Neurocomputing 275 (2018) 1801–1812

Contents lists available at ScienceDirect

Neurocomputing

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

Online discriminative dictionary learning for robust object tracking

Tao Zhou[a], Fanghui Liu[a], Harish Bhaskar[b,e], Jie Yang[a,*], Huanlong Zhang[c], Ping Cai[d]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[b] Department of Electrical and Computer Engineering, Khalifa University of Science Technology and Research, United Arab Emirates
[c] College of Electric and Information Engineering, Zhengzhou University of Light Industry, China
[d] Department of Instrument Science and Engineering, Shanghai Jiao Tong University, China
[e] Zero One Infinity Consulting (ZOIC) Service Ltd., Canada

ARTICLE INFO

Article history:
Received 8 February 2017
Revised 9 August 2017
Accepted 23 October 2017
Available online 29 October 2017

Communicated by Dr Zhu Jianke

Keywords:
Discriminative dictionary learning
Visual tracking
Class-specific sub-dictionary
Likelihood function

ABSTRACT

The discriminative ability of dictionary learning algorithms plays a crucial role in various computer vision applications, particularly in visual object tracking. In this paper, a novel visual tracking algorithm based on an online discriminative dictionary learning technique is proposed. The proposed method incorporates target and background information into dictionary learning in order to separate the target-of-interest from a cluttered background effectively. The dictionary thus learnt, can ensure that each class-specific sub-dictionary has a good representation of the samples associated with its own class and a poor representation of the other classes. In contrast to other dictionary learning mechanisms, the proposed method also introduces an error term that aims to capture outliers (e.g., noise and occlusion) and minimize its effect on tracking. Furthermore, by optimizing a constrained objective function, the learnt dictionary is rendered robust and discriminative, thereby resulting in an accurate tracking framework that can efficiently separate the target from the background. Finally, an effective and simple observation likelihood function based on the reconstruction errors from both positive and negative templates is designed to achieve better tracking performance. Experimental results on a publicly available benchmark dataset demonstrate that the proposed tracking algorithm performs better than several baseline trackers.

1. Introduction

Visual tracking is a challenging research task within computer vision applications. Visual tracking is often complicated due to factors such as changes in lighting conditions, partial or full occlusion, background clutter and target deformation [1–4]. A significant progress has been made in the past decade, for example in Refs. [5–13], in developing a efficient and robust tracking algorithms. Despite such advances, tracking continues to remain a complex task particularly due to appearance variations of the target that make it difficult to distinguish it from the background. In this context, discriminative dictionary learning (DDL) methods have recently been proposed and successfully applied to various computer vision applications such as image restoration, object classification and face recognition, among others.

The aim of DDL is to learn a set of basis from the training data in a manner that the dictionary would faithfully represent a given visual signal. Much research efforts have been spent in developing robust DDL algorithms for distinctively representing test samples. Recently, an unsupervised dictionary learning method based on K-Singular Value Decomposition (SVD) [14], has demonstrated promising performance for image restoration applications. Further, some other discriminative methods [15,16] based on K-SVD have also been developed. Such techniques aim to optimize coding coefficients learned from a discriminative dictionary for classification. In the study by Yang et al. [17], a Fisher criterion was introduced to obtain sub-dictionaries of different classes to learn a structured dictionary. As discussed in [17,18], each sub-dictionary is aimed to provide a good representation ability of the training samples from the associated class and a poor representation ability for other classes. Yang et al. [19] proposed a novel model for dictionary pair learning, in which block-diagonal structures for the coding coefficient matrix and the analysis dictionary have been enforced. In another study, Feng et al. [20] proposed a novel formulation based on graph Laplacian constraint that enabled the exact construction of block-diagonal affinity matrices. One limitation of these existing methods is that they do not consider the effect of noise and outliers, particularly when an occlusion occurs. Without the estimation of such outliers, the learned dictionary cannot effectively estimate coding coefficients with a low measurable error. Further, without a well-specified energy constraint for each atom

* Corresponding author.
E-mail address: jieyang@sjtu.edu.cn (J. Yang).

https://doi.org/10.1016/j.neucom.2017.10.019
0925-2312/

within the learned dictionary, it is not possible to learn a robust and discriminative dictionary.

Despite limitations, DDL continues to be considered a critical part of achieving good tracking performance. Several DDL based trackers [21–24] have been developed to achieve better tracking performance by casting the tracking problem as a binary classification task. However, these trackers always have run into difficulties while distinguishing similar visual patterns, especially for targets that share a similar shape and visual appearance with the background clutter. Some of the other main limitations of the DDL-based tracking models are as follows. First, some methods directly select patches from the target region to constitute the positive dictionary [25], while patches selected away from the target region are considered a part of the negative dictionary. In this case, the dictionaries are not robust enough to represent the estimated target candidate, and the coefficients obtained over the dictionaries are also not discriminative. Second, negative samples that are introduced with outliers could easily result in tracking drifts. Therefore, it is necessary to capture such outliers and minimize their effect during tracking. Finally, some trackers also engage static dictionary [26] or constraint-specific heuristic dictionary that is updated directly using new observation samples as dictionary items [7,25]. Such dictionary update schemes do not allow the tracker to cope with appearance or pose changes.

