Pattern Recognition Letters 80 (2016) 157–164

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Time Varying Metric Learning for visual tracking☆

Jiatong Li[a,b,*], Baojun Zhao[a], Chenwei Deng[a], Richard Yi Da Xu[b]

[a] Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing 100081, PR China
[b] University of Technology, Sydney, 81 Broadway, Ultimo, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 23 February 2016
Available online 7 July 2016

Keywords:
Metric learning
Visual tracking
Wishart process

ABSTRACT

Traditional tracking-by-detection based methods treat the target and the background as a binary classification problem. This two class classification method suffers from two main drawbacks. Firstly, the learning result may be unreliable when the number of training samples is not large enough. Secondly, the binary classifier tends to drift because of the complex background tracking conditions. In this paper, we propose a new model called Time Varying Metric Learning (TVML) for visual tracking. We adopt the Wishart Process to model the time varying metrics for target features, and apply the Recursive Bayesian Estimation (RBE) framework to learn the metric from the data with "side information contraint". Metric learning with side information is able to omit the clustering of negative samples, which is more preferable in complex background tracking scenarios. The recursive Bayesian model ensures the learned metric is accurate with limited training samples. The experimental results demonstrate the comparable performance of the TVML tracker compared to state-of-the-art methods, especially when there are background clutters.

1. Introduction

Visual object tracking is a typical research topic in computer vision that has been widely applied in many areas, such as video surveillance, intelligent traffic and human computer interaction. Although much work has been done on visual tracking, it is still a very challenging problem, and the tracking results may be influenced by many factors including background clutters, appearance changes, illumination variations and abrupt motions.

Over the last few years, a tracking approach called tracking-by-detection (TBD) is proposed which achieves excellent performance [30]. TBD methods have attracted much attention since they are robust to appearance variations, and have great discriminative capacity. The TBD methods usually learn the appearance of an object by training an online binary classifier, where a discriminative model is adopted in the tracking procedure [2,3,11,12,36]. Another main concern of tracking methods is the model update problem. Many effective model update mechanisms have been proposed, such as incremental subspace update [24], online boosting [8,9,21], PN learning [15] and structured SVM [35].

Most of the above existing tracking methods employ a fixed, pre-specified metric during the entire tracking process. For instance, Euclidean metric is most commonly used for feature comparison and chi-square distance for calculating histogram distances. If we can obtain features that are discriminative enough from the distracters, the result is satisfactory in many instances. However, under many circumstances, we often observe that a candidate with the closest match by using pre-specified metric do not always turns out to be the true target-of-interest. Thus finding an appropriate metric is necessary.

Metric learning based tracking methods aim to learn the appropriate metric to better handle the classification between object and the background [14,17,27]. Tsagkatakis and Savakis [27] combine the online metric learning method with nearest neighbor classification to boost the tracking performance. Jiang et al. [14] propose an adaptive metric learning tracking method, where its goal is to find the best extended nearest classifier to maximize the expected number of training data that are correctly classified. Traditional metric learning based methods treat the object and background as binary classification. However, in the real tracking scenarios, there is no need to exert constraint on the entire background patches as one negative class. Imagine that there are two or more background distracters which are distant apart from each other in feature space, then we must learn a classifier that is capable of distinguishing the target from mixture of clusters in the negative class. To avoid this unnecessary complication, in this paper, we

☆ "This paper has been recommended for acceptance by Dr. D. Dembele.".

* Corresponding author at: University of Technology, Sydney, 81 Broadway, Ultimo, NSW 2007, Australia.

E-mail address: Jiatong.Li-3@student.uts.edu.au, ljt_bit@bit.edu.cn (J. Li).

http://dx.doi.org/10.1016/j.patrec.2016.06.017

use the side information that presents a set of pairwise constraints on training data: equivalence constraints that include pairs of "similar" data and inequivalence constraints that include pairs of "dissimilar" data, which can omit background clustering.

Our key motivation is that, in a tracking scenario, it is unnecessary to assume the classes of background. Therefore, it is desirable to provide the so-called "side information" to indicate which data are "similar" or "dissimilar". This side information based metric learning was first introduced in [31]. Furthermore, Yang et al. [33] extend the work and prove that it performs effectively in the Bayesian learning framework with limited training data. In this paper, we introduce the side information based metric learning method into visual tracking, and show that it can effectively promote the tracking performance.

Our main contribution is that we propose a new time varying metric learning model and its Sequential Monte Carlo (SMC) solution for visual tracking. Our method introduces the Wishart Process to model the time varying metric transition and adopts the side information constraint to train the model in a dynamic setting, which provides an appropriate means of defining pair-wise distance between feature samples. In addition, we present a Recursive Bayesian Estimation (RBE) framework to estimate the metric, which can achieve effective learning result with limited training data.

2. Related work

Visual object tracking is one of the traditional research topics in computer vision. During decades of evolution, numerous methods have been proposed. The recent survey and tracking benchmark further promote the development of the field [20,25,30,34].

