ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133

2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

A Review of Visual Tracking with Deep Learning

Xiaoyu Feng*, Wei Mei and Dashuai Hu

Ordnance Engineering College, Shijiazhuang, 050003, China

***Abstract*—Visual tracking is an important research direction in the field of computer vision and has been widely used in military, medical and other fields. In recent years, the upsurge of deep learning in computer vision provides a new way for the realization of visual tracking with higher performance. This paper firstly introduces the concept and research status of visual tracking and deep learning, then focuses on the representative applications of deep learning in visual tracking, and finally summarizes the future development directions and prospects.**

Keywords-deep learning; object tracking; computer vision

I. INTRODUCTION

Visual tracking is a way that detects, extracts, identifies and locates the object in video sequences. Specifically, once the object in the initial frame of the video sequence is determined, the position and the attribute of the object in the following sequences will be given by tracker automatically, or the prompt will be given if the object is out of view [1].

There are essential differences between visual tracking and traditional tracking like radar tracking or satellite tracking. Visual tracking originates from computer vision (CV) and bases on image data. Visual tracking fuses many fields like image processing, pattern recognition, artificial intelligence, and automatic control, and has broad application prospects in military guidance, video encoding, security monitoring, intelligent transportation, medical diagnosis and meteorological analysis etc.

Visual tracking method includes not only the algorithms for all the procedures of object detection, extraction, recognition and localization, but also the scheme how to organize information and make decision in the overall level. All the procedures of tracking are interdependent and restricted mutually. For instance, the methods of object detection, extraction and the form of object expression determine the method of recognition and the efficiency and robustness of tracking, so we need to consider from the system level.

Since 1950s, visual tracking technology has developed to mature for a long time, and produce a large number of excellent algorithms. Generally, after obtaining videos, algorithms need to extract the features of the object, then establish the motion model and observation model to carry on the localization and the recognition to the object, and update models when necessary. The researches show that Mean Shift [2] and Particle Filter [3] as the representative motion models has been used very well, and at present, most of the researches on visual tracking focus on the acquisition and representation of object information, namely how to extract features and design the observation model.

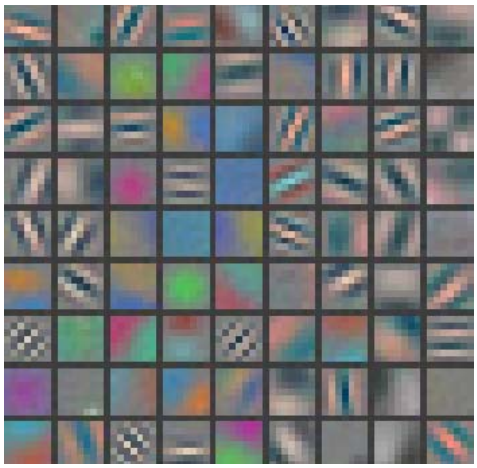

FIGURE I. THE FEATURES EXTRACTED BY A NEURAL NETWORK

Deep learning is the most successful research direction in the field of machine learning. Since proposed in 2006 [4], it has made revolutionary progress and breakthrough in many aspects of information processing such as voice, text, image, video and so on. The advantage of deep learning is mainly reflected in the powerful ability in feature expression. Through the multi-level learning and mapping, deep networks can get high-level abstract features from the edges, colors and other low-level features gradually. These abstract features have high-level dimensions and obvious distinction. Even a simple classifier can achieve high accuracy in classification and regression tasks. Moreover, the feature extraction (shown in Figure 1) is automatic, and don't need professional manual design contrast with traditional feature. Such advantages, which make a huge and complex deep learning network run well, benefit from the large data produced by the Internet and greatly improved computing performance of computer.

In recent years, deep learning improves the performance in computer vision by a huge margin. And in image classification, object detection, object localization, scene classification, almost all the best algorithms are based on deep learning. However, as a traditional computer vision task, visual tracking with deep learning starts later and develops slower than other tasks. There are three main reasons:

- Although deep learning has the advantages in features extraction and observation model expression, but the complex video processing with deep network makes it difficult to achieve real-time tracking.
- The lack of tracking samples limits supervised learning which need a lot of samples.

• Up to date, the effective structure and training method of deep networks special for video processing have been explored.

This paper focuses on the research progress of deep learning in visual tracking direction, introduces the visual tracking algorithms based on deep learning in recent years, and summaries prospect of research finally.

II. METHOD

Visual tracking methods can be divided into two categories according to the observation model: generative method and discriminative method [5]. The generative method uses the generative model to describe the apparent characteristics, and minimizes the reconstruction error to search the object, such as PCA. Discriminative method can be used to distinguish between the object and the background, its performance is more robust, and gradually becomes the main method in tracking. Discriminative method is also referred to as Tracking-by-Detection, and deep learning belongs to this category.

