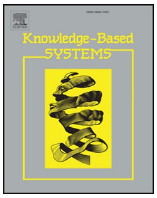
Knowledge-Based Systems 106 (2016) 26–37

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Visual tracking via exemplar regression model

Xiao Ma[a], Qiao Liu[a,b], Zhenyu He[a,*], Xiaofeng Zhang[a,**], Wen-Sheng Chen[c]

[a] School of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[b] School of Mathematics and Computer Science, Guizhou Normal University, Guiyang, China
[c] College of Mathematics and Statistics, Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China

ARTICLE INFO

Article history:
Received 1 December 2015
Revised 29 April 2016
Accepted 13 May 2016
Available online 20 May 2016

Keywords:
Visual tracking
Kernelized ridge regression
Exemplar regression model

ABSTRACT

Visual tracking remains a challenging problem in computer vision due to the intricate variation of target appearances. Some progress made in recent years has revealed that correlation filters, which formulate the tracking process by creating a regressor in the frequency domain, have achieved remarkable experimental results on a large amount of video tracking sequences. On the contrary, building the regressor in the spatial domain directly has been considered as a limited approach since the number of training samples is restricted. And without sufficient training samples, the regressor will have less discriminability. In this paper, we demonstrate that, by giving a very simple positive-negative prior knowledge for the training samples, the performance of the ridge regression model can be improved by a large margin, even better than its frequency domain competitors-the correlation filters, on most challenging sequences. In particular, we build a regressor (or a score function) by learning a linear combination of some selected training samples. The selected samples consist of a large number of negative samples, but a few positive ones. We constrain the combination such that only the coefficients of positive samples are positive, while all coefficients of negative samples are negative. The coefficients are learnt under such a regression setting that makes the outputs fit overlap ratios of the bounding box, where the overlap ratios are measured by calculating the overlaps between the inputs and the estimated position in the last frame. We call this regression *exemplar regression* because of the novel positive-negative arrangement of the linear combination. In addition, we adopt a non-negative least square approach to solve this regression model. We evaluate our approach on both the standard CVPR2013 benchmark and the 50 selected challenging sequences, which include dozens of state-of-the-art trackers and more than 70 datasets in total. In both of the two experiments, our algorithm achieves a promising performance, which outperforms the state-of-the-art approaches.

1. Introduction

Given an initial bounding box of a certain target in the first frame, a visual tracker estimates this target's state, e.g., location and scale, in each frame of the image sequences. Visual tracking is a key component in numerous applications, such as vision-based control, visual surveillance, human-computer interfaces, intelligent transportation, and augmented reality. Although some significant progress has been made in several decades of visual tracking research, most trackers are still prone to failure in challenging scenarios such as partial occlusion, deformation, motion blur, fast motion, illumination variation, background clutter and scale variations.

Among the achievements in visual tracking research, the discriminative approaches [1–9] provide an online mechanism that adapts appearance variations of the target, and achieve better results than their generative rivals [10–13] on some hard sequences [14]. Traditional discriminative approaches [5–7] maintain a classifier trained online to distinguish the target object from its surrounding background. This process can be divided into two stages: searching and updating. During the searching stage, the classifier is utilized to estimate the target's location in a certain search region around the estimated position from the previous frame, typically using a sliding-window [5,7] or particle filtering approach [10]. In the updating stage, the traditional discriminative approaches generate a set of binary labelled training samples which are used to update the classifier online.

Although the traditional discriminative approaches gain convincing results in some hard sequences, it is difficult to arrange the binary labels for training samples in the updating stage. Because it is difficult to determine a pre-defined threshold and rule

[*] Corresponding author. Fax: +86075526032461.
[**] Corresponding author.
E-mail addresses: zyhe@hitsz.edu.cn (Z. He), zhangxiaofeng@gmail.com (X. Zhang), chenws@szu.edu.cn (W.-S. Chen).

http://dx.doi.org/10.1016/j.knosys.2016.05.028
0950-7051/© 2016 Elsevier B.V. All rights reserved.

(e.g., Euclidean distance of one sample from the estimated target location from last frame) to decide whether a sample should be positive or negative.

To avoid the confusion of label arrangement, some discriminative approaches [1–4,8,9] use the regression model instead of the traditional binary classification model during the updating stage. The regression models output a real-value score for each training sample to fit some pre-defined distributions, such as the bounding box overlap ratios [8,9] or a Euclidean distance [1–4]. However, the small sample size training problem makes the traditional regression models, such as the ridge regression, hard to create a robust regressor.

Some progress [1–4] made in recent years has revealed that solving the ridge regression in the frequency domain can achieve a dense sampling since it avoids the small sample size training problem. These approaches are called *correlation filters*. In this paper, different from correlation filters, we propose a simple approach to handle the small sample size training problem in the spatial domain directly. Our approach gives a positive-negative prior knowledge for the training samples. We demonstrate that, by adding such prior knowledge, the performance of the conventional spatial domain ridge regression can be improved by a large margin, even better than its frequency domain competitors-the correlation filters, on most challenging sequences.

