Pattern Recognition 48 (2015) 2885–2905

Contents lists available at ScienceDirect

Pattern Recognition

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

Real-time and robust object tracking in video via low-rank coherency analysis in feature space

Chenglizhao Chen [a], Shuai Li [a,*], Hong Qin [b], Aimin Hao [a]

[a] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
[b] Stony Brook University SUNY, United States

ARTICLE INFO

Article history:
Received 30 July 2014
Received in revised form
7 November 2014
Accepted 27 January 2015
Available online 13 February 2015

Keywords:
Localized compressive sensing representation
Fast lowrank approximation
Lowrank coherency tracking
Visual tracking

ABSTRACT

Object tracking in video is vital for security surveillance, pattern and motion recognition, traffic control, augmented reality, human-computer interaction, etc. Despite the rapid growth of various techniques in recent years, certain technical challenges still exist in terms of efficiency, accuracy, and robustness. To ameliorate, this paper suggests a novel video object tracking approach by first collecting both local and global information from consecutive video observations (i.e., frames) and then exploring the low-rank coherency in the accompanying feature space of targeting objects, which enables real-time and robust object tracking in video while combating certain technical difficulties due to occlusion, deformation, transient illumination, rapid movement, and scale change. Our central idea is to integrate local space-distinctive candidate features and global time-continuous target coherency into a smart low-rank analysis model. For local candidate representation, we propose a simple yet efficient patch-level feature descriptor based on compressive sensing, which is directly derived from the frame color distribution available from video frames. Building upon this powerful local representation, we further organize all the candidates in the frame cache and the yet-to-be-processed new frame to form a space-time feature set, we then employ the low-rank decomposition to enable global coherency voting. Since the low-rank coherency implies the intrinsic co-occurring parts of different target observations, robust tracking can be achieved by employing this principle as the matching criterion even for objects with drastically varying appearance. Furthermore, we progressively incorporate the prior-frames' tracking results into the low-rank approximation in the current frame, which can greatly reduce the most time-consuming computation and guarantee real-time performance. We conduct extensive experiments on several well-known yet challenging benchmarks, and make comprehensive and quantitative evaluations with state-of-the-art methods. All the results demonstrate the superiority of our method in terms of accuracy, efficiency, robustness, and versatility.

1. Introduction and motivation

Visual tracking is still one of the most active research areas in computer vision and pattern recognition, which is extremely valuable in many applications, including surveillance [1], traffic control [2], motion recognition [3], etc. Although visual tracking research has gained great momentum and has achieved significant successes in recent years, it still remains challenging when the goal is to robustly track the target object with high-varying inter-frame appearance and occasional/frequent occlusion in real-time [4]. Generally speaking, current research methods of visual tracking can be roughly categorized into two groups: discriminative tracking [5–9] and generative tracking [10–14].

Discriminative tracking customarily employs binary classification to separate the target object from its background, wherein numerous training samples from the tracking results of previous frames are indispensable to the individual classifiers of current observation [6]. However, in order to achieve accurate and occlusion-invariant tracking, most of the discriminative methods [15,16] avoid updating their classifiers that are varying far away from the initial setting [17], as a result, drift occurs unavoidably when the target object undergoes heavy scale or rapid appearance change.

Unlike discriminative tracking methods, generative tracking methods usually resort to certain appearance models to depict the object-specific observation, and take the candidate having the best compatibility with the appearance model as the tracked object [14,18]. In theory, the generative tracking methods can accommodate any complex appearance variation at the expense of extra computational burden by continuing to enhance the description capacity for the partial appearance model. Hence, for appearance model with limited capacity, the target's representation together with its

[*] Corresponding author.
E-mail address: lishuai@buaa.edu.cn (S. Li).

http://dx.doi.org/10.1016/j.patcog.2015.01.025
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

similarity-matching criteria are the vital issues of generative tracking. Inspired by this rationale, generative methods commonly adopt mid-level patch-based solution (e.g., PCA or histogram) to locally represent the target object, and treat the entire reconstruction error [11] or partial observation [18] of the appearance model as matching criteria, which in some sense can make effective tradeoff between versatility enhancement and drift suppression.

In sharp contrast, we explore the compressive sensing and low-rank analysis theory to facilitate the video object tracking in feature space. It should be noted that [8] suggests to represent the tracking object with Haar-like feature formulation based on compressive sensing. However, the Haar-like feature formulation imposes too much emphasis on the global characteristics distribution to handle the occlusion, which further limits its application scope for generative tracking. To respect object's local characteristics, we propose to employ mid-level patch-based integral histogram to represent the candidate target, which can be seen as a relaxed local version of [8] at the expense of possibly compromising discriminative power. Moreover, based on this novel representation, we have found that the principal characteristics distribution of the target object will not change much within the consecutive frames, even though the appearance may have changed drastically. Therefore, it provides enough rationale for us to leverage such coherency for robust object tracking by resorting to low-rank analysis, and moreover, the low-rank coherency extracted from the tracked instance of our appearance model in previous frames will serve as the matching criteria of current-frame candidate targets. Meanwhile, since the principal low-rank information are redundant, it can be synchronously used to govern the dynamic update of the appearance model with limited capacity (e.g., 100 in our experiments). Benefiting from the elegant integration of space-distinctive candidate features and time-continuous coherency, our method can accommodate high-varying appearance and occasional/frequent occlusion. In particular, the salient contributions of our work documented in this paper can be summarized as follows:

