Journal of New Media JNM, vol.1, no.2, pp.63-74, 2019

Review on Video Object Tracking Based on Deep Learning

Fangming Bi[1, 2], Xin Ma[1, 2], Wei Chen[1, 2, *], Weidong Fang[3], Huayi Chen[1, 2], Jingru Li[1, 2] and Biruk Assefa[1, 4]

Abstract: Video object tracking is an important research topic of computer vision, which finds a wide range of applications in video surveillance, robotics, human-computer interaction and so on. Although many moving object tracking algorithms have been proposed, there are still many difficulties in the actual tracking process, such as illumination change, occlusion, motion blurring, scale change, self-change and so on. Therefore, the development of object tracking technology is still challenging. The emergence of deep learning theory and method provides a new opportunity for the research of object tracking, and it is also the main theoretical framework for the research of moving object tracking algorithm in this paper. In this paper, the existing deep tracking-based target tracking algorithms are classified and sorted out. Based on the previous knowledge and my own understanding, several solutions are proposed for the existing methods. In addition, the existing deep learning target tracking method is still difficult to meet the requirements of real-time, how to design the network and tracking process to achieve speed and effect improvement, there is still a lot of research space.

Keywords: Object tracking, deep learning, neural work.

1 Introduction

With the rapid development of the information age, the emergence and development of computer vision make people realize that computers can replace human beings and the environment to transmit information. Computer vision has become an important subject, covering computer graphics, pattern recognition, neural networks and other technologies. With the development of computer science and digital image processing technology, as well as the increasing demand for video surveillance, more and more scholars pay attention to video object tracking methods, especially deep learning moving object tracking methods, because the deep learning method has achieved good results in the fields of natural language processing and image recognition [Xiong, Shen, Wang et al. (2018); Cui,

[1]College of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China.

[2] Mine Digitization Engineering Research Center of the Ministry of Education, China University of Mining and Technology, Xuzhou, 221116, China.

[3] Key Laboratory of Wireless Sensor Network & Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, 201899, China.

[4] Information Communication Technology Department, Wollo University, Dessie Ethiopia, Ethiopia.

[*] Corresponding Author: Wei Chen. Email: chenwdavior@163.com.

McIntosh and Sun (2018); Tu, Lin, Wang et al. (2018)]. At the same time, great progress has been made. The task of object tracking is firstly to acquire video frames by computer, process and analyze them, find independent moving objects in video images, detect the location of moving object areas in subsequent video frames and mark them, so as to prepare for the later analysis of object trajectory, behavior and other information.

At present, tracking algorithm can be divided into two categories: production and discriminates [Xie, Zhang, Qu et al. (2014)]. The production method uses the generated model to describe the apparent characteristics of the object, and then minimizes the reconstruction error by searching for candidate objects. The more representative algorithms are sparse coding, online density estimation and principal component analysis. The production method focuses on the description of the object itself, ignores the background information, and drifts easily when the object itself changes dramatically or is occluded.

In contrast, discriminates method distinguishes objects and backgrounds by training classifiers. This method is often referred to as tracking-by-detection. In recent years, various machine learning algorithms have been applied to discriminates methods, among which the most representative are multi-instance learning methods, boosting and structural SVM [Babenko, Yang and Belongie (2011); Ning, Yang, Jiang et al. (2016); Son, Jung, Park et al. (2016)]. Discriminates method is more robust because it distinguishes background and foreground information, and gradually occupies the mainstream position in the field of object tracking. It is worth mentioning that most of the current deep learning object tracking methods also belongs to the discriminative framework.

As an important research direction in the field of computer vision, video object tracking technology has important research significance and broad development prospects. The application of moving object tracking technology mainly includes the following aspects:

In the field of social security, object tracking technology has been widely used in the safety monitoring system of hospitals, banks, supermarkets and other public places. Through real-time monitoring and tracking of moving objects, abnormal situations can be found in time and alarms can be quickly raised. In the field of intelligent transportation, vehicles can be detected and monitored in real time. In the field of medical diagnosis, moving object tracking technology can be used to diagnose the motion of heart, kidney, micro vessel and other organs, and identify blood vessels. In the field of military guidance, moving object tracking technology is widely used in UAV detection and missile self-detection. In the field of sports, it is used to acquire sports technology video of athletes in daily training and competitions, and to analyze video information to improve athletes' skills. In the field of sports, it is used in many systems, such as dynamic tracking and positioning, battlefield object recognition, enemy object search and positioning, radar monitoring, artillery fire control and so on.

In addition, the technology of moving object tracking is also widely used in other fields, including human-computer interaction, intelligent robots, virtual reality, object recognition, smart home, scientific detection, remote teaching and so on.

