Knowledge-Based Systems 100 (2016) 13–24

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Discriminative local collaborative representation for online object tracking

Si Chen[a,b], Shaozi Li[a,*], Rongrong Ji[a], Yan Yan[a], Shunzhi Zhu[b]

[a] School of Information Science and Engineering, Xiamen University, Xiamen 361005, China
[b] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China

ARTICLE INFO

Article history:
Received 14 September 2015
Revised 15 December 2015
Accepted 30 January 2016
Available online 2 March 2016

Keywords:
Object tracking
Online learning
Collaborative representation
Local coding
Discriminative tracking

ABSTRACT

Sparse representation has been widely applied to object tracking. However, most sparse representation based trackers only use the holistic template to encode the candidates, where the discriminative information to separate the target from the background is ignored. In addition, the sparsity assumption with the l_1 norm minimization is computationally expensive. In this paper, we propose a robust discriminative local collaborative (DLC) representation algorithm for online object tracking. DLC collaboratively uses the local image patches of both the target templates and the background ones to encode the candidates by an efficient local regularized least square solver with the l_2 norm minimization, where the feature vectors are obtained by employing an effective discriminative-pooling method. Furthermore, we formulate the tracking as a discriminative classification problem, where the classifier is online updated by using the candidates predicted according to the residuals of their local patches. To adapt to the appearance changes, we iteratively update the dictionary with the foreground and background templates from the current frame and take occlusions into account as well. Experimental results demonstrate that our proposed algorithm performs favorably against the state-of-the-art trackers on several challenging video sequences.

1. Introduction

Object tracking plays a critical role in computer vision with wide applications in video surveillance, human computer interaction, vehicle navigation, etc. [1–3]. Given the target object at the first frame, the goal of tracking is to locate this object for the subsequent frames. Despite great progresses made in the last two decades [4], object tracking remains a challenging problem, which is mainly due to the presences of illumination variation, occlusion, fast motion, scale change, as well as background clutter.

Generative and discriminative methods are two major categories for object tracking. Generative methods [5–15] focus on using the appearance model to search for the regions similar to the tracked targets, for example, incremental visual tracking (IVT) [5], adaptive structural local sparse appearance model (ASLSA) [6], visual tracking decomposition (VTD) [7], L1 tracker using accelerated proximal gradient approach (L1-APG) [8] and tracking using local sparse appearance model and k-selection (LSK) [9]. Among these, most trackers are deployed based on sparse representation [8–15] and only maintain target templates to build the appearance model. In these approaches, the sparse assumption is imposed on the data by solving the time-consuming l_1 norm minimization.

Discriminative methods [16–22], on the other hand, cast tracking as a classification problem to separate the target from the background, such as online boosting (OAB) [16,17], online semi-supervised boosting (SemiB) [18], tracking-by-detection with circulant structure (CSK) [19], structured output tracking with kernels (Struck) [20], online multiple instance learning (MIL) [21] and online weighted multiple instance learning (WMIL) [22]. Some discriminative methods are designed based on the online boosting framework. In this framework, a large pool of simple features is generated to iteratively train the weak classifiers online, based on which the best ones are selected over each frame during tracking. Some recent work is further proposed to combine the generative and discriminative methods, e.g., compressive tracking (CT) [23], sparsity-based collaborative model (SCM) [24] and discriminative tracking with local sparse representation (DLSR) [25]. However, the background information, which is critical for effective tracking, is ignored for appearance modeling and the target object is only represented based on the target templates.

In this paper, we propose a robust discriminative local collaborative representation based online object tracking algorithm,

[*] Corresponding author. Tel.: +86 592 2580080; fax: +86 592 2580258.
E-mail addresses: chensi@xmut.edu.cn (S. Chen), szlig@xmu.edu.cn (S. Li).

http://dx.doi.org/10.1016/j.knosys.2016.01.041
0950-7051/© 2016 Elsevier B.V. All rights reserved.

termed DLC. In this algorithm, a novel discriminative local collaborative representation is first proposed by using the local image patches of both the target templates and the background ones to collaboratively encode the candidates via a local regularized least square optimization, where the efficient l_2 norm minimization is leveraged to replace the traditional time-consuming l_1 norm minimization. DLC further obtains the final feature vector representing the candidate by adopting an effective discriminative-pooling strategy. Finally, DLC trains a discriminative classifier based on our discriminative local collaborative representation to distinguish the target from the surrounding background. This classifier is online updated with the candidates from the current frame, whose labels are predicted according to the local regularized residuals of the local image patches. Along with tracking, the dictionary and templates are iteratively updated by using the new tracking result and its surrounding background samples. The adopted update scheme can avoid the influence of occlusions. Both qualitative and quantitative evaluations on challenging video sequences demonstrate that the proposed tracking algorithm outperforms the state-of-the-art trackers.

