J. Vis. Commun. Image R. 28 (2015) 44–52

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Object tracking based on local dynamic sparse model ☆

Zhangjian Ji, Weiqiang Wang *

School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 31 October 2014
Accepted 11 January 2015
Available online 22 January 2015

Keywords:
Object tracking
Local dynamic sparsity
Multi-task sparse model
Occlusion detection
Weighted likelihood probability
Accelerated proximal gradient
Particle filter
Sparse representation

ABSTRACT

Sparse representation has been widely applied in many objecting tracking methods. In this paper, we present a robust and effective object tracking approach based on the local dynamic sparse model, called Local Dynamic Sparse Tracking (LDST). In the proposed method, the local patches of a tracked object are linearly represented by their respective dictionary updated online, and the inter-frame correlation between sparse representations of corresponding patches are modeled in the time domain. To further improve its robustness, the dependency of sparse coefficients between patches in each frame is also characterized by the $\ell_{1,2}$ mixed norms. In addition, for each patch, different weights are exploited in calculating the likelihood probability, in order to eliminate the effect of occluded patches when updating templates. The evaluation experiments on the challenging sequences demonstrate that the proposed method has the better performance compared with some typical state-of-the-art methods.

1. Introduction

Object tracking has long been an important problem in computer vision field, since it plays a crucial role in many practical applications, such as intelligent surveillance, autonomous navigation of vehicles, human–computer interaction, and action recognition. In real-world scenes, a robust object tracking algorithm must handle some challenging factors well, e.g., partial occlusion, illumination variation, pose changes, background clutter, complex motion and object blur. In the last decade, a lot of tracking methods have been presented to deal with the challenges [1].

Generally, the existing object tracking methods can be categorized into two categories, i.e., generative approaches [2–7] and discriminative [8–10] approaches. Due to space limitation, we only briefly review the methods which are most related to ours. Recently, due to the success of sparse representation in face recognition [11], some object tracking methods based on sparse representation have been proposed [12–16]. For example, Mei et al. [14,15] present the holistic appearance of the object can be represented by a series of target templates and trivial templates, and then the object tracking can be formulated as solving the ℓ_1 minimization problem. However, the computational cost of the ℓ_1 minimization is very high. In order to make it suitable for real scenes, the accelerated proximal gradient algorithm (APG) [16] has been proposed to speedup the tracking methods based on the ℓ_1 minimization problem. In [17], the dynamic group sparsity is introduced into the sparse representation to enhance the robustness of the tracker. It should be noted that these methods assume that sparse representations of particles are independent. Ignoring the correlation of sparse representations among particles makes the tracker more prone to drift away from the target. More recently, Zhang et al. [18]. propose an efficient multi-task sparse learning approach for robust visual tracking, and they adopt the $l_{p,q}$ mixed norms to describe the joint representation of all the particles. Further, in order to make tracking methods based on sparse code more efficient, Zhang et al. reduce the dimension of features by exploiting compressed sensing principles [19].

To better handle the temporal variation and occlusion situations, the learning schemes with dynamic update have been exploited by many tracking systems [14,15,18,20–22]. Different from the template updating strategy in [14–16], Wang et al. [23] blend the past information and the current tracking result in a principled way by non-negative dictionary learning strategy, which can automatically detect and reject the occlusion and yield robust object templates. To handle the occlusion situation, Kwak et al. [24] present an active occlusion detection and processing algorithm by learning a general classifier with observation likelihoods. Liu et al. [25] propose a tracking algorithm based on local sparse model which employs the histograms of sparse coefficients and the mean shift algorithm in object tracking. However, this method is based on a static local sparse dictionary and may fail when there are objects with similar appearance in scenes. Jia et al. [26] propose an efficient tracking algorithm with structural

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author.

E-mail addresses: jizhangjian08@mails.ucas.ac.cn (Z. Ji), wqwang@ucas.ac.cn (W. Wang).

http://dx.doi.org/10.1016/j.jvcir.2015.01.008
1047-3203/© 2015 Elsevier Inc. All rights reserved.

local sparse model and adaptive template update strategy. Meanwhile, in order to alleviate the drifting problem, this method uses the alignment pooling in the sparse representation of tracked object. Zhong et al. [27] present a robust object tracking method using a collaborative model, which utilizes both holistic templates and local representations in each frame. However, these approaches do not exploit the temporal consistency of sparse representation of tracked objects, and the stability of target tracking can be affected when the occlusion or appearance change occurs.

