J. Vis. Commun. Image R. 30 (2015) 266–276

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Affine hull based target representation for visual tracking ☆

Jun Wang [a,b], Hanzi Wang [a,c,*], Wan-Lei Zhao [a,c]

[a] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Technology, Xiamen University, Xiamen 361005, China
[b] Cognitive Science Department, Xiamen University, Xiamen 361005, China
[c] Computer Science Department, Xiamen University, Xiamen 361005, China

ARTICLE INFO

Article history:
Received 22 January 2015
Accepted 28 April 2015
Available online 12 May 2015

Keywords:
Visual tracking
Particle filter
Affine hull
Histograms of sparse codes
Appearance model
Target representation
Generative tracking
Template set

ABSTRACT

Handling appearance variations is a challenging issue in visual tracking. Existing appearance models are usually built upon a linear combination of templates. With such kind of representation, accurate visual tracking is not desirable when heavy appearance variations are in presence. Under the framework of particle filtering, we propose a novel target representation for tracking. Namely, the target candidates are represented by affine combinations of a template set, which leads to better capability in describing unseen target appearances. Additionally, in order to adapt this representation to dynamic contexts across a video sequence, a novel template update scheme is presented. Different from conventional approaches, the scheme considers both the importance of one template to a target candidate in the current frame and the recentness of the template that is kept in the template set. Comprehensive experiments show that the proposed algorithm achieves superior performances in comparison with state-of-the-art works.

1. Introduction

Video tracking is to locate a tracked target across the video sequence. It can be applied in a variety of tasks such as human–computer interaction, security and surveillance, video communication and compression, augmented reality, and video editing. Video tracking is usually a computationally expensive process due to a large amount of information encoded in videos. Although much progress has been made in the past decades [1,2], it remains a hard issue due to number of challenges along with the tracked target such as illumination variations, occlusions, out-of-plane rotations, non-rigid object deformation and background clutters.

As a common practice, a target is manually selected in the first frame. Thereby, a tracking algorithm is required to associate the tracked target in the rest frames across the whole video sequence. There are two major components in a typical tracking system: appearance model and motion model. The appearance model is important because it represents the target and is used to evaluate the observation likelihood of each target candidate in the current frame. A good appearance model should be robust to illumination variations, occlusions and other significant appearance variations. In the meantime, computational costs for the target representation should be kept as low as possible such that the whole tracking process can run in real-time.

In visual tracking, a target is usually represented by a set of templates, which is sometimes called as a dictionary. Each target candidate is linearly described as a vector spanned by a pre-defined template set. The observation likelihood of a target candidate belonging to the target is evaluated based on the reconstruction residual (i.e., the distance between the target candidate and the template set). Usually, the target candidate which gives the minimum reconstruction residual is selected as the tracking result in the current frame.

In general, a template set is composed of the most representative instances of the tracking results from previous frames across a video sequence. In [3], a target candidate is linearly represented by a set of basis samples, which are directly sampled from a video sequence. In [4], the representative samples are undergone the principle component analysis (PCA) before they are used as representation templates. Such that a target candidate is represented by a set of PCA basis vectors, which is in low dimension and robust to illumination and pose variations. However, this representation turns out to be vulnerable to background clutters. In [5], a target candidate is represented by both target templates and trivial templates. Different from the L1 algorithm [5], Zhang et al. [6] use both target and background templates to collaboratively represent a target candidate.

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

* Corresponding author at: Computer Science Department, Xiamen University, Xiamen 361005, China.
E-mail addresses: wangjuncv@gmail.com (J. Wang), hanzi.wang@xmu.edu.cn (H. Wang), wlzhao@xmu.edu.cn (W.-L. Zhao).

http://dx.doi.org/10.1016/j.jvcir.2015.04.014
1047-3203/

For all the above-mentioned approaches, the representation for a target candidate is a linear combination of templates. Despite representative instances for a target are included in a template set, this linear combination may fail to cover some of target appearances. It is particularly true in the cases that the target undergoes heavy occlusions and other significant appearance variations. To address this problem, we propose to represent a target candidate with an affine combination of target templates, which is able to cover unseen target appearances.

During the process of tracking, a target object could be embedded in dynamic backgrounds. The shape and appearances of the tracked target also change dynamically. In order to achieve better approximation to the target, the template set should be adaptive to reflect these variations. Traditionally, the template which contributes the lowest weight to the target is swapped out and replaced by the tracking result in the current frame. In our work, a novel template update scheme is proposed, which considers not only the weight of one template that contributes to the target, but also the recentness of the template. Moreover, we propose a novel likelihood evaluation function based on both the reconstruction residual and the template coefficient vector, which makes tracking more stable. Finally, we propose an Affine Hull based Tracking (referred to as AHT) algorithm. Comprehensive experiments on eight challenging video sequences against several state-of-the-art tracking algorithms demonstrate that the proposed AHT algorithm achieves superior tracking performance.