How do we achieve the tracking by detection? Obviously, detect candidate objects for all frames, and use deep learning to recognize the wanted object from the candidates. The following will introduce two kinds of basic network models: stacked autoencoder (SAE) and convolutional neural network (CNN), and show how to use the representative deep networks of them for visual tracking.

A. Stacked Autoencoder

The training data for tracking is regularly only the samples in the initial frame, so the tracking data is very limited. In this situation, non-tracking data is often used to aided the pre-training to obtain the general representation of all kinds of objects. After that, tracking data is used to fine-tune the pre-trained model to obtain the ability of separating object and background for actual tracking. We can see, the idea of migration greatly reduces the training demand of tracking data.

1) DLT. Deep learning tracker (DLT) [6] is the first deep network for tracking task, in which the idea "off-line pre-train+online fine-tune" is proposed. The idea of DLT comes very naturally:

• Off-line unsupervised pre-train the stacked denoising autoencoder using large-scale natural image datasets (Tiny Images dataset) to obtain the general object representation. Stacked denoising autoencoder can obtain more robust feature expression ability by adding noise in input images and reconstructing the original images.

• Combine the coding part of the pre-trained network with a classifier to get the classification network, then use the positive and negative samples obtained from the initial frame to fine-tune the network which can discriminate the current object and background. DLT uses particle filter as the motion model to produce candidate patches of the current frame. The classification network outputs the probability scores for these patches meaning the confidence of their classifications, then choose the highest of these patches as the object.

• In the model updating, DLT uses the way of limited threshold. (i.e. if the highest confidence among all particles is below the threshold, the appearance of object will be considered changing severely, so that the current networks can't work already and need to be updated.)

The main components of DLT are shown in Figure 2.

DLT ranked fifth in the OTB50 dataset, but exists obvious problems: the network structure limits the size of input images at a low level, causing blurry features; SAE is not the best model to track because the classification ability of tracking is not completely same as the reconstruction ability of SAE. So followed by the SO-DLT [7] a deep network with the idea "off-line pre-train+online fine-tune" of DLT uses large-scale CNN for the first time rather than SAE to solve tracking problem. It is shown better than other state-of-the-art. CNN model will be introduced in following.

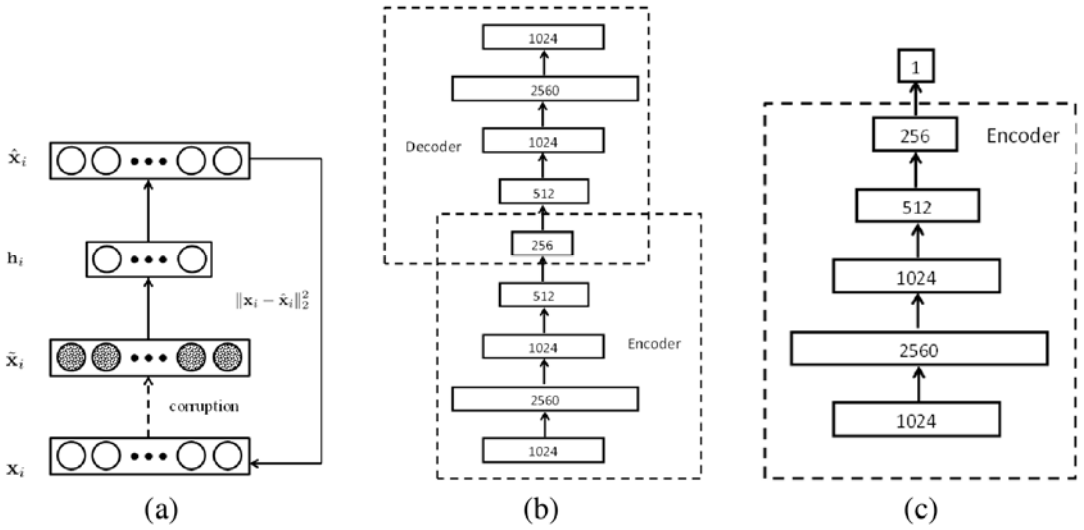

FIGURE II. THE MAIN COMPONENTS OF DLT: (A) DENOISING AUTOENCODER (B) STACKED DENOISING AUTOENCODER (C) ONLINE TRACKING NETWORK [6].

B. *Convolutional Neural Network*

Due to the superiority in image recongnition, CNN has become the mainstream deep model in computer vision, so does it in visual tracking. The general way is using of off-line trained large-scale CNN as both classifier and tracker. Starting from SO-DLT, a large number of CNN-based tracking algorithms appeared, and two representative of them are fully convolutional network tracker [8] (FCNT) and multi-domain convolutional neural network [9] (MD Net).

