Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Visual tracking via shallow and deep collaborative model

Bohan Zhuang*, Lijun Wang, Huchuan Lu

School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China

ARTICLE INFO

Article history:
Received 17 November 2015
Received in revised form
27 July 2016
Accepted 17 August 2016
Communicated by X. Li

Keywords:
Visual tracking
Deep learning
Shallow feature learning
Collaborative tracking

ABSTRACT

In this paper, we propose a robust tracking method based on the collaboration of a generative model and a discriminative classifier, where features are learned by shallow and deep architectures, respectively. For the generative model, we introduce a block-based incremental learning scheme, in which a local binary mask is constructed to deal with occlusion. The similarity degrees between the local patches and their corresponding subspace are integrated to formulate a more accurate global appearance model. In the discriminative model, we exploit the advances of deep learning architectures to learn generic features which are robust to both background clutters and foreground appearance variations. To this end, we first construct a discriminative training set from auxiliary video sequences. A deep classification neural network is then trained offline on this training set. Through online fine-tuning, both the hierarchical feature extractor and the classifier can be adapted to the appearance change of the target for effective online tracking. The collaboration of these two models achieves a good balance in handling occlusion and target appearance change, which are two contradictory challenging factors in visual tracking. Both quantitative and qualitative evaluations against several state-of-the-art algorithms on challenging image sequences demonstrate the accuracy and the robustness of the proposed tracker.

1. Introduction

Visual tracking has long been playing a critical role in numerous vision applications such as military surveillance, human–computer interaction, activity recognition and behavior analysis. The research in designing a robust tracker, which can well handle the challenging factors such as occlusion, illumination variation, rotation, motion blur, shape deformation and background clutter (see Fig. 1), is very attractive.

Current trackers can mainly be categorized into either generative or discriminative approaches. Generative trackers treat the tracking process as finding the candidate most similar to the target object. These methods are mostly based on templates (like [1–7], etc.), subspace (like [8,9]) or inference methods (like [10,11]). Mei and Ling [4] formulate tracking as a sparse coding problem where the target is sparsely represented by the target templates as well as the trivial ones. And in [6], 3D articulated body pose tracking from multiple cameras are proposed to better deal with self-occlusions and pose variations. In [11], Sabirin proposes a novel spatio-temporal graphical models to simultaneously detect and track moving objects for video surveillance.

On the other hand, the discriminative trackers equate tracking as a binary classification problem in order to distinguish the target from the background (like those in [12–16]). In [16], Chu et al. utilize projected gradient to facilitate multiple kernels in finding the best match during tracking under predefined constraints. And further in [12], a set of weak classifiers are combined into a strong one for robust visual tracking. Kalal et al. [13] propose to train a binary classifier from labeled and unlabeled examples which are iteratively corrected by employing positive and negative constraints. Furthermore, several trackers [17–20] are proposed to enjoy the advantages of both generative and discriminative models with good performance. Motivated by this observation, we propose a novel collaborative model, where the generative model employs the shallow feature learning strategy to account for occlusion and the discriminative model adopts the deep feature learning strategy to effectively separate the foreground from the background.

In terms of feature learning, we use deep models to refer to networks that have more than one layer of hidden nodes, and use shallow models to refer to the rest feature learning methods with shallow architectures. Some discriminative tracking methods focus on feature representation by utilizing shallow models. The compressive tracker [21] employs a sparse random measurement matrix to extract the data independent features for the appearance model and separates the object from the background using a naive Bayes classifier. In [24], Grabner et al. propose an online AdaBoost feature selection algorithm to adapt the classifier to the appearance change of the target. Collins et al. [25] use a feature ranking mechanism to adaptively select the top-ranked discriminative

* Corresponding author.
E-mail address: bohan.zhuang@adelaide.edu.au (B. Zhuang).

