Neurocomputing 153 (2015) 77–88

Contents lists available at ScienceDirect

Neurocomputing

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

# Robust visual tracking by metric learning with weighted histogram representations

Jun Wang [a,b,c], Hanzi Wang [a,*], Yan Yan [a]

[a] School of Information Science and Technology, Xiamen University, Xiamen 361005, China
[b] Fujian Key Laboratory of the Brain-like Intelligent Systems (Xiamen University), Xiamen 361005, China
[c] Cognitive Science Department, Xiamen University, Xiamen 361005, China

ARTICLE INFO

*Article history:*
Received 22 May 2014
Received in revised form
2 September 2014
Accepted 22 November 2014
Communicated by Xiaoqin Zhang
Available online 3 December 2014

*Keywords:*
Visual tracking
Distance metric learning
Weighted histogram representations
Particle filters

ABSTRACT

Measuring the similarity between the target template and a target candidate is a critical issue in visual tracking. An appropriate similarity metric can improve the accuracy and robustness of visual tracking. This paper proposes a robust visual tracking algorithm that incorporates online distance metric learning into visual tracking based on a particle filter framework. The appearance variations of an object are effectively learned via an online metric learning mechanism. In addition, we use spatially weighted feature representations using both color and spatial information of objects, which can further improve the tracking performance. The proposed algorithm is compared with several state-of-the-art tracking algorithms, and experimental results on challenging video sequences demonstrate the effectiveness and robustness of the proposed tracking algorithm.



## 1. Introduction

Visual tracking has been well studied in recent decades. The goal of visual tracking is to continually predict the locations of target objects in video sequences. A large number of visual tracking algorithms have been proposed and applied in vehicle navigation, monitor surveillance, human–computer interaction, and so on [1,2]. Although much progress has been made, robust visual tracking remains a challenging problem due to occlusion, background clutters, fast motion, illumination changes, motion blur and rotation (see Fig. 1). Generally speaking, visual tracking algorithms can be categorized as either generative [3–9] or discriminative [10–20].

Generative tracking algorithms search for an image region in each frame that is the most similar to the target template with a maximal similarity score or a minimal reconstruction error. In generative tracking algorithms, an appearance model is used to represent the target object, and then dynamically updated during tracking. Ross et al. [3] used a low-dimensional subspace model to represent an object, which is robust to illumination and pose changes. Wang et al. [4] proposed the least soft-threshold squares (LSS) algorithm to deal with appearance variations. Li et al. [6] used a set of cosine basis functions to build a compact 3D-DCT object representation, where an incremental 3D-DCT algorithm was proposed to achieve robust tracking in challenging environments. In [7], a motion model was decomposed into multiple basic motion models, which aimed to handle motion changes.

Recently, sparse representation based tracking algorithms were also developed [8,9,21,22]. In [8], visual tracking was formulated as a sparsity-based reconstruction problem in a particle filter framework, where a target candidate with the smallest reconstruction error is considered as the tracking result. Zhang et al. [9] generalized the $\ell_1$ minimization tracking algorithm [8] as a multi-task tracking (MTT) algorithm, where visual tracking was formulated as a multi-task sparse learning problem. Jia et al. [21] presented a structural local sparse appearance model for object representation, which exploited both partial and spatial information of the target via an alignment-pooling method. Zhong et al. [22] adopted local representations to build a sparsity-based generative model, which can effectively handle heavy occlusion.

In contrast, discriminative tracking algorithms treat visual tracking as a binary classification problem. These kinds of algorithms consider the differences between an object and its surrounding background. Grabner et al. [10] presented an online boosting algorithm (OAB) to select discriminative features for visual tracking. OAB was extended to a semi-supervised boosting algorithm [11], which effectively alleviated the drifting problem. Avidan [12] proposed an ensemble tracking framework wherein multiple weak classifiers were combined into a strong classifier by using an AdaBoost algorithm. The randomized ensemble tracking (RET) algorithm was proposed by Bai et al. [13], where a set of weak classifiers was combined by using a weight vector that is treated as a

* Corresponding author. Tel.: +86 592 2580063.
*E-mail addresses:* wangjuncv@gmail.com (J. Wang), wang.hanzi@gmail.com (H. Wang), yanyan@xmu.edu.cn (Y. Yan).