1) FCNT. As the representative of CNN used in tracking, FCNT analyses and takes advantage of the feature maps of VGG model successfully which is well pre-trained ImageNet, and gets the following observations:

- CNN feature maps can be used for localization and tracking.
- Many CNN feature maps are noisy or un-related for the task of discriminating a particular object from its background.
- Higher layers capture semantic concepts on object categories, whereas lower layers encode more discriminative features to capture intra class variations.

Due to these observations, FCNT designs the feature selection network to select the most relevant feature maps on the conv4-3 and conv5-3 layers of the VGG network and avoid overfitting on noisy ones, and then designs two channals SNet and GNet for the selected feature maps from two layers' separately. The GNet captures the category information of the object and the SNet discriminates the object from background with similar appearance. Both all the networks are initialized with the given bounding-box in first frame to get heat maps of object, and for new frames, a region of interest (ROI) centered at the object location in last frame is cropped and propagated. At last, through SNet and GNet, the classifier gets two heat maps for prediction and the tracker decides which heat map will be used to generate the final tracking result according to whether there are distractors.The pipeline of FCNT are shown in Figure 3.

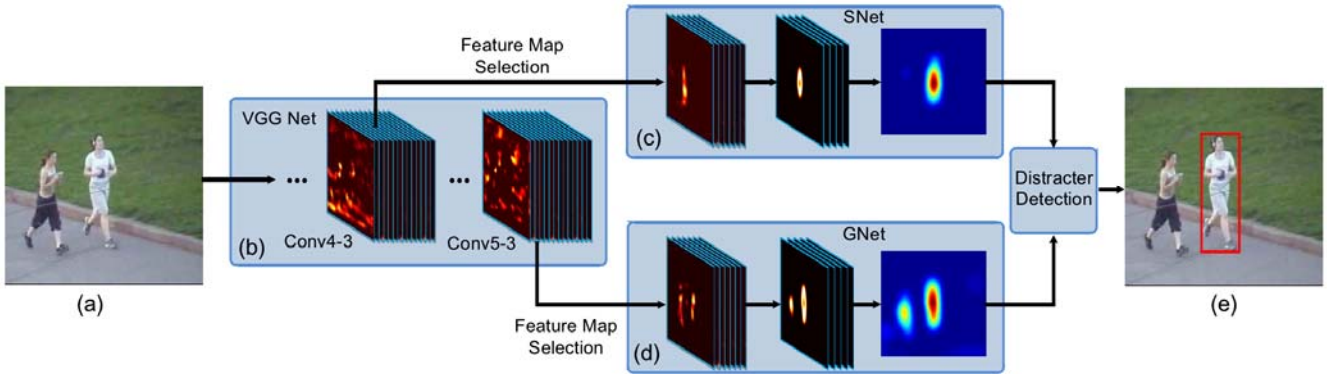

FIGURE III. THE PIPELINE OF FCNT [8]

FCNT constructs a feature selection network and two complementary heat maps prediction networks, which effectively suppresses the tracking drift caused by deformation of objects and distractors. FCNT made a new breakthrough in the OTB50 dataset: precision plots of OPE reaches 0.856, and success plots of OPE reaches 0.599,as shown in Figure 4.

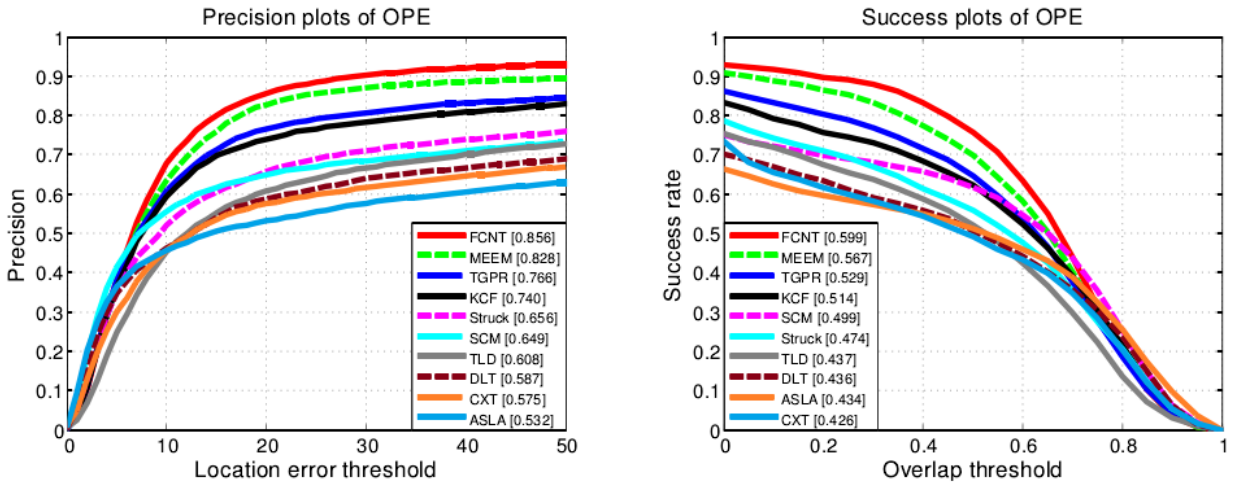

FIGURE IV. THE PRECISION PLOTS AND SUCCESS PLOTS OF OPE FOR THE TOP 10 TRACKERS [8]

