J. Vis. Commun. Image R. 36 (2016) 56–68

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

# Robust residual error consistent tracker with ranking mechanism ☆

Yueen Hou [a,*], Ping Ye [b]

[a] School of Computer, Jiaying University, Meizhou, China
[b] Auditing Department, Jiaying University, Meizhou, China

ARTICLE INFO

*Article history:*
Received 28 August 2015
Accepted 5 January 2016
Available online 13 January 2016

*Keywords:*
Visual tracking
Sparse representation
Particle filter
Residual error
Ranking
Alignment-pooling
Objective function
Target

ABSTRACT

In the paper, we propose a novel structural local sparse representation based residual error consistent ranking tracker. In our tracker, candidate targets are linearly combined by using the structural local sparse appearance model. To encourage temporal consistency, a residual error consistency term is designed to constraint the objective function of sparse representation. Based on the objective function, the similarity information is extracted from both coefficients and residual errors of sparse coding. For extracting similarity information from coefficients, the alignment-pooling algorithm is applied to obtain pooled features. For extracting similarity information from residual errors, we develop a residual error score. For different natures of residual error scores and pooled features, a ranking mechanism is proposed to fuse them. The dictionary updating scheme uses the ranking results of the predicted targets to determine which of them are collected for updating. Our tracker performs favorably against 6 state-of-the-art trackers on 18 challenging sequences.



## 1. Introduction

Visual object tracking technology plays a critical role in numerous application fields, such as visual servo, video surveillance, driver assistance, and human–computer interaction [1]. Though much progress has been made in the past decades, it is still a challenging task due to appearance and illumination changes, occlusions, background clutters and viewpoint variations.

A visual object tracker usually consists of two components: a motion model, which describes the states of targets overtime; an appearance model, which estimates likelihoods of samples belonging to targets.

The motion model is used to forecast target states in successive frames, such as the Kalman filter [2] and the particle filter [3,4]. In this paper, we use the particle filter to construct the proposed the motion model.

The paper focuses on designing a robust appearance model of the tracker. In order to develop a robust appearance model, two important aspects need to be taken into consideration.

The first one is concerned with what kind of features should be used to represent samples. Generally speaking, the sample representation scheme can be categorized into two kinds, which are adopted features and description models, respectively [5]. Adopted features include color [6], intensity [7,8], texture [9], Haar-like features [10], super-pixel based features [11], etc. while the description models contain holistic histogram [8], part-based histogram [12], etc. For the reasons of tracking efficiency, intensity features are used in our proposed tracking algorithm.

The second aspect is representation scheme, which can be classified as either discriminative or generative method.

The generative method aims to find out a candidate target, which is most similar to target templates. Generative methods are used in many tracking algorithms. Some of them perform well in complex scenarios. Black et al. [13] take an off-line subspace model for representing objects. However, the model is unable to adapt to appearance changes. Ross et al. [8] design an incremental learning tracking algorithm, which uses the principal component analysis method (PCA) for updating templates. The algorithm is robust to appearance changes. Recently, the sparse representation theory is widely applied to generative method based trackers. Mei et al. [7] introduce the sparse representation theory into the object tracking field. In the particle filter framework, candidate targets are reconstructed by a template dictionary with the sparsity constraint of coefficients. However, solving the $\ell_1$-minimization problem is time consuming. While for accelerating the solving process of $\ell_1$ minimization problem, several approximate solution approaches are proposed [14–16]. Mei et al. [14] develop a Bounded Particle Resampling (BPR)-$\ell_1$ tracker. The tracker reduces the number of

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
*E-mail address:* houyueen@jyu.edu.cn (Y. Hou).

http://dx.doi.org/10.1016/j.jvcir.2016.01.001
1047-3203/

particles which need to be sparsely reconstructed. In Ref. [15,16], researchers use the Accelerated Proximal Gradient approach (APG) to solve the sparse representation problem. In order to enhance the performance of sparse representation trackers in complex scenarios, some research works focus on modifying sparse representation objective functions [17–20]. Wang et al. [17,18] introduce sparse representation to the principal component analysis method for learning effective appearance models, and develop an iterative algorithm to solve the objective function. Lan et al. [19] develop a visual tracking algorithm based on feature-level fusion using joint sparse representation. The tracker is able to detect the unreliable features, therefore the performance is enhanced. Zhang et al. [20] develop a consistent low-rank sparse learning tracker (CLRST), which uses low-rank representation term, sparse representation term, temporal consistency term and reconstruction error term to construct the objective function. As a reverse thought to conventional sparse representation, Zhuang et al. [16] construct a dictionary that contains a set of candidates to represent target templates. A new regularization term, which indicates the relationship between two candidate features, is added to the objective function. To exploit both spatial and local information of the targets, Jia et al. [21] propose a structural local sparse appearance model based tracking algorithm. This tracking algorithm exploits a dictionary, which contains patches from templates, to reconstruct patches from candidate targets. Differing from other sparse representation tracking schemes, the tracking algorithm uses coefficients of sparse coding to describe the likelihood of candidate states belonging to target states.

