The Application of Neural Network in Multiple Object Tracking

Jiahui Wang, Xiaoshuang Zeng, Wenjie Luo and Wei An

ABSTRACT

Multiple object tracking (MOT) is one of the basic issues in the field of video analysis and monitoring. It has great significance in areas of behavioral event understanding, traffic management and security prevention. Nowadays, there are increasingly deep researches on the application of neural networks in MOT, such as feature extraction, model formulation of both appearance and motion. Compared with the traditional MOT with the employment of hand-crafted features and the design of similarity function between detections, research in neural networks has shown competitive superiority and drawn wide attention from scholars. In this paper, we analyze the current trends and introduce the application of convolutional neural network and recurrent neural network in MOT. We can see that the neural network techniques in MOT have great potential and vast development prospects.

INTRODUCTION

Multiple object tracking is an important matter in the computer vision. It is meaningful for video surveillance, robot navigation and localization, intelligent traffic and other fields. Although some progress has been made in recent years, MOT still faces great challenges due to false alarms, long time occlusion, and camera movement, etc. In a MOT task, objects in the image sequence are detected by effective detector such DPM, then multiple object tracking is transformed into a data-related problem.
Before 2015, researchers mainly looked for a strong, globally optimal model to solve data association problems. The problem of linking multiple object detections into a series of constant trajectories is often modeled as a graph, solved with the K shortest path [1], linear programming [2], conditional random field [3], a cutting graph [4]. There are also many scholars concerned about the motion model.

Compared with other fields, the application of neural networks in multiple object tracking is the main reasons are as follows. Firstly, neural networks usually need a huge number of training data to learn the large parameters, but we can only get relatively small amounts of data actually. Secondly, existing deep networks are mostly trained on image classification tasks, there are significant differences between different types of images. However, it need to distinguish subtle between detection patches in the tasks of MOT. These networks are difficult to apply to detailed tracking tasks. In addition, the data and solutions required in practical applications are complex and diverse in which the length of the video is unknown.

However, tracking performance is heavily depends on the detection results, hand draft feature have a limited ability of expressing the specific object. The performance of the tracking model in the situation of many pedestrians and frequent occlusion is decreased dramatically. Deep neural networks have been successfully applied to image classification [5], object detection [6], image annotation [7] and other fields, and achieved reliable results. The main reason is that the multi-layered structure of the deep neural network learns more abundant features of the target. More and more researchers are trying to apply deep learning to multi-object tracking tasks. The work of deep learning in multi-object tracking has continued to occur. Research on the appearance and movement model has gradually deepened, and has achieved very good results, setting off a new wave of research in the academic community.

**MOT BASED ON CONVOLUTIONAL NEURAL NETWORK**

In view of the great advantages of Convolutional Neural Network (CNN) in feature expression, scholars try to use neural network to extract the features of the object of interest as the basis for data association.

**Feature Learning**

Fengwei [8] and others used the Faster-RCNN network to detect and extract the targets in the video, and used the GoogLeNet network [9] to extract the appearance features of different detections and calculate the distance between different detection features as a measure of the attractiveness in the data association. This model achieves good results in tracking tasks. Kim [10] et al. also used depth features trained on large datasets by deep networks as appearance features for multi-
hypothesis tracking. Tang [4] et al. also used the depth matching feature to enhance the graph cutting effect.

Learning of Feature Distance Measurement Functions

As multi-object tracking is intended to correlate different detections, designing a distance measure function between pairs is also a key factor. The Siamese network is a good way to measure the similarity of the target. The network uses contrast loss, making the distance between different detections belonging to the same target closer. Therefore, the neural network is applied in multi-target tracking.

Leal-Taixe et al [11] used the twitch neural network to learn matching features between paired detections, using image blocks and optical flow features as network inputs, and then comprehensively detecting information such as position and size, using a gradient-enhanced classifier to generate the final classification probability. The model framework is shown in Figure 1

Data Association Learning Model

In addition to using convolutional neural networks to extract features and learning feature distance metrics, convolutional neural networks are also used to learn appearance models, data association models, and so on. The processing algorithm of the model is shown in Fig. 2. For each frame of the video, the pre-trained detector such as DPM detector [12] is used first to detect the objects in each frame. Then the double-threshold strategy [14] is used to generate more reliable small tracklets. Then they train the neural network using some auxiliary data. The pre-trained siames convolutional network is combined with the short-term constraint metric to train the metric matrices respectively applicable to each trajectory segment. Finally, the similarity relationship between trajectory segments is calculated by using the corresponding metric matrix.