Build a deep network suitable for video object tracking. It is necessary to balance the ability of object representation and real-time performance, not only to maintain the advantages of in-depth feature learning, but also to take into account the high real-time requirements of tracking. At the same time, losing spatial information in such operations

as down-sampling in convolution neural network is an obstacle to the tracking task. Therefore, it is necessary to make some improvements in order to truly apply the deep network to tracking [Yin, Chen, Chai et al. (2016); Huang, Chen, Kang et al. (2015); Guan, Xue and An (2016)].

2 Related works

Unlike the trend of in-depth learning in visual field such as detection and recognition, the application of in-depth learning in object tracking is not plain sailing. The main problem is a lack of training data: the magic of depth model comes from the effective learning of a large number of labeled training data, while object tracking only provides the bounding-box of the first frame as training data. In this case, it is difficult to train a depth model from scratch for the current object at the beginning of tracking [Kristan, Eldesokey, Xing et al. (2017); Kristan, Leonardis, Matas et al. (2016); Kristan, Pflugfelder, Leonardis et al. (2014); Kristan, Pflugfelder, Leonardis et al. (2014); Gao, Yang, Zhang et al. (2016); Li, Bi, Zha et al. (2016); Li, Bi, Yang et al. (2015)].

People began to use an online learning method to train the tracking detector, for example, to establish an initial detection model at the location where the first frame target appears, and to update the model in subsequent frames to adapt to the target change. A typical example is a long-tracking TLD [Kalal, Mikolajczyk and Matas (2012)] method proposed by Surrey University doctoral student Z. Kalal.

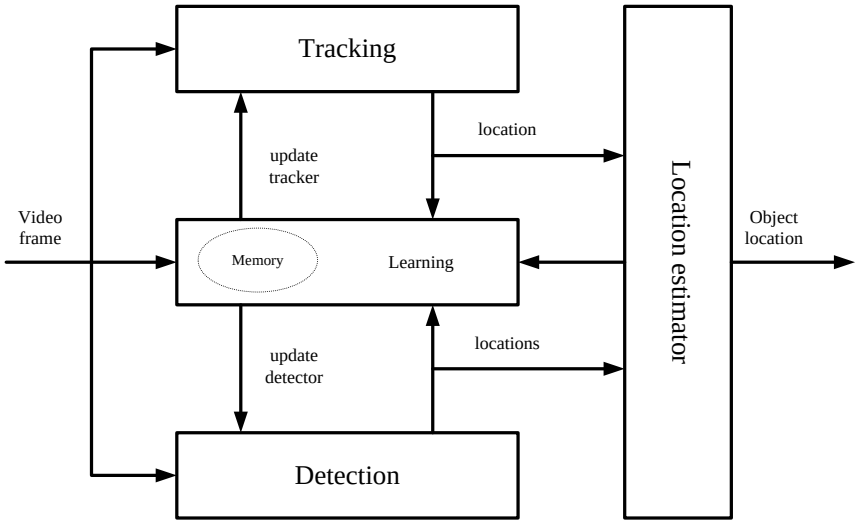

Figure 1: Architecture of TLD

When the training data of object tracking is very limited, the auxiliary non-tracking training data is used for pre-training to obtain the general representation of object features. In actual tracking, the pre-training model is fine-tuned by using the limited sample information of the current tracking object, which makes the model more classified for the current tracking object. It can greatly reduce the need for training samples and improve the performance of tracking algorithm. The representative works in this respect are DLT and SO-DLT [Wang

and Yeung (2013); Wang, Li, Gupta et al. (2015)], both from Dr. Wang Naiyan, Hong Kong University of Science and Technology. DLT is the first tracking algorithm that applies depth network to single object tracking. First, the idea of "off-line pre-training + on-line fine-tuning" is proposed, which largely solves the problem of insufficient training samples in tracking. SO-DLT continues DLT's strategy of using non-tracking data pre-training and online fine-tuning to solve the problem of insufficient training data in the tracking process, and at the same time improves the existing problems of DLT greatly.

Since 2015, there has been a new trend in the application of deep learning in the field of object tracking. That is to say, the CNN trained on large-scale classification databases such as ImageNet [Deng, Dong, Socher et al. (2009)] is directly used to obtain the feature representation of the object, such as VGG-Net [Simonyan and Zisserman (2014)], and then the tracking results are obtained by classifying the observation model. This approach not only avoids the difficulty of training large-scale convolution neural networks directly when tracking, but also makes full use of the powerful representation ability of depth features. Ma Chao's work in ICCV2015 [Ma, Huang, Yang et al. (2015)] combines the feature maps of different layers in convolution neural networks and tracks them in the framework of relevant filtering. In the shallow network, the spatial resolution is higher, but the semantic information of features is less. With the increase of the layers of convolution neural network, the semantic information of features extracted from the deep network is more and more abundant, but the spatial resolution will be lower, which is not conducive to the location of the object. Therefore, the fusion of different layer features is conducive to improving the accuracy of object tracking.