http://dx.doi.org/10.1016/j.neucom.2016.08.070
0925-2312/

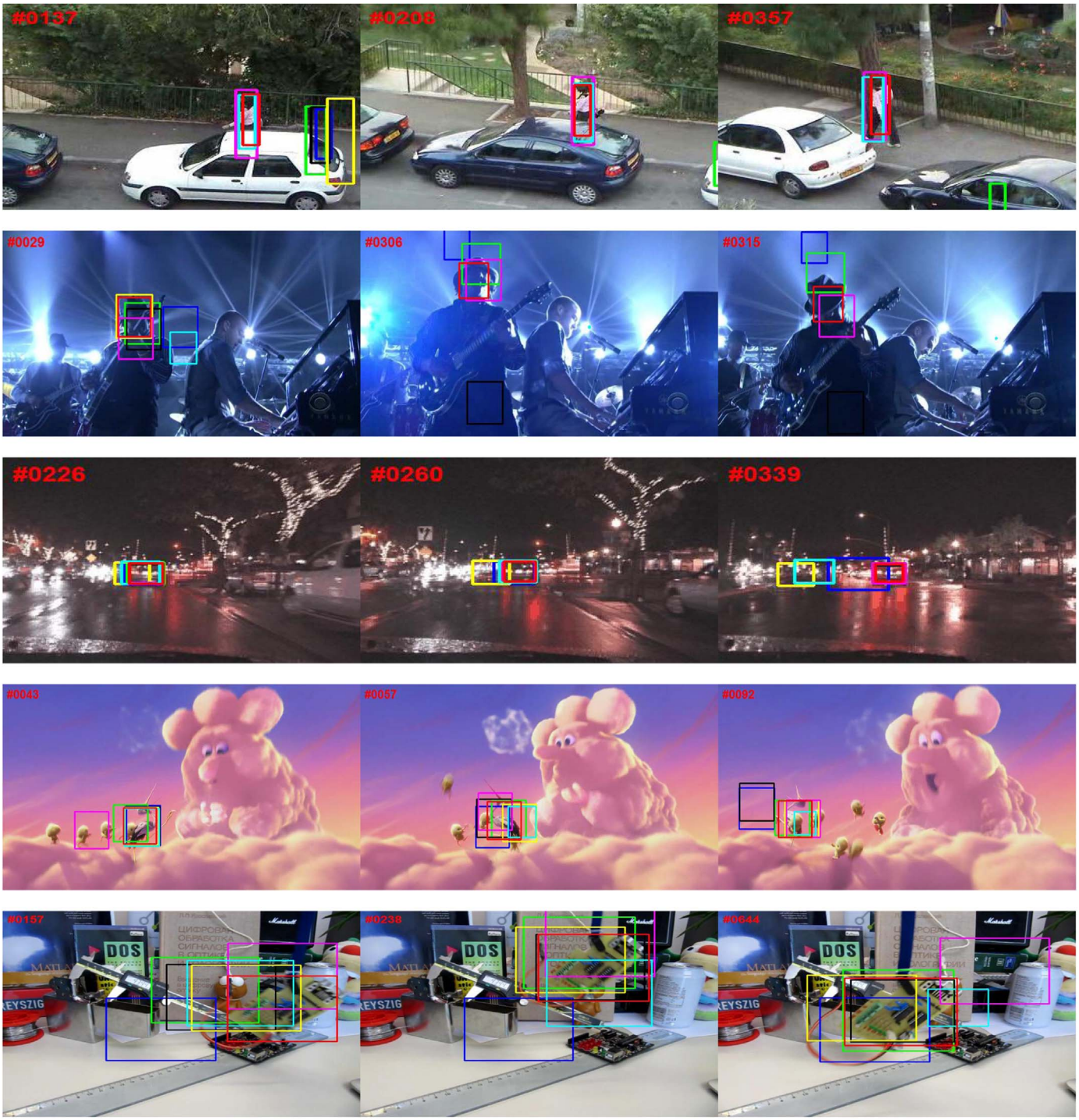

Fig. 1. Challenges during tracking in real-world environments, including heavy occlusion (*woman*), abrupt motion (*shaking*), illumination change (*carDark*), pose variation (*bird*) and complex background (*board*). We use blue, green, black, yellow, magenta, cyan and red rectangles to represent the tracking results of the IVT [9], ASLA [3], OSPT [8], CT [21], Struck [22], DLT [23] and the proposed method, respectively. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

features from multiple feature spaces for tracking. In [26], keypoint descriptors in the region of the interested object are learned online with background information being considered in the meantime. However, due to the difficulty in representing complex functions using limited samples and the restricted capability in generalizing complicated classification problems, the performance of shallow models in tracking scenarios is not satisfactory.

On the other hand, deep learning has been successfully introduced to several computer vision applications, such as image classification [27], face recognition [28] and object-class segmentation [29]. Its aim is to replace the hand-crafted features with the high-level and robust features learned from raw pixel values [30–34]. The deep feature learning strategy demonstrates a strong capability to extract essential characteristics from massive auxiliary data by layer-wisely training a deep nonlinear network. The rich invariant features learned in this way can be further employed in classification and prediction problems, and are empirically shown by our experiments to improve the accuracy and

Please cite this article as: B. Zhuang, et al., Visual tracking via shallow and deep collaborative model, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.08.070