The tracklet association is regarded as a Generalized Linear Allocation (GLA) problem and solved by the soft assign algorithm [12]. In order to solve the trajectory cracks and identity switch problems in the tracklet association model, Wang Bing et al. [13] used the twinning neural network and short-term constraints to model the small trajectory appearance model.
Training Based on Parts of Body

In addition to extract the features of the objects using the CNN network, some scholars used the key points [15] or gestures [16] to achieve multi-object tracking.

Insafutdinov [17] uses the CNN network to train body parts detectors to detect various parts of the body such as the head, shoulders, elbows, ankles, etc., and he build a graph to show the relationship between body parts, which is represented by edges.

MOT BASED ON RECURRENT NEURAL NETWORK

How to recover the objects after recoclusion from the blocked state is a problem which is need to be solved. Therefore, the data association of multi-object tracking should be based on the state of the previous states of the object, while the convolutional neural network cannot record the information of the previous moment, and the recurrent neural network (RNN) has the ability to maintain information, so the recurrent neural networks are also widely used.

Milan [19] applied a recurrent neural network to multi-object tracking model, including target state prediction, data association, and state updates. The combination problem of data association was solved by Long Short Term Memory (LSTM). End-to-end multi-target tracking with deep learning is firstly implemented. The model structure is shown in Figure 3.
In addition, Alahi [20] proposed the LSTM model to learn pedestrian movement characteristics and predict its future trajectory. Sadeghian [21] uses the RNN network to calculate the similarity score between the target and the detection. Which is composed of 3 RNNs, which respectively model the appearance, motion and target interaction, and the objects’ timing information, are used in the modeling process.

CONCLUSION

The application of neural networks in multi-object tracking has become more and more broadly. With the huge extensive data sources in information society, multi-object tracking based on deep learning will be an important research direction,
but we have to do a lot of exploration on the model of deep network training can be robustly applied to various scenarios and achieve better tracking performance.

REFERENCES

1. Berclaz J, Fleuret F, Turetken E, et al. Multiple object tracking using k-shortest paths optimization[J]. IEEE transactions on pattern analysis and machine intelligence, 2011, 33(9): 1806-1819.
2. Leal-Taixé L, Pons-Moll G, Rosenhahn B. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker[C]//Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. IEEE, 2011: 120-127.
3. Yang, B. and R. Nevatia, Multi-Target Tracking by Online Learning a CRF Model of Appearance and Motion Patterns. 2014, 107(2): p. 203 - 217.
4. Tang S, Andres B, Andriluka M, et al. Multi-person tracking by multicut and deep matching[C]//European Conference on Computer Vision. Springer, Cham, 2016: 100-111.
5. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.
6. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.
7. Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015: 3156-3164.
8. Yu F, Li W, Li Q, et al. POI: multiple object tracking with high performance detection and appearance feature[C]//European Conference on Computer Vision. Springer, Cham, 2016: 36-42.
9. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Computer Vision and Pattern Recognition (CVPR), 2015.
10. Kim C, Li F, Ciptadi A, et al. Multiple hypothesis tracking revisited[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4696-4704.
11. Leal-Taixé L, Canton-Ferrer C, Schindler K. Learning by tracking: Siamese CNN for robust target association[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 33-40.
12. Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 32(9): 1627-1645.
13. Wang, B., et al., Joint Learning of Convolutional Neural Networks and Temporally Constrained Metrics for Tracklet Association. 2016, IEEE. p. 386 - 393.
14. Huang C, Wu B, Nevatia R. Robust object tracking by hierarchical association of detection responses[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2008: 788-801.
15. Gold S, Rangarajan A. Softmax to softassign: Neural network algorithms for combinatorial optimization[J]. Journal of Artificial Neural Networks, 1996, 2(4): 381-399.
16. Iqbal U, Milan A, Gall J. Posetrack: Joint multi-person pose estimation and tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, 1.
17. Insafutdinov E, Andriluka M, Pishchulin L, et al. Arttrack: Articulated multi-person tracking in the wild[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, 4327.
18. P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In ICCV’13.
19. Milan A, Rezatofighi S H, Dick A R, et al. Online Multi-Target Tracking Using Recurrent Neural Networks[C]//AAAI. 2017: 4225-4232.

20. Alahi A, Goel K, Ramanathan V, et al. Social lstm: Human trajectory prediction in crowded spaces[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 961-971.

21. Sadeghian A, Alahi A, Savarese S. Tracking the untrackable: Learning to track multiple cues with long-term dependencies[J]. arXiv preprint arXiv:1701.01909, 2017, 4(5): 6.