Engineering Applications of Artificial Intelligence 65 (2017) 406–420

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Deep learning to frame objects for visual target tracking

Shuchao Pang [b], Juan José del Coz [a], Zhezhou Yu [b], Oscar Luaces [a,*], Jorge Díez [a]

[a] Artificial Intelligence Center, University of Oviedo, 33204 Gijón, Spain
[b] Coll. Computer Science and Technology, Jilin University, Changchun, 130012, China

ARTICLE INFO

Keywords:
Deep convolutional networks
Deep learning
Target tracking visualization

ABSTRACT

We present a new approach to deal with visual tracking target tasks. This method uses a convolutional neural network able to rank a set of patches depending on how well the target is framed (centered). To cover the possible interferences our proposal is to feed the network with patches located in the surroundings of the object detected in the previous frame, and with different sizes, thus taking into account eventual changes of scale. In order to train the network, we had to create an ad-hoc large dataset with positive and negative examples of framed objects extracted from the Imagenet detection database. The positive examples were those containing the object in a correct frame, while the negative ones were the incorrectly framed. Finally, we select the most promising patch, using a matching function based on the deep features provided by the well-known AlexNet network. All the training stage of this method is offline, so it is fast and useful for real-time visual tracking. Experimental results show that the method is very competitive with respect to state-of-the-art algorithms, being also very robust against typical interferences during the visual target tracking process.

1. Introduction

Among all applications of computer vision, visual target tracking is probably the most challenging problem in recent years, which has also been attracting more and more attention. A wide range of applications are based on visual tracking technology, like vehicle navigation, augmented reality and video safe surveillance (see more examples in Wu et al., 2013). The goal is to locate and follow the trajectory of a moving target in a video sequence starting with little a priori knowledge about the object, mainly the bounding box (e.g. location and size) in the first frame of the sequence. Due to severe visual appearance changes caused by different reasons such as geometric deformation, illumination variations, partial and full occlusions, motion blur, scale changes and/or fast motion, visual target tracking problem is a tough and a challenging task (Wu et al., 2015). These circumstances make target tracking an active research field in the last years.

Among the different tasks that must be tackled to accomplish a visual target tracking system, some authors think that *feature extraction* and *representation* are the most important (Li et al., 2013; Black and Jepson, 1998; Ross et al., 2008). Based on this idea, many algorithms were proposed (Comaniciu et al., 2003; Dalal and Triggs, 2005; Tuzel et al., 2006; Wu et al., 2012; Grabner et al., 2006), and further promoted the development of generative models for target tracking (Liu et al., 2011; Zhong et al., 2012; Jia et al., 2012; Kwon and Lee, 2010, 2011; Ross et al., 2008). In addition to these traditional and hand-crafted features, nowadays, deep features are showing a strong ability in image representation, in the same way than deep neural networks have been successfully applied on many computer vision applications recently (Ma et al., 2015; Wang et al., 2015; Wang and Yeung, 2013; Taigman et al., 2014; Ouyang et al., 2016; Qi, 2016; Carreira et al., 2016).

On the other hand, other papers payed more attention on the classification ability of algorithms in order to decide which candidate image patch could be classified and ranked as the final real target in each frame. This kind of methods generate a discriminative model. For example, some good discriminative models can be found in Babenko et al. (2009), Zhong et al. (2012), Avidan (2004), Hare et al. (2011) and Henriques et al. (2015). However, all the methods proposed only address the tracking problem up to a certain extent. Moreover, analyzing the existing works in this field, we can conclude that (i) they usually extract target features firstly and then find the most similar image patch (called Generative Model) or (ii) they train and fine-tune a classifier to distinguish the positive and negative image patches (called Discriminative Model).

In this paper, we present a robust visual target tracking approach based on deep learning, in which we propose a fusion between both discriminative and generative models.

* Corresponding author.
E-mail addresses: pangshuchao1212@sina.com (S. Pang), juanjo@uniovi.es (J.J. del Coz), yuzz@jlu.edu.cn (Z. Yu), oluaces@uniovi.es (O. Luaces), jdiez@uniovi.es (J. Díez).

http://dx.doi.org/10.1016/j.engappai.2017.08.010
Received 12 May 2017; Received in revised form 10 July 2017; Accepted 17 August 2017
0952-1976/

In our framework, we firstly use a deep learning network to obtain a discriminative object location model. This network will be trained to discriminate from well framed and poorly framed objects. Then, we construct a matching score function to verify which object in the current frame matches the target object set in the first frame. Therefore, the *motion object location* combined with *target verification* is the essence of the visual target tracking proposed in this paper.

This method improves the performance of the tracking process, its training is completely offline, and it is able to deal with a large number of disturbing interferences.

In order to accurately detect motion objects similar to the original target established at the beginning of the tracking process, we use a deep learning network specially tuned to detect correctly framed objects, based on a domain transferred deep convolutional neural network (DT-DCNN) architecture. We call it Deep Framer Network (DFN). To train our DFN, we manually built more than half a million positive (well framed) and negative (poorly framed) object patches. After training our network, we propose a matching score function based on deep features to verify whether the objects from next frame are the tracked target or not. The detailed procedure steps are reported in the next sections.