The main contributions of this work are summarized as follows.

(1) Different from the existing sparse representation methods, our object representation method makes full use of the local and the background information to encode the candidates with the efficient l_2 norm minimization, and thereby it achieves very competitive tracking performance and the low computational cost.
(2) As for the local codes, the proposed discriminative-pooling strategy can better describe the similarity between each candidate and the target, thus improving the discriminability of our algorithm.
(3) By combining generative and discriminative models, the proposed object representation method can facilitate the discriminative classifier to effectively separate the foreground from the cluttered background.
(4) The dictionary update approach not only adapts to the dramatic appearance changes, but also improves the robustness of appearance representation when occlusions occur.

The paper is organized as follows: Section 2 reviews the related work and describes the motivation of our work. Our proposed DLC tracking algorithm with discriminative local collaborative representation is introduced in Section 3 in details. Section 4 presents the experiments and Section 5 concludes this paper.

2. Related work

Current online tracking algorithms can be coarsely categorized into generative and discriminative methods. Generative methods [5–15] learn an appearance model to represent the target object and then use it to search for the candidate with the minimal reconstruction error, such as using sparse representation. Discriminative methods [16–22], treat tracking as a classification problem, which popularly performs a tracking-by-detection process to find the candidate with the maximal classification response as the tracking result. Both generative and discriminative approaches can be combined to further improve the tracking performance [23–30].

2.1. Sparse representation based generative methods

Recently, sparse representation [31–33] has largely promoted the development of generative tracking methods [8–11]. Under the sparsity assumption, a signal is encoded over a dictionary such that it can be represented in the form of a linear combination of only a few basis vectors [31]. The sparse codes can be measured by l_0 norm, which counts the number of non-zeros in the sparse vector. Since the l_0 norm minimization is NP hard, the l_1 norm minimization [34], as the closest convex function to the l_0 norm minimization, is widely employed in sparse representation based trackers [8–11]. For example, the L1 trackers [10, 11] are proposed, where the target candidate is modeled by a sparse linear combination of target and trivial templates. In [9], Liu et al. employed the histograms of local sparse representation and used mean-shift to locate the target object. Bao et al. [8] developed a refined l_1 norm minimization model by adding the l_2 norm regularization, which can be solved by using an accelerated proximal gradient descent method.

To obtain the final histogram feature vector over the local codes, various pooling methods are proposed. In [25], the final feature vector is obtained by concatenating all local codes, called concatenating-pooling, which is sensitive to image noises and the dimensionality of the final feature vector is extremely high. In [35], a max-pooling operator is used to compute the final feature vector whose element is the maximum coefficient of each basis over all patches. The max-pooling method loses the discriminative ability because it ignores other non-maximum coefficients. In [36], Zhang et al. exploited an average-pooling method over local codes, but it loses the spatial layout of the corresponding local patches. In [6], Jia et al. proposed a structural local sparse coding model with an alignment-pooling method, termed ASLSA, where it is assumed that the local appearance variation of a patch is described by the blocks at the same positions of the template. However, the alignment-pooling method is not effective enough to handle the dramatic pose or scale variations.

Despite much demonstrated success of these sparse representation methods in tracking, several problems remain open. First, although the l_1 norm minimization is much more efficient than the l_0 norm minimization, it is still time-consuming. Second, most existing methods only consider the holistic representation of the target, while local information that is important to handle the dramatic appearance changes is usually ignored. Third, the background information is neither used to encode the candidates nor considered in the dictionary construction.

2.2. Discriminative tracking methods

One popular trend in discriminative methods is the so-called tracking-by-detection [16–19,21,22], which iteratively carries out two steps, i.e., (1) classifier updating with the samples drawn from the current frame and (2) object detection using the updated classifier at the next frame. It is clear that the online updating scheme serves as the key towards robust tracking.

In [16], the online boosting framework was proposed to boost the multiple weak classifiers constructed on several pools of features. However, only one positive sample and a few negative samples are used to update the classifier, which may lead to the drifting problem due to the mis-aligned samples introduced. In [17], a global feature pool is shared for all selectors to speed up the online boosting process. In [18], Grabner et al. developed an online semi-supervised boosting algorithm, which uses the unlabeled samples drawn from the current frame to update the classifier. However, the unlabeled samples are predicted only relying on the labeled samples from the first frame, where the information from the subsequent frames is ignored. Babenko et al. [21] proposed an online MIL tracker to handle the inherent ambiguity of labeling, which introduces multiple instance learning to update the classifier by using positive and negative bags. Zhang and Song [22] developed an online WMIL tracker which considers the importance of the instances in the positive bag during the learning process. However, both the MIL and WMIL trackers [21,22] only update the classifier with the positive labels for all the instances in the positive bag, thereby degrading the tracking performance.