Many trackers train an online binary classifier to distinguish the object from background. One representative method is Support Vector Tracker [2], which uses the idea of SVM combined with optical flow to enhance the performance. Hare et al. [10] further extent the work to Structured SVM to learn the samples with structured labels, which achieves promising result. Babenko et al. [3] introduce multiple instance learning to collect positive and negative samples into bags to learn a discriminative classifier, so as to overcome the drift problem. Zhang et al. [36] adopt the random projection method to reduce the feature dimension which achieves real-time tracking.

In addition to the binary classifier method. Incremental subspace learning and boosting methods are also introduced to the online tracker. Ross et al. [24] propose an incremental learning method for object tracking based on PCA representation. This method can efficiently learn and update a low-dimensional subspace which is composed by PCA eigenvectors. Boosting-based appearance models have been widely used for visual tracking [8,9,15,21] due to their efficient discriminative learning capabilities. Practically, discriminative haar-like features are selected, and weak classifiers are correspondingly generated and pooled together as a strong classifier for object location. Grabner et al. [8] demonstrate that an online boosting-based feature selection and classification method can improve tracking performance dramatically than off-line classifier based algorithms. In [37], an efficient instance probability optimization based feature selection scheme was exploited for better tracking performance with low computational complexity.

Multiple models or trackers are further adopted to obtain more robust performance. Wei et al. [28] propose to combine generative model with discriminative model to jointly handle complex tracking conditions. Kwon and Lee [18] enrich the Bayesian tracking model to multiple state so as to adapt to complicated tracking scenarios. Kalal et al. [16] address the long term tracking problem by using dual experts to handle the positive and negative distracters, which achieves excellent result. In [19,35], multiple trackers are integrated to form a tracker ensemble to enhance the tracking performance.

Our work inherits the advantages of tracking-by-detection methods but with great differences. Firstly, our method trains the samples with side information constraint, which focuses on discriminating the object from background by omitting the negative samples clustering. Secondly, unlike the gradient based metric learning methods [14,27], our method estimates the metric accurately with limited samples in the Bayesian framework, which is more suitable to the visual tracking.

The rest of the paper is organized as follows: in Section 3, we first introduce the Wishart Process and side information constraint, and then propose our Time Varying Metric Learning model with its SMC solution. In Section 4, we explain how to apply the proposed model to visual tracking. In Section 5, model validation is conducted on synthetic data, then the proposed tracker is evaluated in the 50 sequences tracking benchmark. Finally, we conclude the paper.

3. Proposed model

In this paper, we aim to learn a distance metric to calculate the distance between two feature vectors $\mathbf{x}$ and $\mathbf{y}$, which can be defined as:

$$d_M(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_M = \sqrt{(\mathbf{x}-\mathbf{y})^T M(\mathbf{x}-\mathbf{y})}, \tag{1}$$

where M is a positive semi-definite (PSD) matrix.

3.1. Wishart process

Wishart Process [23] is a stochastic process which is able to generate a sequence of random PSD matrices M_1, ..., M_t over the time. According to the definition of Wishart Process, the relationship between M_t and M_{t-1} is defined as:

$$\begin{aligned} &M_t|\nu, S_{t-1} \sim Wishart(\nu, S_{t-1}) \\ &where \quad S_{t-1} = \nu\Big(\frac{1}{\nu}A^{\frac{1}{2}}\Big)(M_{t-1})^d\Big(\frac{1}{\nu}A^{\frac{1}{2}}\Big)^T, \end{aligned} \tag{2}$$

where M_t is the PSD matrix at time t, and $Wishart(\nu, S_{t-1})$ is the Wishart distribution parameterized by ν and S_{t-1}, which are the number of degrees of freedom and the time-dependent scale parameter respectively. A is a positive definite symmetric parameter matrix that is decomposed by Cholesky decomposition as $A = (A^{\frac{1}{2}})(A^{\frac{1}{2}})^T$. d is the scalar parameter.

Wishart Process is able to model the dynamic behavior of a set of PSD matrices across time. The scale parameter S_t not only defines the time variation of the PSD matrix but also ensures the proposed matrices are positive definite. The parameters A and d control the variation behavior of the PSD matrices. A is interpreted as revealing how each element of PSD matrix depends on the previous PSD matrix. While parameter d denotes the overall strength of the metric evolution relationship. d is proved to be theoretically bound between $(-1, 1)$ [23]. And in practice, d is usually within (0, 1). More details about Wishart Process and its parameter interpretation is referred to [23,29].

3.2. Side information constraint

As mentioned above, we aim to learn a metric to better distinguish positive (target) samples from negative (background) samples by putting no constraint to the negative training data. To achieve this goal, we propose to use pair-wise data constraint instead of treating all the training data as two classes. As a result, by setting the similar pair-wise constraint to positive data, and the dissimilar constraint between pair-wise positive and negative data,