Pattern Recognition 46 (2013) 1762–1771

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Robust visual tracking with discriminative sparse learning

Xiaoqiang Lu, Yuan Yuan *, Pingkun Yan

Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China

ARTICLE INFO

Available online 23 November 2012

Keywords:
Visual tracking
Sparse representation
Particle filter
Non-local self-similarity

ABSTRACT

Recently, sparse representation in the task of visual tracking has been obtained increasing attention and many algorithms are proposed based on it. In these algorithms for visual tracking, each candidate target is sparsely represented by a set of target templates. However, these algorithms fail to consider the structural information of the space of the target templates, i.e., target template set. In this paper, we propose an algorithm named *non-local self-similarity* (NLSS) *based sparse coding algorithm* (NLSSC) to learn the sparse representations, which considers the geometrical structure of the set of target candidates. By using *non-local self-similarity* (NLSS) as a smooth operator, the proposed method can turn the tracking into sparse representations problems, in which the information of the set of target candidates is exploited. Extensive experimental results on visual tracking have demonstrated the effectiveness of the proposed algorithm.

1. Introduction

Visual tracking is now a widely used technique in many applications such as security and surveillance, vehicle navigation, human computer interaction, and so on [1,2]. In order to design a robust visual tracking method under change conditions, the challenges are caused by the presence of scale, occlusion, pose variations, background clutter and illumination changes [3,4]. A detailed review can be found in Refs. [16,30,31]. Generally, the visual tracking problem can be classified in two different categories: generative and discriminative. The generative tracking methods adopt an appearance model to express the target observations. Some generative tracking methods include eigentracker [7], mean shift tracker [10], incremental tracker [15], and covariance tracker [14]. Ross et al. [15] present a tracking method that trains a low-dimensional subspace representation, and fits online changes in the target appearance. However, the appearance model needs to be often dynamically updated to fit the target appearance variations due to the rotation changes and scale variations.

Discriminative tracking methods address the tracking as a classification problem [25,26]. The strategy of tracking is to search the target location, which optimally extracts the target from the background [32,33]. Avidan et al. [5] form a feature vector by every pixel in the reference image and an adaptive ensemble of classifiers is trained to separate the object from the background. Collins and Liu [9] build a confidence map by searching the most discriminative RGB color combination in each frame. Yu et al. [18] propose a mixed combination of a generative model and a discriminative classifier to capture appearance variations. Babenko et al. [6] adopt online multiple instances learning to be robust to occlusions and other image corruptions. Yin and Collins [17] adopt global mode to search the object, and reinitialize the local tracker. Aran and Akarun [8] use an image fusion approach for discrimination and a generative approach for the target updates.

Recently, sparse representation has been introduced for tracking in Refs. [9,29] and later exploited in Ref. [12]. In Ref. [13], a target candidate is sparsely represented as a linear combination of target templates and trivial templates that only have one nonzero element in each of them. The sparse representation problem is solved through a L1 minimization problem to solve the model tracking problem [13,29]. However, based on the L1 sparse representation, these similar target candidates often have very different estimates due to the potential instability of sparse decompositions, which can result in bad tracking performance. In this case, it is necessary to exploit the geometry of the target candidate set to stabilize the sparse decompositions. Moreover, L1 methods assume that sparse representations of particles are independent. The structure relationships that ultimately constrain particle representations can be ignored in the L1 methods, which will result in bad tracking performance in cases of significant changes in appearance [37].

Recently, researches have shown that the geometrical structure of the data can improve the learning performance for discriminative training [34–36]. Felzenszwalb et al. [35] demonstrate an object detection system using mixtures of multiscale deformable part models. By using latent information and matching deformable

* Corresponding author.
E-mail address: yuany@opt.ac.cn (Y. Yuan).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.11.016

models to images, the proposed system is both efficient and accurate. Han et al. [36] present a statistical model called statistical local spatial relations (SLSR) to analyze the local region relations, which can resist some geometry transforms such as rotation, scale, viewpoint changes and part occlusion.

Motivated by recent progress of non-local self-similarity and sparse coding [22], in this paper, we propose a novel algorithm, called *non-local self-similarity regularized sparse coding* (NLSSSC), which explicitly considers the geometrical structure of the target and templates set. NLSSSC builds a *k*-nearest neighbor to encode the structure information in the target. The *non-local self-similarity* (NLSS) is regarded as a regularizer, which is incorporated into the sparse coding algorithm to preserve the structure information of the target. In order to obtain a proper sparse representation, the L2-norm regularization term of NLSS regularizer is replaced with L1 norm due to the success of compressed sensing [23]. Through preserving the structure information in the target candidate set, NLSSSC can have more discriminating power compared with the traditional algorithms and improve visual tracking performance.

