Pattern Recognition 48 (2015) 2964–2982

Contents lists available at ScienceDirect

Pattern Recognition

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

Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle

Jason Kuen, Kian Ming Lim *, Chin Poo Lee

Faculty of Information Science and Technology, Multimedia University, Malaysia

ARTICLE INFO

Article history:
Received 20 June 2014
Received in revised form
9 January 2015
Accepted 17 February 2015
Available online 26 February 2015

Keywords:
Visual tracking
Temporal slowness
Deep learning
Self-taught learning
Invariant representation

ABSTRACT

Visual representation is crucial for visual tracking method's performances. Conventionally, visual representations adopted in visual tracking rely on hand-crafted computer vision descriptors. These descriptors were developed generically without considering tracking-specific information. In this paper, we propose to learn complex-valued invariant representations from tracked sequential image patches, via strong temporal slowness constraint and stacked convolutional autoencoders. The deep slow local representations are learned offline on unlabeled data and transferred to the observational model of our proposed tracker. The proposed observational model retains old training samples to alleviate drift, and collect negative samples which are coherent with target's motion pattern for better discriminative tracking. With the learned representation and online training samples, a logistic regression classifier is adopted to distinguish target from background, and retrained online to adapt to appearance changes. Subsequently, the observational model is integrated into a particle filter framework to perform visual tracking. Experimental results on various challenging benchmark sequences demonstrate that the proposed tracker performs favorably against several state-of-the-art trackers.

1. Introduction

Visual tracking is one of the most important research topics in computer vision because it is in the core of many real-world applications. Applications of such include human–computer interactions, video surveillance, and robotics. Due to the need for generality, recent years have seen the rise of online model-free visual tracking methods which attempt to learn the appearance of the target object over time, without prior knowledge about the object. Despite much research efforts have been made, visual tracking is still regarded as a challenging task due to various appearance changes of the target object and background distractions. Illumination variations, occlusion, fast motion, and background clutters are some challenges in visual tracking.

A typical visual tracking method is dependent on its two major components [1], namely dynamic model (motion estimation) and observational model. A dynamic model is used to model the states and state transition of the target object, whereas an observational model describes the target object and observations based on certain visual representations. To deal with the abovementioned visual tracking challenges, most recent tracking methods tend to put focus on adopting or developing more effective representations. However, variants of image representations (e.g., Histogram of Oriented Gradients (HOG), Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBP)) developed in the computer vision domain are not universally effective on wide-range of vision tasks, and they lack of customizability. One recent and highly effective approach to have better task-specific representations, is to learn representations from raw data itself. Representation learning techniques seek to bypass the conventional way of labor-intensive feature engineering, by disentangling the underlying explanatory factors for the observed input. Thus, representation learning will be the main focus of our approach.

Objects in a video are likely to be subject to small transformations across frames but the content remains largely unchanged. Our work presented in this paper aims to exploit temporal slowness principle to learn an image representation which change slowly over time, thus making it robust against these local transformations. Making use of a big amount of unlabeled tracked sequential data, generic local features invariant to transformations commonly found in tracking tasks can be learned offline. To that end, a complex-cell-like autoencoder model with temporal slowness constraint is proposed for learning separate representations of invariances and their transformations in the image sequences. To learn more complex invariances, a deep learning model is

* Corresponding author. Tel.: +606 2523066; fax. +606 2318840.
E-mail addresses: jason7fd@gmail.com (J. Kuen), kmlim@mmu.edu.my (K.M. Lim), cplee@mmu.edu.my (C.P. Lee).