As a representative work of applying CNN features to object tracking, one of the highlights of FCNT [Wang, Ouyang, Wang et al. (2016)] is that it deeply analyses the performance of pre-trained CNN features on ImageNet in object tracking tasks, and designs the follow-up network structure based on the analysis results. Based on the analysis of the characteristics of different layers of CNN, FCNT constructs a feature selection network and two complementary heat-map prediction networks, which can effectively shorten the interference duration and prevent the drift of the tracker, and at the same time is more robust to the deformation of the object itself.

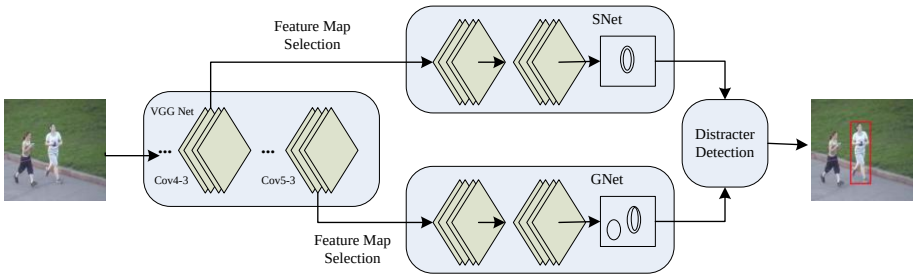

Figure 2: Architecture of FCNT

In the above algorithm, SNet and GNet are initialized in the first frame by minimizing the loss function as follows:

$$L = L_S + L_G,$$
$$L_U = \left\| \hat{M}_U - M \right\|_F^2 + \beta \left\| W_U \right\|_F^2, \tag{1}$$

In addition, SNet is updated with the results of the first frame tracking and the current frame, by minimizing the following formula:

$$\min \beta \left\| W_S \right\|_F^2 + \sum_{x,y} \left\{ \left[\hat{M}_S^1(x,y) - M^1(x,y) \right]^2 + \left[1 - \Phi^t(x,y)\right] \left[\hat{M}_S^t(x,y) - M^t(x,y) \right]^2 \right\} \tag{2}$$

The strategies of solving the problem of insufficient training data and the task of object tracking introduced by the previous algorithms deviate from each other, while MDNet [Nam and Han (2016)], the VOT 2015 champion, gives a better method. MDNet uses a convolution neural network to learn a powerful classifier to separate the object from the background. The network regards each training sequence as a single domain, and each domain has a two-level classification layer (fc6) for it, which is used to distinguish the foreground and background of the current sequence. All the layers before the network are shared by the sequence. In this way, the shared layer achieves the goal of common feature expression in learning and tracking sequences, while the domain-specific layer solves the problem of inconsistent classification objectives in different training sequences.

The procedure of above tracking algorithm is presented as follows:

Algorithm 1 MDNet online tracking algorithm

Input: Pretrained CNN hidden layer three convolutions and two full connections)+initial target state;

Output: Estimated target status

<1> Randomly initialize the last layer of fc6 parameters;

<2> Train a bb regression device;

<3> Get positive and negative samples;

<4> Training, update weight w4, w5, w6;

<5> Update long-term, short-term frame number

<6> **repeat**

1 Obtain candidate targets;

2 Calculate the score to find the best;

3 **if** the score is greater than 0.5 **then**

Update positive and negative samples;

Update the number of long-term, short-term frames;

Bb returned to the finishing position.

4 **if** the score is less than 0.5 **then**

Use short-term positive and negative samples to train update weights

5 **if** the number of frames is mod 100 ==0, **then**

Use the positive samples of long-term and the negative and negative samples of short-

term to train the update weights.

CVPR2016 proposes a SiamFC network for object tracking through similarity learning. The biggest feature is that the trained network is directly used in tracking without updating. The network also uses different layers of feature fusion and border regression to improve the performance of object tracking. There is also a SiamFC network work on ECCV2016, which differs from the previous one in that it uses full convolution network. The advantage of this method is that only one forward operation is needed to get the score of all convolution regions through cross-correlation layer, and the size of the search image is not required to be the same as that of the object image [Bertinetto, Valmadre, Henriques et al. (2016); Tao, Gavves and Smeulders (2016)].

