Pattern Recognition 47 (2014) 1395–1410

Contents lists available at ScienceDirect

Pattern Recognition

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

Visual tracking via weakly supervised learning from multiple imperfect oracles

Bineng Zhong [a,f], Hongxun Yao [b], Sheng Chen [c], Rongrong Ji [d], Tat-Jun Chin [e], Hanzi Wang [f,*]

[a] Department of Computer Science and Engineering, Huaqiao University, China
[b] Department of Computer Science and Engineering, Harbin Institute of Technology, China
[c] Oregon State University, USA
[d] Department of Electronic Engineering, Columbia University, USA
[e] School of Computer Science, The University of Adelaide, Australia
[f] School of Information Science and Technology, Xiamen University, China

ARTICLE INFO

Article history:
Received 6 June 2012
Received in revised form
28 April 2013
Accepted 2 October 2013
Available online 11 October 2013

Keywords:
Visual tracking
Weakly supervised learning
Information fusion
Online learning
Adaptive appearance model
Drift problem
Online evaluation

ABSTRACT

Notwithstanding many years of progress, visual tracking is still a difficult but important problem. Since most top-performing tracking methods have their strengths and weaknesses and are suited for handling only a certain type of variation, one of the next challenges is to integrate all these methods and address the problem of long-term persistent tracking in ever-changing environments. Towards this goal, we consider visual tracking in a novel weakly supervised learning scenario where (possibly noisy) labels but no ground truth are provided by multiple imperfect oracles (i.e., different trackers). These trackers naturally have intrinsic diversity due to their different design strategies, and we propose a probabilistic method to simultaneously infer the most likely object position by considering the outputs of all trackers, and estimate the accuracy of each tracker. An online evaluation strategy of trackers and a heuristic training data selection scheme are adopted to make the inference more effective and efficient. Consequently, the proposed method can avoid the pitfalls of purely single tracking methods and get reliably labeled samples to incrementally update each tracker (if it is an appearance-adaptive tracker) to capture the appearance changes. Extensive experiments on challenging video sequences demonstrate the robustness and effectiveness of the proposed method.

Crown Copyright © 2013 Published by Elsevier Ltd. All rights reserved.

1. Introduction

Visual tracking has attracted significant attention due to its wide variety of applications, including intelligent video surveillance, human machine interfaces, robotics, and so on. Much progress has been made in the last two decades; an overview is given in the next section. However, designing robust visual tracking methods is still a challenging problem. Challenges in visual tracking problems include non-rigid shapes, appearance variations, occlusions, illumination changes, cluttered scenes, low frame rate, etc.

Years of research in visual tracking have demonstrated that significant improvements may be achieved by using more sophisticated feature selection or target representation, more elaborate synergies between tracking and classification, segmentation or detection, and taking into account prior information of the scenes and the tracked objects. Since each kind of tracking method has its strengths and weaknesses and is applicable for handling one or a few types of challenges, it is difficult, if not impossible, for a single tracking method to work under a variety of tracking scenarios. Many methods often use sequentially cascaded or parallel majority voting frameworks to fuse the outputs of a number of tracking methods. One of the main challenges affecting these two kinds of fusing schemes is how to measure the performance of a tracker when there is no ground truth available.

In this paper, the proposed tracking method is conceptually different and is based on a new strategy; in contrast to using sequentially cascaded or parallel majority voting schemes, we consider visual tracking in a novel weakly supervised learning scenario where labels are provided by multiple imperfect oracles (i.e., different trackers), and no ground truth is given. A probabilistic method is proposed to explore the alternatives of fusing multiple imperfect oracles for visual tracking, and simultaneously infer the most likely object position and the accuracy of each imperfect oracle.

The inspiration for this work comes from a recently developed machine learning area in weak supervision, where the task is to jointly learn from multiple labeling sources [1–6]. This task

* Correspondence to: School of Information Science and Technology, Haiyun Campus, Xiamen University, Xiamen 361005, China. Tel.: +86 592 2580063.
E-mail address: hanzi.wang@xmu.edu.cn (H. Wang).

0031-3203/$ - see front matter Crown Copyright © 2013 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.10.002

underlies several subfields such as data fusion, active learning, transfer learning, multitask learning, multiview learning, learning under covariate shift and distributed inference, which are receiving increasing interest in the machine learning community. The problem of learning from multiple labeling sources is different from the unsupervised, supervised, semi-supervised or transductive learning problems, in that each training instance is given a set of candidate class labels provided by different labelers with varying accuracy, and the ground truth label of each instance is unknown. In practice, a variety of real-world problems can be formalized as multi-labeler problems. For example, there have been an increasing number of experiments using Amazon's Mechanical Turk [7] for annotation. In situations like these, the performance of different annotators can vary widely. Without the ground truth, how to learn classifiers, evaluate the annotators, infer the ground truth label of each data point, and estimate the labeling difficulty of each data point are the main issues addressed by the task of learning from multiple labeling sources. Other examples of a multi-labeler scenario include reCAPTCHA [1], computer-aided diagnosis [4] and search-engine optimizers [5].

