does not include the patches from the background images to describe the structure of the background, which may be helpful to discriminate the target from the background. In addition, Oron et al. [30] develop another method to