In this paper, we present a robust and effective object tracking approach based on the local dynamic sparse model in the particle filer framework. The highlights in this contribution are summarized as follows. First, a sparse innovation term is introduced into the objective function to characterize the inter-frame correlation between sparse representations of corresponding patch in the time domain. Second, we adopt the $\ell_{1,2}$ mixed norm to characterize the dependencies of sparse coefficients between patches in each frame to enhance the robustness of the proposed method. Third, we eliminate the influence of occlusion patches in the update of templates, and in calculating the likelihood probability, we give different weights of different patches.

This paper is organized as follows. Section 2 formulate the proposed local dynamic sparse tracking model in details. Then it is evaluated and the related experimental results are reported in Section 3. Section 4 concludes the whole work.

2. Local dynamic sparse tracking model

This section presents the proposed object tracker based on local dynamic sparse model. First, we introduce the sparse representation of tracked objects and formulate dynamic sparse tracking model in Section 2.1. Further, the local sparse representation based on patches of tracked object and the corresponding local dynamic sparse tracking model are described in details in Section 2.2. Then we present how to exploit the Bayesian inference framework to estimate the target states, i.e., motion parameters, in Section 2.3. A new template update scheme is presented to reduce the influence of occlusion in Section 2.4. Finally, we introduce how to utilize the Accelerated Proximal Gradient (APG) method to solve the proposed local dynamic sparse tracking model in Section 2.5.

2.1. Sparse representation of objects

It is known that the object appearance under different environments can be approximately embedded in a low dimensional subspace. Here the appearance subspace can be spanned by a group of representative images, called templates, $\mathbf{t}^i \in \mathbb{R}^d, i = 1, 2, \ldots, n$. Given the image set of target templates at time $t, \mathbf{T}_t = [\mathbf{t}_t^1, \mathbf{t}_t^2, \ldots, \mathbf{t}_t^n] \in \mathbb{R}^{d \times n}, d \gg n$, A candidate particle $\mathbf{y}_t \in \mathbb{R}^d$ at time t can be approximated by the linear combination of target templates in $\mathbf{T}_t$, that is,

$$\mathbf{y}_t = \mathbf{T}_t\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ denotes the column vector of target coefficients, and column vector $\mathbf{e} \in \mathbb{R}^d$ represents the error or residual term. The large nonzero entry in $\mathbf{e}$ indicates the corresponding pixel in $\mathbf{y}_t$ is possibly corrupted or occluded. In [14], Mei et al. adopt the trivial templates $\mathbf{I}$ to capture the occlusion, so Eq. (1) can be rewritten as,

$$\mathbf{y}_t = [\mathbf{T}_t \quad \mathbf{I}]\begin{bmatrix}\mathbf{x}\\ \mathbf{e}\end{bmatrix}, \tag{2}$$

where $\mathbf{I}$ is a $d \times d$ identity matrix, and here $\mathbf{e}$ should be sparse, since outliers usually occupy very few pixel locations of tracked objects. Consequently, for a candidate target $\mathbf{y}_t$, the sparse solution of Eq. (2) can be obtained by solving the following ℓ_1-regularized least squares problem.

$$\min_{\mathbf{c}} \frac{1}{2}||\mathbf{y}_t - \mathbf{B}_t\mathbf{c}||_2^2 + \lambda||\mathbf{c}||_1, \quad s.t. \quad \mathbf{c} \succcurlyeq 0 \tag{3}$$

where $\mathbf{B}_t = [\mathbf{T}_t, \mathbf{I}, -\mathbf{I}] \in \mathbb{R}^{d \times (n+2d)}, \mathbf{c} = [\mathbf{x}^T, \mathbf{e}_+^T, -\mathbf{e}_-^T]^T$, and $\mathbf{c} \succcurlyeq 0$ means each entry in $\mathbf{c}$ is nonnegative. Since the dimension of matrix $\mathbf{B}_t$ is generally very large, solving the Eq. (3) needs very high time cost. To handle this problem, motivated by handling the sparse outliers term in [28], we model the error vector $\mathbf{e}$ in formula (1) as the addition of two independent random vectors, i.e., the Gaussian noise vector $\mathbf{n}$ and the Laplacian noise vector $\mathbf{s}$. Thereby, we have