The major contributions of our work in this paper are:

- we built a large object dataset, with 668 404 images, that can be used to train models to discriminate well framed from poorly framed objects (we will make this dataset public available when publishing the paper—10 GB),
- we deeply analyzed the essence of visual target tracking from a large object dataset and construct the Deep Framer Network (DFN) to discover and frame objects under difficult situations with numerous distracting factors during the tracking process,
- we developed an adaptive matching score function to verify the tracked target in the current frame from a group of candidate framed objects, with no online model update.

2. Related work

In this section, we shortly introduce some works related to our proposed method, including an overview to visual target tracking, as well as to deep neural networks.

2.1. Visual target tracking overview

In the past decade, visual target tracking has been extensively studied and significant progress has also been made in the area of computer vision (Wu et al., 2015). This section provides an overview for visual target tracking from two perspectives, which are (i) tracking models and (ii) target feature representations.

From the point of view of *tracking models*, most tracking methods fall into generative or discriminative models. Those generative methods describe the target appearance by a generative model and search for the candidate with maximum likelihood with respect to the tracked target in the next frame. LSK method (Liu et al., 2011) proposed a local sparse appearance model to enhance visual target tracking robustness by combining with the mean shift algorithm to locate targets. In Jia et al. (2012), the authors regarded the target as the composition of different local image patches with spatial layout based on sparse codes. To deal with drastic lighting changes and fast motion, Kwon and Lee (2010) used multiple observation and motion models to construct a target tracking decomposition approach, which accounted for a relatively large appearance variation. In Kwon and Lee (2011), the sampling of trackers using Markov Chain Monte Carlo was proposed to search for more suitable trackers, which was an extension work based on (Kwon and Lee, 2010). Due to the ceaseless appearance changes of the tracked target, Ross et al. (2008) developed a system that incrementally updates the eigenbasis and it adapts to the changes in the tracking process.

In contrast, discriminative modeling algorithms regard the target tracking problem as a kind of classification problem using a built model to distinguish the target from the background. Babenko et al. (2009) proposed an online target tracking algorithm by constructing the Multiple Instance Learning framework (MIL). In Avidan (2004), a trained Support Vector Machine (SVM) classifier was integrated in an optical flow framework to deal with target appearance changes. Struck algorithm (Hare et al., 2011) utilized kernelized structured SVM to design the target tracking model which can exploit the constraints of the predicted outputs. Henriques et al. (2015) derived a new Kernelized Correlation Filter (KCF) with Discrete Fourier Transform to reduce both storage and computation by several orders of magnitude for target tracking tasks, which showed a powerful discriminative capability between the target and the surrounding environment. In addition, Zhong et al. (2012) tried a fusion target tracking model between a sparsity appearance generative model and a discriminative classifier. Generally, those tracking methods by detection and deep learning methods for target tracking also can be regarded as the special category of discriminative models.

From the view of *feature representations*, there are several features which are used into visual target tracking task. Generally speaking, we can divide the features into two categories, *hand-crafted features* and *deep features*. Hand-crafted features can be usually regarded as artificial, low-level features. They were widely used in target tracking process to address the challenging problem of interferences up to a certain extent. Belonging to this category of features, we can mention, for instance, color histograms (Comaniciu et al., 2003), histograms of oriented gradients (HOG) (Dalal and Triggs, 2005), covariance region descriptors (Tuzel et al., 2006; Wu et al., 2012), and Haar-like features (Grabner et al., 2006).

Hand-crafted features are designed at the pixel-level of the image, which can be unstable due to the numerous disturbances during the tracking process. Moreover, the key point is that pixel-level features belong to shallow features that are not capable to capture deep and semantic information of the target, which will lead to target tracking drift in some challenging video sequences.

Deep features became more attractive for computer vision tasks, since neural networks were again in the spotlight. These compact and semantic feature representations are discovered by a ceaseless iterative training on large image datasets. Many vision tasks use deep features which also proved to be very effective at extracting semantic features and classifying objects of various categories. or visual target tracking, on the one hand, deep features are used as a black box to represent the tracked target, like SDAE (Wang and Yeung, 2013; Hong et al., 2015). On the other hand, several recent visual tracking algorithms began to directly train their deep model with existing target tracking sequences for learning real and semantic target features, such as Nam and Han (2016), Bertinetto et al. (2016) and Zhang et al. (2016).

2.2. Deep neural networks in computer vision

In recent years, deep neural networks have been developed and applied to numerous computer vision tasks, for example, image classification and recognition (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), object detection (Girshick et al., 2014; Ouyang et al., 2016), image segmentation (Long et al., 2015; Qi, 2016), face verification (Taigman et al., 2014), and human pose estimation (Toshev and Szegedy, 2014; Yang et al., 2016; Carreira et al., 2016). The main reasons for deep neural networks to be more popular these days are their accuracy, efficiency and flexibility. Worth of mention are the advances in hardware, which make possible for deep neural networks to deal with problems whose size would prevent us from using such techniques some years ago. More specifically, the great success of deep neural networks in the field of computer vision is mostly attributed to their hierarchical hidden feature layers that outperforms hand-crafted features. However, the popularity of deep neural networks is still not very extended for visual target tracking task since it is hard to collect a large number of labeled images to deal with this problem.