Given a candidate sample, our modified ridge regression model learns its score by calculating a linear combination of weighted similarities of some selected samples , as called *support samples* in this paper. Among those weights of similarities, we constrain that only the support samples in the estimated positions in the past frames are positive and the rest of the weights are all negative. This constraint gives us a large number of negative weights in the linear combination, but a few positive ones. We call the modified ridge regression *exemplar regression*. We show that the support samples and the weights can be solved by a non-negative least square method.

The main contribution of this paper is summarized as follows:

- We provide a simple positive-negative constraint method for a common kernelized ridge regression model to construct a robust visual tracker – we call exemplar regression tracking (ERT). The experiments show that the proposed ERT approach gains the state-of-the-art results under the standard CVPR2013 benchmark [1] [14] and other challenging sequences.
- We provide an easy-to-implement approach to solve the ERT based on the off-the-shelf non-negative least square method.

The rest of the paper is organized as follows: Section 3.2 describes the proposed ERT approach. The implementation details of the proposed approach are introduced in Section 4. In Section 5 we discuss the method. In Section 6 we perform an extensive experimental comparison with the state-of-the-art visual trackers and we draw a conclusion in section 7.

2. Related works

We refer to [15–17] for the detailed visual tracking surveys. In this section, we briefly review the most related online single object tracking, especially for the regression based approaches. The visual tracking approaches can be generally categorized as either generative [10,11,13,18–24], or discriminative [1–9,25,26] based on their appearance models.

Generative approaches build an appearance model then use this model to find the optimal candidate samples with a certain region in the image frame which has the minimum construction error. Black et al., [20] learn an off-line subspace model to represent the object of interest for tracking. In [22], every image sample is fragmented into several patches, each of them is represented by an intensity histogram and compared to the corresponding patch in the target region by the Earth Movers Distance. Ross et al., [10] introduce the incremental PCA (Principal Component Analysis) to capture the full range of appearances of the target in the past frames. Mei et al., [11,27] employ the sparse representation to build a dictionary which contains the appearances from past frames, then select the optimal appearance from this dictionary to estimate the target's positions. Li et al., [18] further extend the *l*1 tracking [11] by using the orthogonal matching pursuit algorithm to solve the optimization problems efficiently. In [19], an accelerated version of *l*1 tracking [11] is proposed based on the accelerated proximal gradient (APG) approach. Jia et al., [13] and Wang et al., [12] introduce the patch-based sparse representation to enhance the tracker's robustness. In [21], Kwon et al., combine multiple observation and motion models to handle large appearance and motion variation.

The traditional discriminative approaches pose the tracking problem as a binary classification task with local search and determine the decision boundary for separating the target object from the background. Collins et al., [26] provide an approach which learns the discriminative features online to separate the target object from the background. In [25], an online boosting algorithm is proposed to select features for tracking. Babenko et al., [7] introduce the multiple instance learning approach to decide the labels of the training samples, and integrate an online boosting approach to select the Haar-like features for tracking. A semi-supervised learning approach [6] is proposed in which positive and negative samples are selected via an online classifier with structural constraints. Zhang et al., [5] employ a random projected compressed features space with a very high dimension to represent the target's appearances, and a binary classifier is used to find the optimal image sample in this compressed domain.

One of the typical regression based trackers is Struck [9], which builds a regression model upon the Structural Support Vector Machine (SSVM) framework, and embeds the novel LaRank [28] algorithm into the updating stages. In [9], a regressor is formulated by the weighted linear combination of a set of support vectors. These support vectors might come from different frames during the tracking process. Therefore, Struck has a good ability to adapt of appearance variances. [9] uses the Haar-like features and sliding-window searching strategy, which makes it difficult to handle scale variances of the target. In [8], the authors integrate the Lie group theory to promote Struck's scale-aware ability.

Another popular regression model is Correlation Filters (CFs) [1–4]. In a few years, CFs has proved itself to be competitive with far more complicated approaches. The use of Fast Fourier Transform (FFT) makes it run at a very high speed online(about hundreds of frames-per-second). CFs takes advantage of the famous theorem that the convolution of two image samples in a spatial domain is equivalent to an element-wise product in the Fourier domain. By formulating the CF's objective function in the Fourier domain, it can achieve dense sampling during the updating and searching stage without being very much time consuming.

Comparison of paradigms between the exemplar regression model, Struck and CFs is shown in Table 1. The original Struck [9] applies a sliding-window to find the optimal sample. The regressor in Struck is a linear combination of weighted kernel functions between support samples and the test sample, which is very similar to ours. Struck introduces the LaRank [28] to construct its regressor. CFs use the convolution between the constructed regressor (or filter) and the search region's image patch to find the optimal position. However, a single correlation filter can not

[1] This benchmark is first published in IEEE conference on Computer Vision and Pattern Recognition (CVPR) in 2013: https://sites.google.com/site/trackerbenchmark/benchmarks/v10