Neurocomputing 191 (2016) 273–285

Contents lists available at ScienceDirect

Neurocomputing

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

Individual adaptive metric learning for visual tracking

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

College of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, PR China

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form
17 November 2015
Accepted 19 January 2016
Communicated by Jinhui Tang
Available online 11 February 2016

Keywords:
Metric learning
Classification
Visual tracking

ABSTRACT

Recent attempts demonstrate that learning an appropriate distance metric in visual tracking applications can improve the tracking performance. However, the existing metric learning methods learn and adjust the distance between all pairwise sample points in an iterative way, which raises the time consumption issue in real-time tracking applications. To address this problem, this paper proposes a novel metric learning method and applies it to visual tracking. The main idea of the proposed method is to adapt the distance from each individual sample point to a few anchor points instead of the distance between all pairs of samples, so as to reduce the number of distances to be adjusted. Based on this idea, we construct a convex matrix function that collapses the sample points to their class centers and maximizes the inter-class distance. Given n training samples in d-dimensional space, the equation can be solved in a closed form with the computational complexity of $O(d^2n)$. This is much more computationally efficient than traditional methods with the computational complexity of $O(dn^2)$ in each iteration (normally in tracking applications, $d \ll n$). Furthermore, the proposed method can be learned in an online manner which is able to accelerate the learning process and improve the matching accuracy in visual tracking applications. Experiments on UCI datasets demonstrate that the proposed learning method is comparable with the traditional metric learning methods in term of classification accuracy but much more time efficient. The comparison experiments on benchmark video sequences show that the tracking algorithm based on our learning approach can outperform the state-of-the-art tracking algorithms.

1. Introduction

During the visual tracking process, matching the target of interest between consecutive frames is critical to tracking performance, due to the fine and dynamic distinction between the visual appearance of the target and background caused by occlusion, background clutter, illumination variation, motion blur and pose change. Instead of matching under a pre-specified or fixed distance metric [1–3], recent attempts [4–10] incorporate metric learning into tracking applications, which learn and adjust the distance metric, so as to yield small distance between different appearances of the target and large distance between the target's appearances and background's appearances. These methods demonstrate that learning an appropriate distance metric in tracking applications can significantly improve the tracking performance.

Normally, distance metric learning methods focus on learning and adapting a Mahalanobis distance between pairs of training samples [11–17]. This can be viewed as to linearly transform the original feature space to a new transformed space. And in the new space, the inter-class distance (calculated by the distance between pairwise samples in different classes) is constrained to be larger than the intra-class distance (calculated by the distance between pairwise samples in the same class), so that the feature samples of the target are separated from feature samples of the background.

However, a straightforward integration of distance metric learning algorithms into tracking methods may be still problematic. Conventional metric learning need to adjust the Mahalanobis distance between all pairs of feature samples in an iterative way. As shown in Fig. 1 (a), for n training samples of dimension d, there are $O(n^2)$ distances to be adjusted in every iteration, i.e., a single iteration of looping through all constraints costs $O(dn^2)$ [18,19]. With the increase of available samples during a tracking process, a straightforward concatenation of these metric learning algorithms and tracking methods may lead to a serious time complexity issue. Additionally, the change of distance between any sample pair would cause the change of distances between other sample pairs, i.e., the adjustment for the distance between any pair of samples would affect other distances. As a result, the objective under the adaption for all pairwise samples simultaneously may be difficult to achieve.

To address these two problems, this paper proposes a novel metric learning approach, which adapts the distance from each individual sample point to a few anchor points instead of the

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

http://dx.doi.org/10.1016/j.neucom.2016.01.052
0925-2312/

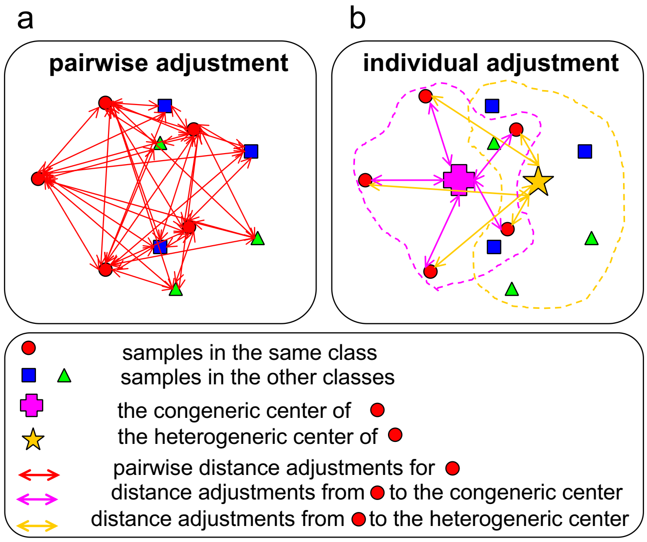

Fig. 1. (a) Conventional metric learning methods adapt the distance between all pairs of sample points. Thus for n training samples, there are $n(n-1)/2$ distances to be adjusted. (b) The proposed method adapt the distance from each individual sample point to its congeneric center and the distance from each point to its heterogeneric center. Thus for n training samples, there are only $2n$ distances that need to be adjusted. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

distances between all pairwise samples, so as to reduce the number of distances to be adapted. As shown in Fig. 1(b), for any sample point, when selecting its class mean (called congeneric center) and the mean of points belonging to the other classes (called heterogeneric center) to be anchor points, the intra-class distance can be expressed by the distances from feature samples to their congeneric centers, and the inter-class distance can be expressed by the distances from samples to their heterogeneric centers. Then the number of distances to be adapted is significantly reduced.