The tracked target are often corrupted by noise or occluded in many visual tracking scenarios, which result unpredictable error. Although most trackers adopt different measure to reduce the error, they fail and cannot track the target with severe occlusions. The poor performances of most trackers are caused by the presence of occlusion, illumination changes, scale changes and varying view points. The performance depends on the degree of the similarity between the target and the templates. Hence, it is necessary to designing a robust visual tracking algorithm by consider the prior knowledge of the target and the templates. To further improve robustness, we propose an algorithm named non-local self-similarity (NLSS) based sparse coding algorithm (NLSSC) to learn the sparse representations, which considers the geometrical structure of the set of target candidates. The prior knowledge of the geometrical structure of data is successfully applied into the image process [34]. By using non-local self-similarity (NLSS) as a smooth operator, the proposed method demonstrates very promising performances. The main contributions of this paper are as follows:

1. A novel algorithm named non-local self-similarity (NLSS) based sparse representations is developed by considering the geometrical structure of the set of target candidates. To the best of our knowledge, few publications utilize such a framework in visual tracking scenarios. In addition, the obtained new penalty can generate more stable solutions than the L1 penalty.
2. The different video sequences involving scale, occlusion, pose variations, background clutter and illumination changes are tested to show that the proposed method demonstrates very excellent performances compared to other trackers.

The rest of this paper is organized as follows. The original L1-tracker is reviewed in Section 2. Section 3 introduces the NLSSSC algorithm for visual tracking, as well as the optimization scheme, including learning sparse representations. The experimental results on visual tracking are presented in Section 5. Finally, Section 6 concludes this paper.

2. L1-tracker with sparse representation

In this section, we will first review the original L1-Tracker framework [13], which effectively combines the particle filter and the sparse representation.

2.1. Particle filter

The L1-Tracker addresses the visual tracking as a sparse representation problem in the particle filter framework [11]. For frame at time t, z_t is denoted as the state variable describing the location and shape of a target, which can be modeled by the velocity components and the affine transformation parameters. In order to propagate the particles, the parameter of the velocity is introduced into the objection motion. The Gaussian distribution around the previous state z_{t-1} is exploited to approximate the transformation parameter of the state variable z_t. The tracking problem is represented as estimation of the state probability $(z_t|y_{1:t})$, where $y_{1:t}=(y_1,y_2,\cdots,y_t)$ represents the observations from previous t frame [11]. A two-stage Bayesian sequential estimation can be used for the tracking process. Applying Bayesian theorem, the filtering distribution can be recursively updated as

$$p(z_t|y_{1:t-1})=\int p(z_t|z_{t-1})p(z_{t-1}|y_{1:t-1})dz_{t-1}, \tag{1}$$

$$p(z_t|y_{1:t})\propto p(y_t|z_t)p(z_t|y_{1:t-1}), \tag{2}$$

where $p(z_t|z_{t-1})$ denotes the state transition probability, and $p(y_t|z_t)$ denotes the observation likelihood. The variable y_t is the region of interest cropped from the image, which can be normalized to be the same size as the target templates. It is practically intractable to directly calculate the above distribution. In the particle filter, the posterior $p(z_t|y_{1:t})$ is approximated by a finite set of M particle samples $\{z_t^i\}_{i=1}^M$ with importance given to weights. For each frame, the samples need to be updated and resampled.

In the L1-Tracker, the state variable z_t is modeled by six parameters of the affine transformation [13]. The state transition of z_t are formulated independently as a Gaussian distribution around the previous state variable z_{t-1}. The state transition model $p(z_t|z_{t-1})$ can generate the M candidate samples. The observation model $p(y_t|z_t)$ can be approximated by a Gaussian distribution, which indicates the approximation error between a target candidate and the target templates. The approximation error can be represented in the sparse representation described as following.

2.2. Sparse representation

To formulate the approximation error via observation likelihood $p(y_t|z_t)$, a patch is extracted from y_t corresponding to state z_t. The patch can be reshaped to a 1D vector x [13]. For a set of M particle samples $Z_t=\{z_t^1,\cdots,z_t^M\}$, the patch matrix $X=[x_1,\cdots,x_M]\in R^{d\times M}$ is the corresponding target candidate set. The sparse representation of X is formulated as a regularized L1 minimization function:

$$\min_S \|X-BS\|_F^2+\alpha\|S\|_1, \tag{3}$$

where $B=[T,I]$ consists of the target template set T and trivial template set I. The matrix whose columns in $T=[t_1,\cdots,t_n]\in R^{d\times n}$ are target template, and I is an identity matrix. $S=[A;E]$ consists of target coding coefficient matrix $A=[a_1,a_2,\cdots,a_n]$ and trivial coding coefficient matrix $E=[e_1,e_2,\cdots,e_n]$ respectively. Finally, the observation likelihood can be obtained from the reconstruction error of z_t^i as

$$p\left(y_t|z_t^i\right)=e^{-\varphi\|T\hat{a}_i-x_i\|_2^2}, \tag{4}$$

where the vector $\hat{a}_i$ can be gotten by solving the L1 minimization (3), φ is a constant controlling the similarity. For tracking at time instant t, the target candidate of the maximum observation