J. Vis. Commun. Image R. 25 (2014) 423–434

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Collaborative object tracking model with local sparse representation

Chengjun Xie [a,b], Jieqing Tan [a], Peng Chen [c,*], Jie Zhang [b], Lei He [a]

[a] School of Computer & Information, Hefei University of Technology, Hefei 230009, China
[b] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
[c] Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China

ARTICLE INFO

Article history:
Received 16 August 2013
Accepted 7 December 2013
Available online 16 December 2013

Keywords:
Object tracking
Discriminative model
Generative model
Sparse representation
Appearance model
Collaborative model
Sparse coding histogram
Similarity measure

ABSTRACT

There existed many visual tracking methods that are based on sparse representation model, most of them were either generative or discriminative, which made object tracking more difficult when objects have undergone large pose change, illumination variation or partial occlusion. To address this issue, in this paper we propose a collaborative object tracking model with local sparse representation. The key idea of our method is to develop a local sparse representation-based discriminative model (SRDM) and a local sparse representation-based generative model (SRGM). In the SRDM module, the appearance of a target is modeled by local sparse codes that can be formed as training data for a linear classifier to discriminate the target from the background. In the SRGM module, the appearance of the target is represented by sparse coding histogram and a sparse coding-based similarity measure is applied to compute the distance between histograms of a target candidate and the target template. Finally, a collaborative similarity measure is proposed for measuring the difference of the two models, and then the corresponding likelihood of the target candidates is input into a particle filter framework to estimate the target state sequentially over time in visual tracking. Experiments on some publicly available benchmarks of video sequences showed that our proposed tracker is robust and effective.

1. Introduction

Object tracking is one of the most important components in computer vision and arises in many practical applications such as video surveillance, human motion understanding, and interactive video processing, and so on. Although many trackers have been proposed and have made successes under various scenarios, object tracking is still challenging because the appearance of an object may be changed drastically while undergoing significant pose change, illumination variation and/or partial occlusion. Such a thorough review can be found in [1,12], where tracking algorithms were categorized into generative and discriminative approaches. Generative methods formulated the tracking problem as searching for the most similar regions to the target model. Discriminative methods treated the tracking problem as a binary classification problem which attempts to design a classifier to distinguish the target object from the background. In this paper, we concentrate mainly on designing a robust tracking model that confronts the aforementioned challenges by combining tracking outputs of the generative and discriminative models.

Recently, sparse representation [4] has been successfully applied in visual tracking, and a plethora of sparse representation-based tracking methods have been proposed [2,5–10,19,21,23]. Among these generative appearance models based on sparse representation, tracking problems were formulated to attempt to jointly estimate the target appearance by finding a sparse linear combination over a dictionary containing the target and trivial templates. Further experiments showed that sparse representation was efficient and adaptable to the aforementioned challenges, especially to partial occlusion. However, those sparse representation-based trackers only considered global templates, did not make full use of local representations, and hence failed in tracking target when the templates directly cropped from target image are very limited [24]. Therefore, Local patch-based sparse representation models were introduced in [5,7,9]. In [5,7] the object appearance was modeled by histograms of local sparse representation, however, both of their methods were based on a static local dictionary obtained from the first frame and may fail in dynamic scenes. Afterwards, Jia et al. [9] adopted an alignment pooling method scanning across local patches based on sparse coefficients for robust tracking. Although these trackers with local sparse representation have demonstrated good robustness in many videos, they may fail in the discrimination between the target and the background more possible when there are some challenging factors, such as background clutter and the background regions with similar

* Corresponding author.
E-mail addresses: cjxie@iim.ac.cn (C. Xie), bigeagle@mail.ustc.edu.cn (P. Chen).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.12.012

appearance to the object class. In comparison, discriminative appearance models based on sparse representation posed visual object tracking as a binary classification issue. In [2], local image patches of a target object were represented by their sparse codes with an over-complete dictionary constructed online, and the sparse codes were treated as training samples. The key idea of the model was to train a classifier by learning the sparse codes and to maximize the separability between the object and non-object regions discriminately. Nevertheless, a major limitation of the discriminative appearance models is in that they were heavily relied on training sample selection. Moreover, most discriminative appearance models took the current object location as one positive sample, and its neighborhoods as negatives. However, the imprecise current object location could degrade the appearance model and cause drift.