Ideally, under an optimal distance metric, sample points will be collapsed to their congeneric centers and farthest from the other classes. Thus we construct a convex function that projects most sample points to the congeneric centers through a sparsity regularization and maximizes the inter-class distance. This matrix function has a closed form solution with the computational complexity of $O(d^2n)$. Normally, in tracking applications, the number n of training samples is much larger than the feature dimension d [20–25]. Thus the proposed method is much more computationally efficient than traditional metric learning methods (which cost $O(dn^2)$ in each iteration, as discussed before).

Intuitively, the distances to be adjusted on individual samples (represented by the Magenta and Golden lines in Fig. 1) are much less than the distances to be adjusted (represented by the Red lines in Fig. 1) on pairwise training samples, and the adjustment for each individual point is independent to the others. Thus the optimal solution of the proposed individual adaptive metric learning (IAML) algorithm is also easier to achieve than that of conventional metric learning algorithms.

Furthermore, IAML can be learned in an online manner, which is able to accelerate the learning process and improve the matching accuracy in visual tracking applications. We combine IAML with the binary classification based tracking framework [7,9,24,26]. Experimental results on both synthetic data and benchmark video sequences show that IAML can improve the matching accuracy, and its application in visual tracking can outperform the state-of-the-art tracking methods.

This paper contributes to the research of metric learning based tracking in the following ways. (1) A novel distance metric learning method is proposed, which reduces the computational complexity by adjusting the distance from each individual sample to the congeneric center and the heterogeneric center instead of the distances between all pairs of sample points. (2) Based on the idea of adapting individual sample points, we construct a convex matrix function, which collapses the sample points to their respective congeneric centers and maximizes the inter-class distance. This function has a closed form solution with low computational complexity. (3) An online version of the proposed learning method is presented, which can accelerate the learning process and improve the tracking accuracy in visual tracking applications. (4) The proposed metric learning method is incorporated into visual tracking framework. The performance improvement is demonstrated by extensive experiments.

The rest of the paper is organized as follows. Section 2 provides an overview of the related visual tracking and metric learning algorithms. The formulation and solution of our metric learning method is given in Section 3. The use of our metric learning method in visual tracking is presented in Section 4. Experiment results are given in Section 5 and the paper is concluded in Section 6.

2. Related work

Recent attempts of incorporating subspace learning into visual tracking have produced encouraging results, and various objectives for subspace adaption have been proposed. For example, Wang et al. [27] parameterize the instantaneous image motion by a subspace motion model. Ho et al. [28] propose linear subspaces to represent the appearance of the target. Lim et al. [29,30] incrementally learn a low dimensional eigenspace representation to reflect appearance changes of the target. Elgammal et al. [31–33] characterize the target with the manifold structure. Liu et al. [34] segment the subspaces by low-rank representation. Mei et al. [35–37] model the target's appearance by the sparse approximation over template sets. Li et al. [38] transform the feature space to a 3D-DCT subspace. Shirazi et al. [39] learn the target's appearance in the non-Euclidean geometry subspace by a Grassmann approach. However, these methods do not take the background information into account, which may lead to tracking failure when the background shares a similar appearance with the foreground.

By contrast, an increasing number of trackers adopt the binary classifiers to separate the target from its background. For examples, Collins et al. [20] adapt the color space to distinguish between the target and its surrounding environment. Grabner et al. [21,40] track the target with an online boosting classifier. Babenko et al. [22] learn the multiple instances with the bag information. Bai et al. [23] distinguish between the target and the background by a weakly supervised ranking support vector machine. Zhang et al. [24] learn multiple weak classifiers to separate the target and the background in each dimension. Yao et al. [41] represent the visual appearance by the Gaussian mixture model and distinguishes the foreground and the background by the KL divergence. Henriques et al. [25] classify the target and the background with a kernel correlation filter.

Based on the classification strategy, learning the distance metric between the feature samples representing the target and the background has produced encouraging results, and various objectives for metric adjustment have been proposed. For example, Wang et al. [4] learn the distance metric by collapsing classes [19]. Another example is that Jiang et al. [5] learn the metric by maximizing the k-NN classification accuracy [18] for differential tracking. There are many other distance metric