It is worth noticed that the classification task needs to recognize objects in different species, what's more, the tracking task also needs to recongnize one object in different appearances. The difference between two tasks motivate to fuse of multi-layer features of CNN, because CNN for the classification more focuses on the differences between cluster, ignore the differences within cluster.

2) MD Net. Different from the idea of FCNT, MD Net uses all the sequences of a video to pre-train to adopt tracking task. The networks mentioned above use irrelevant image data to reduce the training demand of tracking data, and this idea has some deviation from tracking. The object of one class in this video can be the background in another video, so MD Net proposes the idea of multi-domain to distinguish the object and background in every domain independently. And a domain means a set of videos which contain the same kind of object.

As is shown in Figure 5, MD Net is divided into two parts: the shared layers and the K branches of domain-specific layers. Each of the branch contains a binary classification layer with softmax loss, which is used to distinguish the object and background in each domain. and the shared layers sharing with all domains to ensure the general representation. In order to validate the generalization ability, MD Net use the training data and validation data alternately between the OTB100 dataset and VOT dataset of 2013~2015. In addition, MD Net adopts some strategies from the detection task to avoid drifting, such as hard negative mining, bounding-box regression and so on. Finally MD Net won the VOT 2015 champion, also achieved an amazing score in the OTB50 dataset: precision plots of OPE reaches 0.942, and success plots of OPE reaches 0.702. As for the design of end-to-end network and the real-time tracking, MD Net still has great room for improvement.

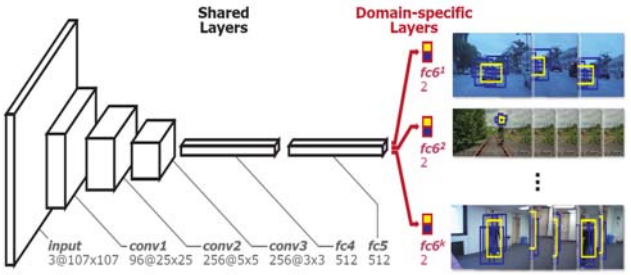

FIGURE V. THE STRUCTURE OF MD NET[9]

III. SUMMARY

Deep learning is a powerful features learning method. Although its progression in visual tracking is less than its in video recognition and video object detection, however researchers try different ways make deep learning adapt to features of visual tracking task. For the research of deep learning applied in tracking, there are many directions worth exploring:

- Apply other network models: there have been the tracking models of DBN [10] and RNN [11] which are non-mainstream but have achieved good results.
- Design the network structure: adapt to the video processing and the end-to-end learning, meanwhile enhance the tracking effect.
- Optimize the process, structure and parameters: ensure the real-time tracking and balance the performance between speed and effect.
- Combining deep learning with the traditional method of CV or the achievement from related areas.
- Design the training for video tracking: dig the spatial and temporal correlation of object data in the video.

This paper firstly introduces the concept and research status of visual tracking and deep learning, then focuses on the representative applications of deep learning in visual tracking, and finally summarizes the future development directions and prospects. Hope to researchers who is interested in related fields for reference.

REFERENCES

[1] A. Yilmaz, O. Javed, Object tracking: a survey, ACM Computing Surveys, 2006, 38(4): 1-45.

[2] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, Trans on Pattern Analysis and Machine Intelligence, 2003, 25(5): 564-577.

[3] J. Carpener, P. Clifford, P. Fearnhead, An improved particle filter for non-linear problems, Radar, Sonar Navigation, 1999, 146(1): 1-7.

[4] G. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks, Science, 2006, 313(5786): 504-507.

[5] R. Raina, Y.R. Shen, N. Andrew, Classification with hybrid generative/discriminative models, NIPS. (2003)

[6] N.Y. Wang, D.Y. Yeung, Learning a deep compact image representation for visual tracking. In NIPS. (2013)

[7] N.Y. Wang, S.Y. Li, A Gupta, Transferring Rich Feature Hierarchies for Robust Visual Tracking, CVPR.(2016)

[8] L. Wang, W. Ouyang, X. Wang. Visual tracking with fully convolutional networks, ICCV.(2015)

[9] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, CVPR.(2016)

[10] G.E. Hinton, S Osindro, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Computation,2006, 18(7): 1527-1554.

[11] A. Milan, S.H. Rezatofighi, Online multi-target tracking using recurrent neural networks, CVPR.(2016)