$$\mathbf{y}_t = \mathbf{T}_t\mathbf{x}_t + \mathbf{n}_t + \mathbf{s}_t, \tag{4}$$

where the Gaussian component models small dense noise and the Laplacian one aims to handle outliers.[1] Correspondingly, we can optimize the following objective function to obtain the sparse representation for a candidate target $\mathbf{y}_t$,

$$\min_{\mathbf{x}_t,\mathbf{s}_t} \frac{1}{2}||\mathbf{y}_t - \mathbf{T}_t\mathbf{x}_t - \mathbf{s}_t||_2^2 + \lambda||\mathbf{x}_t||_1 + \gamma||\mathbf{s}_t||_1, \tag{5}$$

where λ and γ are the penalty parameters of coefficient sparsity and the abnormal error respectively.

The aforementioned methods assume that the sparse representation of the tracked target is irrelevant between consecutive frames [14–16,13]. In practice, the tracked target varies very small between consecutive frames, so its inter-frame coefficients have strong relevance. In our model, this prior can be used to overcome the drifting problem due to some significant changes, e.g., illumination change. The inter-frame change of sparse coefficient representations of tracked objects can be regarded as a dynamic system. Thereby, we model the dependency of sparse representation coefficients of a tracked target at time t and $t-1$ as

$$\mathbf{x}_t = \mathbf{F}_t\mathbf{x}_{t-1} + \boldsymbol{\varepsilon} \tag{6}$$

where $\mathbf{F}_t$ denotes the state transition matrix from time $t-1$ to time $t, \boldsymbol{\varepsilon}$ denotes the prediction error or residue.

To solve the dynamic system defined by Eqs. (4) and (6), Inspired by Adam et al.'s work in [30], we introduce the innovation sparse term into the optimization problem defined by Eq. (5) and have

$$\min_{\mathbf{x}_t,\mathbf{s}_t} \frac{1}{2}||\mathbf{y}_t - \mathbf{T}_t\mathbf{x}_t - \mathbf{s}_t||_2^2 + \lambda||\mathbf{x}_t||_1 + \gamma||\mathbf{s}_t||_1 + \xi||\mathbf{x}_t - \mathbf{F}_t\mathbf{x}_{t-1}||_1, \tag{7}$$

where λ, γ and ξ are the controlling factors of sparsity.

2.2. Local dynamic sparse representation of tracked objects

The tracked object is also represented by a set of patches in our system. For each target candidate $\mathbf{y}_t \in \mathbb{R}^d$ at time t, it is divided into N non-overlapped patches, and each of them is represented as a vector $\mathbf{p}_j \in \mathbb{R}^{d/N}, j = 1, 2, \ldots, N$, which is normalized by ℓ_2 norm. In order to calculate the sparse representation of each patch $\mathbf{p}_j$, we need to construct the corresponding dictionary $\mathbf{D}_j$ respectively. For each dictionary $\mathbf{D}_j \in \mathbb{R}^{d/N \times n}, j = 1, 2, \ldots, N$, it is made up of the corresponding patch $\mathbf{d}_j^i \in \mathbb{R}^{d/N}$ of each target template $\mathbf{t}^i \in \mathbb{R}^d$, each of which is obtained by using non-overlapped sliding windows on each target template. Furthermore, we apply the formulation of Eq. (7) to each patch $\mathbf{p}_{t,j}$ at time t, and have

$$\min_{\mathbf{x}_{t,j},\mathbf{s}_{t,j}} \frac{1}{2}||\mathbf{p}_{t,j} - \mathbf{D}_{t,j}\mathbf{x}_{t,j} - \mathbf{s}_{t,j}||_2^2 + \lambda||\mathbf{x}_{t,j}||_1 + \gamma||\mathbf{s}_{t,j}||_1 + \xi||\mathbf{x}_{t,j} - \mathbf{F}_t\mathbf{x}_{t-1,j}||_1, \tag{8}$$

[1] The similar ideas of decomposing the noise into two components (e.g. decomposing the noise into sparse and non-sparse [13,29] for achieving robust motion estimation) have appeared in computer vision community.