To the best of our knowledge, none of the earlier studies have viewed visual tracking as the problem of learning from multiple labeling sources. A new weakly supervised learning based information fusion method is proposed for integrating trackers and leads to encouraging results when it is applied to the task of visual tracking. While most existing fusion-based tracking methods utilize multiple features, the proposed method integrates the results of existing tracking methods which naturally have intrinsic diversity due to their different design strategies. The proposed method has the following advantages:

(1) The proposed method presents a natural way of fusing multiple imperfect oracles to get a final reliable and accurate tracking result. The imperfect oracles can be some imperfect tracking methods in the literature. This avoids the pitfalls of depending on a single tracking method.
(2) The proposed method gives an estimate of the ground truth labeling of training data during tracking in a robust probabilistic inference manner and thus can alleviate the tracker drift problem.
(3) The proposed method can evaluate online the accuracy and trustworthiness of the different tracking methods in the absence of ground truth [8,9]. This allows the best individual method to be used at each time instance.

It is vital to recognize that we are not proposing a single new tracking method, but a new weakly supervised framework to integrate the results of multiple trackers. Therefore, previously established and newly developed trackers can also be potentially exploited by our framework.

Our work is based on the initial version [10]. However, there are a lot of important differences between this work and our previous work, which can be summarized as follows. First, the present paper is rewritten to make it more clear and include a more comprehensive review of related works. Second, additional experimental comparisons with more other state-of-the-art methods on more challenging video sequences are performed to better illustrate the superiority of the proposed method. Third, more discussions about the robustness of the proposed method and perturbations of the solution under simulated dummy trackers have been discussed in the paper.

The rest of the paper is organized as follows. An overview of the related work is given in Section 2. Section 3 introduces our weakly supervised learning formulation for visual tracking, and presents the probabilistic method that jointly estimates the most likely object position and each tracker' accuracy. The detailed tracking method is then described in Section 4. Experimental results are given in Section 5. Finally, it concludes this work in Section 6 and gives suggestions for future research in Section 7.

2. Related work

This section gives a brief review of related tracking methods using online learning and multi-cue fusion techniques.

Although numerous methods have been proposed, robust visual tracking remains a significant challenge. Difficulties in visual tracking include non-rigid shapes, appearance variations, occlusions, illumination changes, cluttered scenes, low frame rate, etc. To solve these challenges, most top-performing methods rely on online learning-based methods [11–15] to adaptively update target appearance. In these methods, visual tracking is formulated as an online binary classification problem and the target model is updated using the images tracked from previous frames. Compared with the methods using fixed target models, such as [16–18], these adaptive methods are more robust to appearance changes. However, the main drawback of these appearance-adaptive methods is their sensitivity to drift, i.e., they may gradually adapt to non-targets.

One popular technique to avoid tracker drift is to make sure the current tracker does not stray too far from the initial appearance model. Matthews et al. [19] are among the first to utilize that technique and provide a partial solution for template trackers. In [20], discriminative attentional regions are chosen on-the-fly as those that best discriminate the current object area from the background region. In that work, tracker drift is unlikely, since no on-line updates of the attentional regions. Furthermore, Fan et al. [63] propose a robust tracking method based on discriminative spatial attention. Grabner et al. formulate tracking as an online semi-supervised learning problem [21]. Combining with a prior classifier, this method takes all incoming samples as unlabeled and uses them to update the tracker. Despite their success, these methods are limited by the fact that they cannot accommodate very large changes in appearance.

To balance semi-supervised and fully adaptive tracking, Stalder et al. [22] present a method using object specific and adaptive priors. In [23], Babenko et al. propose to use a multiple instance learning based appearance model for object tracking. Instead of using a single positive image patch to update a traditional discriminative classifier, they use one positive bag consisting of several image patches to update a multiple instance learning classifier. This method is robust but can lose accuracy if the patches do not precisely capture the object appearance information. In [24–26], co-training is applied to online multiple-tracker learning with different features. The trackers collaboratively classify the new unlabeled samples and use these newly labeled samples with high confidence to update each other. However, independence among different features is required in co-tracking and this condition is too strong to be met in practice.

To incrementally learn from multiple noised data, Lou and Hamprecht [27] propose a structured learning method for cell tracking which formulates tracking by assignment as a constrained binary energy minimization problem. However, the method requires exhaustive assignment annotations of pairs of frames. To address this limitation, they further propose a structured learning method [28] using partial annotations, which has achieved a performance comparable to that obtained from exhaustive annotation. Wang et al. [29] propose an active learning method for solving the incomplete data problem in facial age classification by the furthest nearest-neighbor criterion. In [30], a novel active learning framework is also proposed for video annotation and interactive tracking, in which active learning is used to intelligently query a worker to label only certain objects at

Download English Version:

https://daneshyari.com/en/article/10360411

Download Persian Version:

https://daneshyari.com/article/10360411

Daneshyari.com