Information Sciences 364–365 (2016) 33–50

Contents lists available at ScienceDirect

Information Sciences

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

Online similarity learning for visual tracking

Sihua Yi, Nan Jiang, Bin Feng, Xinggang Wang, Wenyu Liu*

School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, PR China

ARTICLE INFO

Article history:
Received 21 August 2015
Revised 13 April 2016
Accepted 15 May 2016
Available online 18 May 2016

Keywords:
Metric learning
Classification
Visual tracking

ABSTRACT

Incorporating metric learning in visual tracking applications has been demonstrated to be able to improve tracking performance. However, the optimal metric is mainly derived based on annotated feature vectors by studying their magnitude and intersection angle. In complex scenarios, the magnitude of feature samples may change drastically, confining the matching performance of the distance metric. Moreover, most distance learning methods are optimized in a time-consuming iterative way, which limits their applications in real-time visual tracking. To address these problems, this paper proposes a novel metric called sphere similarity metric, which normalizes the magnitude of feature vectors and measures the distance between pairs of vectors by their intersection angle. Such metric is robust even when the magnitude of feature vectors changes drastically in complex scenarios. We formulate the proposed metric by a convex matrix function, which does not require adapting samples' magnitude iteratively, and can be optimized in a closed form with low computational complexity. Additionally, the proposed similarity metric can be learned in an online manner, which accelerates the learning process and improves the tracking accuracy in visual tracking applications. Experimental results on synthetic data and benchmark video sequences show that the proposed metric learning method achieves better classification accuracy, and its application in visual tracking outperforms the state-of-the-art tracking methods.

1. Introduction

Matching the target over consecutive frames robustly is critical for visual tracking. Because the target of interest may suffer the challenging appearance changes caused by occlusion, motion blur and illumination variation in complex environments. Under the improper distance metric, the estimated region may not be the true target of interest, but one sharing the similar appearance with the target in the background. Instead of using a pre-specified or fixed distance metric [5,7,22], distance adjustment seems to be a feasible way to improve the matching performance, most commonly used being Mahalanobis distance metric learning. Recent attempts [26–28,32,33,50,54] have integrated the Mahalanobis distance metric learning into visual tracking. These methods train and adjust the distance metric adaptively to separate the target from its distracters in the background, and report quite promising results. However, many studies have indicated that it does not always guarantee a better performance in complex scenarios to straightforwardly integrate conventional distance metric learning algorithms into visual tracking [4,28,41,60,68]. Moreover, most existing metric learning methods are computationally intensive, which limits their application in real-time tracking.

* Corresponding author. Tel.: +86 2787543236.
E-mail address: liuwy@hust.edu.cn (W. Liu).