http://dx.doi.org/10.1016/j.patcog.2015.02.012
0031-3203/

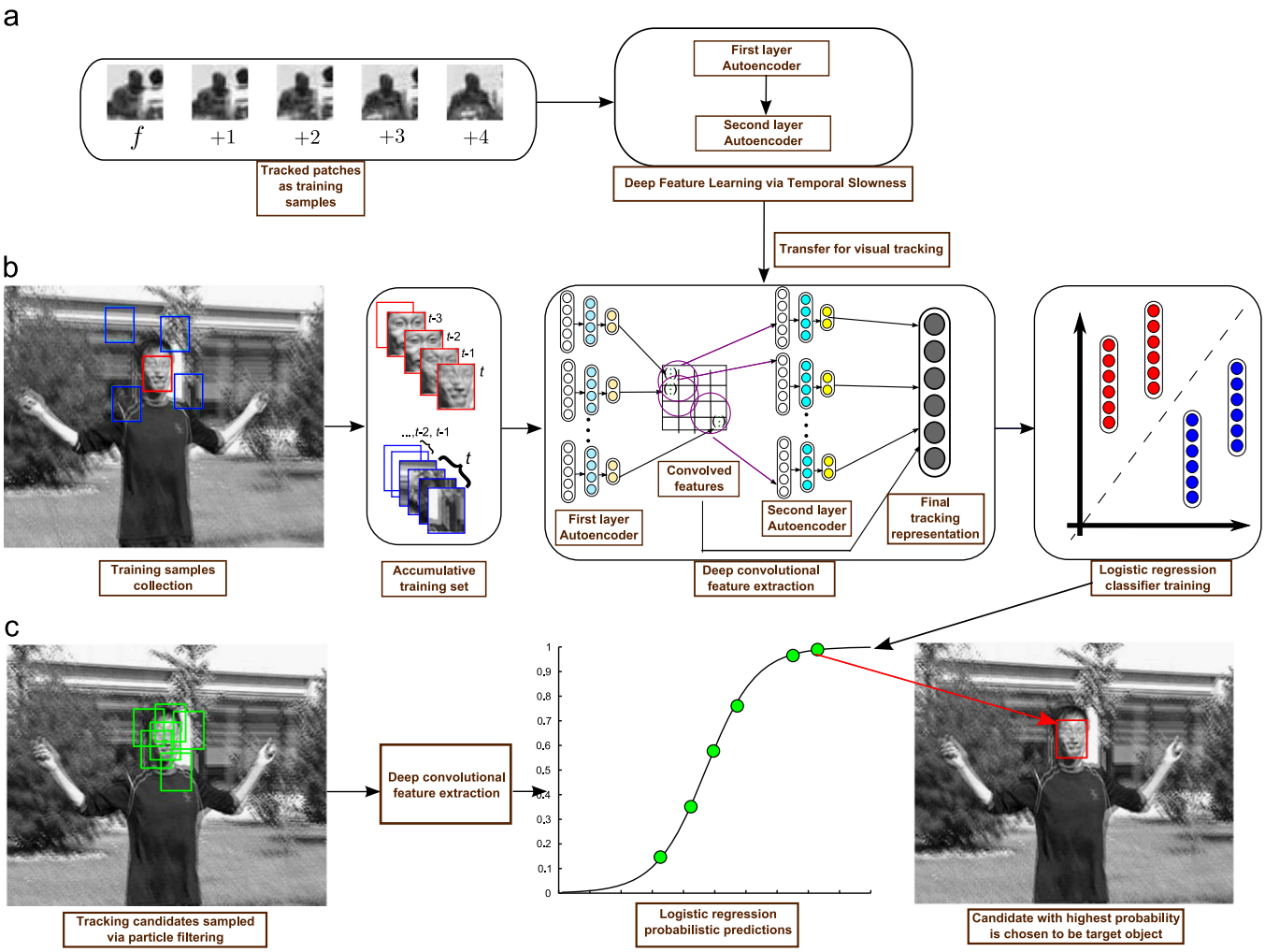

Fig. 1. Overview of proposed tracker in terms of three major stages: (a) offline learning of deep slow representations, (b) observational model update, and (c) tracking.

formed by training a second autoencoder with the convolved activations of the first autoencoders on larger image patches.

The overview of our proposed method with its three major components is illustrated in Fig. 1. Firstly, in Fig. 1(a), tracked image patches are used to train the deep stacked autoencoders via temporal slowness constraint (refer to Fig. 2 for visual details). The trained stacked autoencoders are then transferred to an adaptive observational model for visual tracking (Fig. 1(b) and (c)). Based on certain conditions during tracking, the observational model is updated online to account for appearance changes. Fig. 1 (b) describes the steps for observational model update, whereby logistic regression classifier is trained on an accumulative training set, with the features obtained from the transferred stacked autoencoders. In Fig. 1(c), tracking is performed by sampling tracking candidates via particle filtering. With the learned representation and trained logistic regression, the candidate with the highest predicted probability is chosen to be the target object.

The main contributions of this paper are

1. We present an autoencoder algorithm to learn generic invariant features offline for visual tracking. To train the model, we perform tracking on unlabeled sequential data and obtain tracked image patches as training data. Transformation-invariant features are learned by enforcing strong temporal slowness between tracked image patches. With subspace pooling, we construct a complex-valued representation which separates invariances from their transformations. We further add another autoencoder layer to construct a stacked convolutional autoencoders' model for learning higher-level invariances. The stacked autoencoders are then transferred for use in visual tracking, based on self-taught learning paradigm [2].
2. With the learned representations, we propose an adaptive observational model for tracking. Both first and second layers of the stacked autoencoders are transferred to form a final tracking representation. For better discriminative tracking, the proposed observational model is equipped with a novel negative sampling method which collects more relevant negative training samples. Besides, to alleviate visual drift, we propose a simple technique for the observational model to retain early and recent training samples.
3. We integrate the proposed adaptive observational model into a particle filter framework and evaluate our proposed tracker on a number of challenging benchmark sequences, comparing with several state-of-the-art trackers. Results demonstrate that the proposed tracker performs favorably against the competing trackers.

2. Related work

An observational model, also known as an appearance model is undoubtedly the most crucial component in visual tracking. In this section, literature review is done for existing trackers in terms of