Motivated by some of these challenges aforementioned, in this paper, a novel tracking framework via online discriminative dictionary learning is proposed. The proposed framework is an online learning algorithm that aims to obtain a robust and discriminative dictionary, in which the sub-dictionary associated with the target has good representation ability of the target and a poor representation ability for the background. The proposed algorithm will take into account background information, which when integrated, can adaptively separate the target from the background thereby being more robust against appearance variations of a target. Constraints are imposed to ensure that coding coefficients of all samples from the same class are closer to their mean, thereby reducing the variation of the coding coefficients within each class. Under this scheme, the coding coefficients are made to have smaller within-class scatter, but larger between-class scatter, thus making the learned dictionary more discriminative and help in achieving a better tracking performance. Also, an error term that can capture outliers (e.g., noise and occlusion) is introduced in the dictionary learning step which reduces the effect of outliers efficiently and enables the tracker to avoid drifts. Further, with an energy constraint imposed on each atom of the learned dictionary, it is demonstrated that a more accurate and stable dictionary can be built, and thus obtain more discriminative coding coefficients of the test samples over the learned dictionary. The basic flow diagram of the proposed tracking algorithm is shown in Fig. 1.

In what follows, the related work is described in Section 2. A detailed description of the proposed discriminative dictionary learning strategy is presented in Section 3. Further, the tracking framework that incorporates the learned discriminative dictionary is detailed in Section 4. Finally, experimental results on public datasets are presented and analyzed in Section 5 followed by some brief conclusive remarks in Section 6.

2. Related work

Target tracking via online subspace learning [27,28] has received significant attention in recent years. For example, Ross et al. [12] proposed the Incremental Visual Tracking (IVT) method, an online approach to efficiently learn and update a low dimensional PCA subspace representation of the target. Although the IVT method based on subspace coding has been proven to be robust to lighting and pose changes, it also has demonstrated sensitivity to partial occlusion and background clutter. The primary reason for this sensitivity is due to the noise term that cannot be modeled effectively using small variances, particularly when outliers exists.

Sparse representation (SR) has also been used in visual tracking to achieve promising results against target appearance variations and occlusion. SR, for the first time, was introduced into tracking by Mei and Ling [29]. Here, a candidate was constructed using both the target and trivial templates. Li et al. [30] had used compressed theory to reduce the computational overhead of real-time tracking through adaptive sparse signal recovery. Similarly, Wang et al. [31] engaged a trained linear classifier using sparse codes with an over-complete dictionary for image patch-based target representation. However, the ℓ_1 optimization resulted in a high computational load, hence could not satisfy the real-time demand for visual tracking. Also, the above methods did not take into consideration the background information, and therefore could not distinguish the target from any complex background due to the lack of discriminative power. Li et al. [30] proposed to adopt background templates rather than the over-complete dictionary [29]. The method was helpful in improving the tracking performance particularly because the representations of the target and the background were well separated. Zhong et al. [7] introduced background templates to establish a sparsity-based discriminative classifier, which could exploit the distinction between the foreground and the background to improve tracking performance. Although these methods considered the use of background information [7,30], the dictionary remained either fixed or updated using selected patches from the target and background directly. In this manner, the dictionary was not robust to represent the estimated target candidate, and the coefficients obtained over the dictionaries were also not discriminative.

Sparse coding techniques are usually challenged during the process of dictionary learning [32]. The primary aim of dictionary learning is to effectively use the coding coefficients obtained from a discriminative dictionary as features to distinguish the target from the background. Most current dictionary learning methods are based either on online dictionary learning scheme [33,34] or SVD [14]. For tracking, Wang et al. [21] have proposed a robust non-negative dictionary learning algorithm for updating target templates online so that each learned template could capture a distinctive aspect of the tracked target. In [35], the dictionary was directly learned from patches that were sampled from a target image. Xing et al. [22] proposed to learn a multi-lifespan dictionary that effectively utilized tracking data. However, the state-of-the-art methods did not consider handling outliers (e.g., noise and occlusion), hence resulted in inaccurate dictionary learning and tracking drift. To handle this issue and effectively capture noise, Liu et al. [36] proposed to introduce a sparse noise matrix into low-rank representation. Further extensions including the use of ℓ_1, ℓ_2 and $\ell_{2,1}$-norms on the noise matrix proved promising results. In [37], the ℓ_2-norm was replaced by $\ell_{2,1}$-norm within the Non-negative Matrix Factorization (NMF) reconstruction function to make the NMF robust to outliers. Besides, some related research has also indicated that the $\ell_{2,1}$-norm has been especially helpful for handling noise-like outliers [38].

3. The proposed online discriminative dictionary learning algorithm

Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N] \in \mathbb{R}^{l\times N}$ be a set of the N samples, $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_1, \ldots, \mathbf{d}_M] \in \mathbb{R}^{l\times M}$ be the dictionary where each column represents an atom, and $\mathbf{C}$ is the coefficient matrix of $\mathbf{X}$ over $\mathbf{D}$. The goal of dictionary learning is to obtain a dictionary, such that each sample can be linearly reconstructed using a relatively smaller subset of dictionary atoms, while keeping the reconstruction error as