http://dx.doi.org/10.1016/j.ins.2016.05.019
0020-0255/

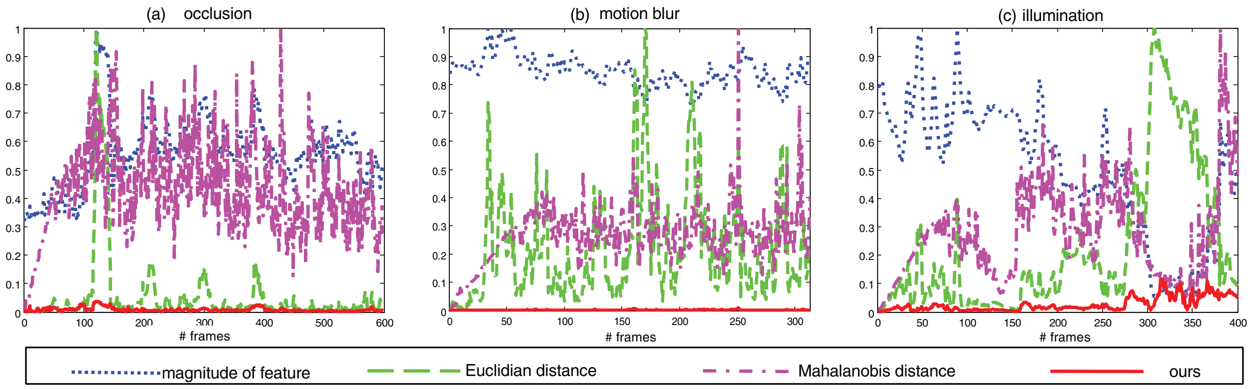

Fig. 1. The matching performance under different distance measurements in complex environments. The *x*-axis indicates the frame number. The blue line indicates the magnitude of target feature in each frame. The green and magenta lines respectively indicate the Euclidian distance and Mahalanobis distance between the predefined target and the target in current frame. And the red line indicates the proposed distance measurement (which is calculated as the difference between 1 and the proposed spherical similarity). Ideally, we expect the distance between the predefined target and the current target to approximate zero, which implies that the target of concern in each frame can be identified consistently and robustly. The target in video sequences (a) *woman*, (b) *jumping* and (c) *skating1* respectively suffers heavy occlusion, motion blur and illumination change [56]. The magnitude, Euclidian distance and Mahalanobis distance change dramatically in these scenarios. At the same time, the proposed distance measurement persistently reports promising results. The target is detected by the ground truth and the distance measurement is performed based on the identical feature extract process in our test. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Given a training data set $\{x_i\}_{i=1,\dots,n}$, distance metric learning distinguishes the feature samples among different classes by adapting their distance in the Mahalanobis form, which is written as

$$\begin{aligned} &d_M(x_i, x_j) \\ &\quad = (x_i - x_j)^\top A (x_i - x_j) \\ &\quad = (Lx_i - Lx_j)^\top (Lx_i - Lx_j) \\ &\quad = |Lx_i|^2 + |Lx_j|^2 - 2|Lx_i||Lx_j| \cos\theta_{Lx_i, Lx_j} \\ &where\ A = L^\top L \end{aligned} \tag{1}$$

where $|Lx_i|$ denotes the magnitude of Lx_i and θ_{Lx_i,Lx_j} denotes the intersection angle between Lx_i and Lx_j. From Eq. (1), distance metric learning projects feature samples onto a new metric space by a linear projection L, and adapts their Euclidian distance, i.e., the magnitude and the intersection angle, in the new space.

With this general form of Mahalanobis distance metric, the existing metric learning methods optimize the classification problem towards various objectives. By far as we know, there are very limited methods, if not none of them, being especially designed for visual tracking applications. In experiments, we observe that the magnitude of feature samples in the complex scenarios may fluctuate drastically, which leads to the uncertain Mahalanobis distance d_M in Eq. (1) during tracking process. As shown in Fig. 1, the target in video sequences (a) *woman*, (b) *jumping* and (c) *skating1* respectively suffers heavy occlusion, motion blur and illumination changes [56]. The magnitude fluctuates sharply in these scenarios. And the Euclidean/Mahalanobis distance [36] between the predefined target and the current target in the current frame changes drastically (when the distance is approximated to zero, the target of concern in each frame is expected to be identified consistently and robustly), which may confine the matching performance and limit the tracking accuracy in complex environments. Moreover, most existing metric learning methods learn and adapt the magnitude and the intersection angle in the projected space between all pairs of training samples in an iterative way. Thus they suffer from heavy time consumption, which limits their applications in real-time visual tracking.

To address these problems, the main idea of this paper is to learn a new metric, which is robust when the magnitude of feature vectors changes drastically in complex environments. We achieve the objective by normalizing the magnitude of feature samples in the projected space. This normalization can be viewed as to project feature samples onto a hypersphere, and measure the distance metric between pairs of samples by their similarity, which we call the spherical similarity. The spherical similarity between any pair of samples depends merely on their intersection angle, which is insensitive to the length of vectors and has no positive semi-definite constraint. As shown in Fig. 1, the proposed similarity metric is much more robust than the Mahalanobis/Euclidian distance in some complex scenarios. The hypersphere projection can be constructed by a convex matrix function. This function normalizes the length of all feature vectors, maximizes the similarity between feature samples in the same class, and minimizes the similarity between samples in different classes. Since it is unnecessary to adapt the length of sample vectors during the adjustment of their intersection angle, the matrix function can be solved in a closed form with low computational complexity. Moreover, it can be learned in an online manner, which accelerates the learning process and improves the tracking accuracy in visual tracking applications.