- We propose a versatile, real-time, and robust video object tracking method, which can neatly accommodate the object's varying appearances caused by local occlusion, large deformation, transient illumination, rapid movement, drastic scale change, etc.
- We define an efficient yet discriminative patch-based appearance model based on compressive sensing, which can compactly represent the intrinsic characteristics distribution information of the local object parts in a very low dimensional feature space.
- We propose a novel low-rank decomposition based cross-frame coherency analysis model to robustly capture video object that may undergo large appearance variation, and such method can also be used to govern the dynamic update of our appearance model.
- We formulate a series of sparsity-measuring based criteria to accelerate tracking performance, assist appearance model update, and handle occlusion effectively.

2. Related work

Based upon the feature representation and the matching criteria, we further classify the large variety of discriminative and generative tracking methods into global-representation based tracking methods, local-representation based tracking methods, and hybrid tracking methods. Now we briefly review them as follows.

Global-representation based tracking methods: Most of the global-representation based tracking methods usually resort to certain types of color or intensity based histogram for feature representation. Since the histogram implies discriminative color distribution to distinguish the target object from its surroundings, the tracking problem may be converted to a binary classification problem by globally making a decision for boundary [19–21]. Meanwhile, global tracking criteria derived from all previous frames are oftentimes used to facilitate current-frame tracking [22,23]. Thus, such methods give rise to high tracking accuracy and low computational cost [5,7]. Despite some special advantages of the global-representation based tracking methods, several common problems remain to be solved. First, because global representation is sensitive to occlusion, such methods tend to mistakenly consider occlusion as reasonable appearance variation when updating their basic classifiers, which may easily result in tracking drift [6,16]. Second, because the global feature representation is discontinuous in nature, learning based global tracking solutions (or matching criteria) are usually hard to accommodate fast appearance change [24].

Local-representation based tracking methods: Local-representation based tracking methods commonly decompose the target object into many discriminative patches/regions, and employ the patch-to-patch or region-to-region matching strategy to conduct object tracking. For example, Adam et al. [13] represented the target with multiple regular image fragments, which locally describe its different components. Wang et al. [10] represented the target object with the irregular super-pixels based SLIC method [25], and employed clustering based matching criteria together with the special voting or integrating strategy [26] to improve the robustness of object tracking. Although local-representation based tracking methods can well solve the occlusion problem [27], the absence of global spatial-distribution information may weaken the distinguishability of object representation. Therefore, it may at times lead to tracking drift, and the patch-wise matching operation may further deteriorate the tracking result. To combat such limitations, Erdem et al. [12] proposed to represent the target object with higher-level object regions, wherein they adopted the grid-based region representation and searched the target object in a region-to-region manner. Most recently, He et al. [28] used a locally sensitive histogram based region representation to combat the illumination variation, and obtained rather amazing results. Meanwhile, Yao et al. [9] leveraged latent variables based online learning to facilitate the region-weighted representation of target objects, which achieves more robust tracking results. Although local-representation based tracking methods demonstrate their special advantages in handling occlusion and fast partial appearance, however, these methods are still hard to accommodate drastic appearance variation, and the inevitable computational complexity involved in such methods heavily limits their real-time tracking capability (it may be noted that in such cases, FPS $\ll$ 15).

Sparse-representation based tracking methods: Based on the sparse representation (SR) theory, Liu et al. [11] employed the image patches based local representation and global reconstruction error based matching criteria to locate the target object (i.e., local representation with global tracking), wherein the basis functions used for reconstruction are learned from previous-frames' tracking results. Similarly, Zhong et al. [29] introduced a sparsity-based generative model by alternatively formulating sparsity based local feature, and further integrated the global spatial information of each individual patch into an occlusion handing scheme. Xu at el. [14] proposed a sparse coding pool based hybrid representation and taken into account multiple templates during target matching, which achieves improved tracking results. And then, they [18] further proposed an occlusion-handling method by coupling additional noise templates (which is local) with a batch of PCA based individual phototypes (which is global) [15]. Zhang et al. [30] also obtained comparable tracking performance by introducing the low-rank constraint into the formulation of SR basis functions. Recently, Zhang et al. [31] concentrated on localized tracking solution, wherein their SR basis functions are learned within multiple observation constraints. Meanwhile, following the diametrically opposed rationality (global representation with local tracking) of sparse representation, Zhang et al. proposed compressive sensing based global tracking methods [8] via globally representing the target object and