In actually, generative and discriminative appearance models have their respective pros and cons, and are complementary to each other to a certain extent. Therefore, we propose an efficient tracking algorithm incorporating the information from a developed generative model and a discriminative model, based on local sparse representation. Our proposed algorithm mainly contain a local dictionary obtained by sampling local image patches within the target region from the first frame; a sparse representation-based generative model (SRGM) with local sparse template which is represented by histograms of local sparse representation based on the dictionary and is updated online; a sparse representation-based discriminative model (SRDM) with a linear classifier which is trained by learning the sparse codes based on the dictionary and is updated online; and a similarity function fusing the information from the generative model and the discriminative appearance model. The discriminative model is able to investigate informative samples as support vectors for object/non-object classification, resulting in a strong discrimination. Although our SRDM module has a good generalization ability to distinguish object and background, the local sparse representation-based classifier may be affected when updated with the background information as positive samples. However, the SRGM module can alleviate the influence since it is distinct to be foreground or background with local sparse coding histograms. Thus, the SRDM and SRGM module are complementary to each other to some extent.

The main contributions of this paper are:

- A novel target appearance modeling method by combining the generative model and the discriminative appearance model based on local sparse representation.
- A new similarity measure between the candidates by fusing the information from the generative model and the discriminative appearance model.

2. Related works

There was a rich literature on appearance modeling and representation [12]. An effective object representation should have a strong description or discrimination ability to distinguish targets from background. Most of recent tracking algorithms focused on object representation schemes with generative appearance models [3,6–10,13,14,16,17,19–23,33,38] and discriminative models [2,5,11,15,25,26,37].

Generative methods formulated the tracking problem as searching for the most similar regions to the target model. Intensity histogram was perhaps the simplest way to represent object appearance in many tracking algorithms [3], but it missed the spatial information of object appearance, which makes it sensitive to noise as well as occlusion in many tracking applications. To solve these problems, Nejhum et al. [17] modeled the target appearance as a small number of rectangular blocks with histograms, whose positions within the tracking window are determined adaptively. More recently, He et al. [22] presented a tracking framework based on a locality sensitive histogram that was computed at each pixel location and a floating-point value was added to the corresponding bin for each occurrence of an intensity value. In addition, to cover a wide range of pose and illumination variation, Ross et al. developed an online subspace learning model to account for appearance variation [13]. Recently, the sparse representation framework [4,18] has attracted considerable interests in object tracking due to its robustness to occlusion and image noise. Following the pioneer work, many methods adopted sparse representation model for tracking objects [5–10,19–21,23]. In [6], each target candidate was represented as a linear combination of a set of online updated templates, consisting of target templates and trivial templates, and the candidate with the smallest error to target template reconstruction is regarded as the tracking result. More recently, Zhang et al. [8] presented a multi-task sparse optimization framework. Instead of treating test samples independently, the framework explored the interdependencies between test samples by solving a regularized group sparsity problem. Besides the high computational cost, another drawback of these trackers is to model object appearance as global sparse templates. Since local representations can capture the local structural target appearance, the local visual representations [7,9,22] were robust to global appearance changes caused by illumination variation, shape deformation, and partial occlusion. In [24], extensive experiments have demonstrated that local sparse representation-based trackers outperformed those with global sparse templates. Therefore, Jia et al. [9] adopted an alignment pooling method scanning across local patches based on sparse coefficients for robust tracking. As these methods exploited generative representation of target objects only and did not take the background into account, they were less effective for tracking in cluttered background.

By training a model via a discriminative classifier, discriminative methods [2,11,15,25,26,37] have shown good performance in discriminating object from the background. Avidan et al. [11] developed an online boosting method for tracking targets, which was an ensemble tracker that yielded a strong classifier by a set of weak classifiers. Bai et al. [25] treated object tracking as a weakly supervised ranking problem, which can avoid the heuristic and unreliable step of training sample selection towards the true target samples. In contrast with them, Babenko et al. [15] used multiple instance learning (MIL) instead of traditional supervised learning to handle ambiguous binary data obtained online. Zhang et al. [26] proposed an online weighted multiple instance tracker, which incorporated the important information of samples into the online multi-instance boosting learning process, resulting in robust tracking results. Despite the success of the discriminative methods, a major challenge is how to choose positive and negative samples when updating the adaptive appearance model. Since most discriminative trackers took the current object location as one positive sample and sampled its neighbors as negatives, it might degrade the appearance model and cause drift due to the imprecise current object location.

In this paper, we propose an effective object tacking method involving a generative model and a discriminative appearance model based on local sparse representation. The proposed method consists of three main parts: a generative appearance model for object representation, which is composed of local patch templates with the corresponding histograms of local sparse representation, and thus provides a more flexible mechanism to deal with the problem of appearance change; a discriminative appearance model for object representation, which is obtained by learning the local sparse codes of the negative and positive samples, and thus is capable of discriminating object from the background powerfully; a similarity mea-