The discriminative method treats tracking as a classification task. A discriminative model based tracker trains a binary classifier to discriminate the target and the background. The classifier is trained by both positive and negative samples which are updated by observations obtained online. In recent years, many research works use discriminative methods for object tracking, and obtain desirable performances. In [10], a multiple instance learning tracker based on the boosting framework is proposed. A strong classifier composed by many weak classifiers is employed to detect targets. Furthermore, the weak classifiers with more discriminative power are adaptively chosen to compose the strong classifier. Zhang et al. [22] develop a real-time compressive tracker, which is trained by both positive and negative samples. As for the efficiency of the tracker, a very sparse measurement matrix is adopted to extract the features for the appearance model. Kalal et al. [23] propose a TLD tracker based on the tracking–learning–detection framework. The TLD tracker combines tracking and detection together for object tracking, hence it is suitable for long time tracking tasks. Many sparse representation trackers apply the discriminative method to improve the robustness in complex scenarios. Wang et al. [24] exploit real-world images learned off-line for building a dictionary. A linear classifier is learned online by using sparse coding coefficients of tracking results to detect the target and the background. In [25], a linear classifier based on local sparse representation is learned to discriminate the target and the background. A two-stage algorithm using both the ground truth information and observations is applied to alleviate the drift problem. Bai et al. [26] embed classification results into the sparse representation objective function. Zhong et al. [27] combine a sparsity-based discriminative model (SDM) and a sparsity-based generative model (SGM) together for object tracking.

Motivated by the demonstrated success of [21], we use the structural local sparse appearance model for object tracking. For considering the temporal consistency, a residual error consistency term is used to constrain the sparse representation objective function. Both coefficients and residual errors of sparse coding are used to describe the likelihood of candidates belonging to targets. The likelihood information from coefficients and residual errors is fused by a novel ranking mechanism. What's more, the ranking results are used to select the correct tracking results which are collected for updating the dictionary.

The contributions of the paper are as follows:

(1) Most of the sparse representation based tracking algorithms use residual errors to measure the similarity of the candidates and the targets set. However, little concerns with the relationship of residual errors between two consecutive frames. Different from traditional sparse representation tracking algorithms, we use a residual error consistency term to constraint the sparse representation objective function, which effectively enforces temporal consistency for visual tracking.
(2) Ref. [21] uses only coefficients of sparse coding to determinate target states, but ignores the similarity information that hides in the residual errors of sparse coding. For the purpose of enhancing the detective power of the tracker, we develop a kind of residual error score that extracts the likelihood information from residual errors of sparse coding.
(3) Given different natures of coefficients and residual errors of sparse coding, a new ranking mechanism is designed to fuse the likelihood information from them. Moreover, the ranking results are applied to select the 'right' tracking results for dictionary updating.

The main components of the proposed tracking algorithm are shown as Fig. 1.

## 2. Related works

Our proposed tracking algorithm is based on the theory of sparse representation, which is widely applied in numerous vision applications[28–30,7,21]. Mei et al. [7], who introduce sparse representation into object tracking areas, propose a $\ell_1$ tracker. The $\ell_1$ tracker considers the object tracking task as a linear combination work. In the particle filter framework, candidates are reconstructed by atoms of a dictionary with $\ell_1$ minimization and non-negativity constraints. The dictionary contains target templates as well as trivial templates which enable the tracker to handle partial occlusions. The candidate, which has the smallest value of residual error, is considered to be the target. The $\ell_1$ tracker makes a good performance. However, solving the $\ell_1$ minimization function is time consuming. Yang et al. [29] discuss the relationship of linear combination functions with $\ell_1$ and $\ell_2$ minimization constraints. They obtain a conclusion that the linear combination functions with the $\ell_1$ minimization constraint and $\ell_2$ minimization constraint obtain similar performance in face recognition tasks. Nevertheless, solving the $\ell_2$ minimization constraint function is much faster than solving the $\ell_1$ minimization constraint function. In order to speed up the tracking algorithm, we apply $\ell_2$ minimization to constrain the linear combination function.

Another drawback of the $\ell_1$ tracker is that the tracker only considers the holistic representation and does not exploit local information of targets. To overcome the drawback, Jia et al. [21] propose a structural local sparse appearance model, which exploits both partial and spatial information of targets. In the appearance model, over-lapped local image patches are sampled. By using an alignment-pooling method, candidates' similarity information is extracted from coefficients of spares coding. Our proposed tracking algorithm is based on the structural local sparse appearance model, which bears some similarities with [21]. Different from [21], we use not only coefficients but also residual errors to measure the similarity, and propose a novel ranking mechanism to fuse the likelihood information from them. For further robustness, a residual