The remainder of this paper is organized as follows. Section 2 summarizes the related work. Section 3 briefly reviews the particle filter framework for visual tracking. Section 4 presents the proposed tracking algorithm, which includes the target representation, the likelihood evaluation and the template update scheme. Section 5 evaluates the performance of the proposed AHT algorithm in comparison to several state-of-the-art tracking algorithms on some challenging video sequences. Section 6 concludes the paper.

2. Related work

Visual tracking is an important research topic in computer vision. Based on the types of appearance models adopted, tracking algorithms can be broadly categorized into two groups, namely generative tracking and discriminative tracking. Generative tracking algorithms [4,5,7–12] usually learn an appearance model to represent a target, and use the learnt appearance model to search for the region which is the most similar to the target model in each image frame. Unlike generative tracking algorithms, discriminative tracking algorithms [13–24] formulate visual tracking as a binary classification problem, in which a classifier is learnt to separate a target from its surrounding backgrounds. Here, we briefly review some representative tracking algorithms and techniques related to our work.

Recently, several tracking algorithms based on sparse representation techniques [25] have been proposed [5–7,26–30]. These tracking algorithms represent each target candidate by a sparse linear combination of atoms (or templates) in a dictionary. In [5], a target candidate is linearly represented as a linear combination of both target and trivial templates, where trivial templates are used to describe outliers or occlusions. The L1 tracking algorithm [5] is robust to various appearance variations (e.g., partial occlusions and illumination variations), however, it is computationally expensive due to the requirement of solving the ℓ_1-regularized least-square problem. Bao et al. [29] extend it via an accelerated proximal gradient algorithm, which achieves better tracking performance and faster speed.

In order to further improve tracking stability and accuracy, Zhang et al. [6] utilize the background information around the target to build background templates, and use both target and background templates to collaboratively represent each target candidate. Wang et al. [30] use target templates to represent a target candidate, where the target templates are updated by a dictionary trained online. In [26], visual tracking is formulated as a multi-task sparse learning (MTT) problem, where the representations of target candidates are jointly learnt by using the accelerated proximal gradient algorithm. The MTT tracking algorithm [26] is robust to illumination and pose variations. However, these tracking algorithms [5,26,29] normalize each target candidate to a low resolution image patches (e.g., 12×15 in [5]), which are incapable of capturing sufficient visual information to represent the target. Although sparse representation based tracking algorithms have shown interesting results, expensive computation requirement due to solving the ℓ_1 minimization problem limits the performance of these tracking algorithms in real time.

In [31], a target is represented by multiple non-overlapping image patches. The Frag tracking algorithm [31] can effectively alleviate drift problem, however, because the target template is fixed, it is not robust to significant appearance variations. Kwon and Lee [32] propose a compound (VTD) tracker by combining multiple basic trackers, which are designed by associating a basic motion model and a basic appearance model. VTD is robust to drastic motion variations. In [33], Kwon and Lee sample a set of trackers to deal with severe appearance and motion variations. Oron et al. [34] propose a tracker by estimating the amount of local (dis) order in a target, which turns out to be robust to non-rigid object deformation.

Wang et al. [9] propose a Least Soft-thresold Squares (LSS) distance based tracking algorithm. In [9], a target candidate is modeled by a dictionary, which consists of a set of PCA basis vectors. The reconstruction residual between a target candidate and the dictionary is computed by the LSS distance. The LSS tracking algorithm is robust to partial occlusions and background clutters. In [8], a target is represented via a PCA subspace, and the continuous outliers are depicted by using a probability continuous outlier model.

For generative tracking algorithms, a robust appearance model plays an important role. Usually, a target candidate is linearly approximated by a set of templates or basis vectors. However, this turns out to be insufficient to cover possible appearances of a target due to significant appearance variations during visual tracking.

Inspired by the convex hull representation that is used in image set classification [35] and face recognition [36,37], a target candidate is approximated by an affine combination on a set of templates in our work. The proposed affine hull based target representation has several advantages. First, target appearance models based affine combinations are more discriminative than other linear combinations based appearance models, since the affine hull can account for unseen target appearances. Such a target representation is particularly appealing in the cases that heavy occlusions and other significant appearance variations are in presence. Moreover, the proposed affine hull based target representation can be simplified as a projection operation, which is approximated efficiently by using the ridge regression technique [38].

3. Particle filter for visual tracking

The particle filter framework [39], also known as a sequential Monte Carlo method for importance sampling [40], is popular for visual tracking due to its simplicity and effectiveness. Given the