http://dx.doi.org/10.1016/j.neucom.2014.11.050
0925-2312/

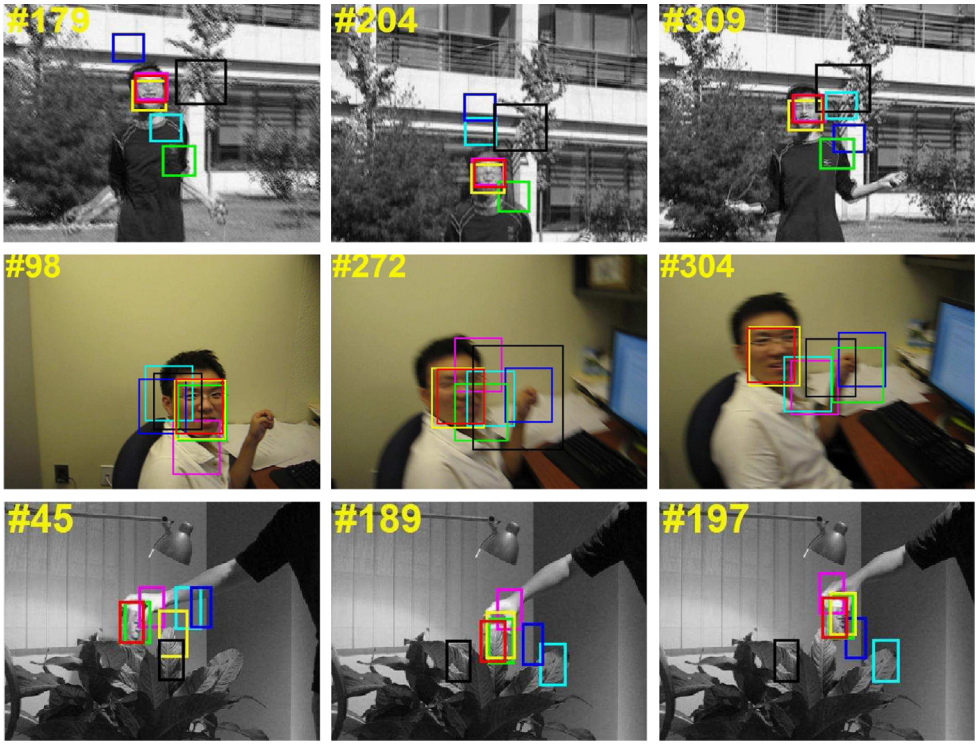


**Fig. 1.** Tracking in several challenging situations including motion blur (the first row: *Jumping*), fast motion (the middle row: *Face*), and occlusion (the last row: *Coke can*). The tracking results of CT [15], Frag [5], L1 [8], MTT [9], TLD [19], VTD [7] and the proposed tracking algorithm are represented by magenta, cyan, blue, green, yellow, black and red rectangles, respectively. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

distribution of confidence among the weak classifiers. Collins et al. [14] adaptively selected the best discriminative feature from multiple features via an online ranking mechanism, which can deal with partial occlusions, background clutters and illumination changes.

Zhang et al. [15] proposed a real-time compressive tracking (CT) algorithm by adopting random projection to project a datum in high-dimensional space to a low-dimensional vector. In order to alleviate the drifting problem, CT used a spatial sampling scheme to obtain several positive samples to train a classifier, instead of using one positive sample as in [10,14]. Babenko et al. [16] used positive and negative bags to learn classifiers, where multiple instance learning was introduced into visual tracking. Hare et al. [17] proposed a tracking-by-detection algorithm based on a structured output SVM learning technique. Yang et al. [18] presented a superpixel based appearance model for visual tracking to handle pose variations.

The measure of similarity between the target template and a target candidate is an important issue, which might have influence on the accuracy and robustness of visual tracking algorithms. In this paper, we present a simple and robust tracking algorithm that is able to handle fast motion, background clutters, motion blur, occlusion, etc. The proposed algorithm employs an online distance metric learning technique [23], instead of a predefined metric to measure the similarity between the target template and a target candidate. In order to further improve the robustness for visual tracking, we use spatially weighted histogram-based feature representation. The intensity values are used in the spatially weighted histogram. The computational complexity of our algorithm is low because the dimensionality of the proposed feature representation is low. Therefore, the computational cost of online distance metric learning and object tracking is reduced. By using the online distance metric learning algorithm [23], we compute the measure of similarity between the target template and a target candidate and track an object in a particle filter framework.

The remainder of this paper is organized as follows. Section 2 summarizes the related work. Section 3 presents the feature representation and proposes an online distance metric learning based tracking (referred to as OMLT) algorithm. Section 4 presents experimental results, and evaluates the performance of the proposed tracking algorithm and several competing algorithms. Section 5 concludes the paper.

## 2. Related work

The performance of most tracking algorithms greatly depends on a distance metric or the measure of similarity between the target template and a target candidate. Most existing tracking algorithms use a pre-determined metric, e.g., the EMD metric [5], the histogram intersection [22], or the Bhattacharyya coefficient metric [24,25]. A predefined distance metric cannot adapt to appearance variations, and may lead to tracking failure.

Recently, much research in the fields of image retrieval and pattern recognition has demonstrated that an appropriate distance metric can significantly improve the retrieval or classification performance. Distance metric learning algorithms have attracted much interest in visual tracking [26,28,30,32]. The goal of metric learning is to learn a metric for measuring the similarity between the target template and a target candidate.

Jiang et al. [26] proposed a discriminative tracking algorithm using the neighborhood component analysis (NCA) metric learning algorithm [27]. NCA learned a distance metric and reduces the dimensionality of the feature space. However, NCA could suffer from spurious local maxima. Wang et al. [28] presented an object tracking algorithm based on the maximally collapsing metric learning (MCML) [29]. MCML learned a distance metric by collapsing samples with the same class label together and pushing away samples with different class labels. MCML assumed that samples with the same class labels have a unimodal class distribution. Tsagkatakis and Savakis [30] used the information-theoretic metric learning (ITML) algorithm [31] for visual tracking, where visual tracking was considered as the nearest neighbor classification problem. ITML required a large number of