The algorithm uses a sample image and a larger search image to train a full convolution network. This will produce a score of scores (v map) that can efficiently generate many examples. We define the loss function of a score map as the mean of each loss:

$$\mathrm{L}(y,v) = \frac{1}{|D|}\sum_{u \in D} \ell(y[u], v[u]) \tag{3}$$

On the score map, for each position u, a true label $y[u] \in +1,-1$ is required. The parameters of the network need to be trained by the SGD method to solve the following problems:

$$\arg\min_{\theta} \mathrm{E}_{(z,x,y)} L(y, f(z,x;\theta)) \tag{4}$$

If the element of the score map satisfies the following conditions, the object is considered to be positive examples:

$$y[u] = \begin{cases} +1 & if k \|u - c\| \le R \\ -1 & otherwise \end{cases} \tag{5}$$

The network achieves good results on VOT 21015 data set. Compared with other deep learning methods, the speed of the network has a great advantage [Dai, Li, He et al. (2016); Laina, Rupprecht, Belagiannis et al. (2016); Milletari, Navab and Ahmadi (2016); Long, Shelhamer and Darrell et al. (2014); Dai, He, Li et al. (2016); Wang, Wang, Lu et al. (2016)].

In recent years, RNN, especially LSTM and GRU with gate structure, has shown outstanding performance in sequential tasks. Many researchers began to explore how to use RNN to solve existing problems in tracking tasks. When exploring how to use RNN to solve the existing problems in object tracking tasks, RTT uses multi-directional recurrent neural network to model and mine reliable object parts which are useful for overall tracking. In fact, it is a RNN model on two-dimensional plane, and ultimately solves the tracking drift problem caused by accumulation and propagation of prediction errors. It is also an improvement and exploration of part-based tracking method and correlation filtering method. RTT is the first tracking algorithm that uses RNN to model complex large-scale association relations in part-based tracking tasks. Compared with other correlation filter algorithms based on traditional features, RNN is more effective in mining association relations and constraining filters [Graves (2012); Graves, Mohamed, Hinton et al. (2013); Cui, Xiao, Feng et al. (2016)].

Since 2014, the tracking method based on correlation filtering has become a hot topic in object tracking. Among them, Martin from Sweden has done a series of good jobs. Martin proposed an effective fusion method for different spatial resolution features in ECCV2016. By using continuous convolution operation, the interpolation calculation of implicit feature maps was carried out, which solved the training problem for different resolution feature maps. ECO is Martin's latest work. This method uses fewer filters to obtain object tracking results, which can greatly improve the speed of operation [Choi, Chang and Yun (2017); Danelljan, Bhat and Khan (2016); Galoogahi, Fagg and Lucey (2017)]. The starting point of this article is actually to improve time efficiency and space efficiency. In the past one or two years, many methods with good results have been based on correlation filtering. The earliest use of the relevant filtering is the MOSSE of Bolme et al. in 2010 CVPR, which is very fast. After MOSSE, things like KCF, DSST, CN, SRDCF, C-COT, etc. are all based on correlation filtering. As the feature dimension becomes higher and higher, the algorithm becomes more and more complex. Although the tracking effect is gradually improving, it is at the expense of tracking speed. Martin Danelljan has accumulated a wealth of experience in the algorithms related to filtering, and analyzed the three most important factors for speed reduction: model size, training set size and model update. In response to the above three issues, Martin has given coping strategies and improved from three aspects. Summarize the reasons for the good ECO effect: 1. Comprehensive features (CNN, HOG, CN), whose contribution to the results is very huge; 2. Correlation filters are more representative to prevent over-fitting; 3. Training samples have Diversity, reduce redundancy; 4. Update the model of every frame to prevent model drift.

In addition to the research of tracking algorithms, we know that a good data set has greatly improved the research in a certain direction. For example, ImageNet data set has a great contribution to the task of object classification and detection. Recently, some data sets have been published in the field of object tracking. For example, the UAV aerial data set on ECCV2016. And Google's YouTube-Bounding Boxes data set, which is suitable for object detection tasks in video and training and testing of object tracking algorithms.

3 Application research

The task of object tracking is to predict the size and location of the object in the subsequent frames given the size and location of the object in the initial frame of a video sequence. This basic task flow can be divided into the following frameworks: Input the initialization object box, generate many candidate boxes in the next frame, extract the characteristics of these candidate boxes, then score these candidate boxes, and finally find a candidate box with the highest score as the prediction object in these scoring, or enter multiple prediction values. Line merging provides better prediction objects.

Based on the above framework, we can divide the object tracking into 5 main research contents. The first is the motion model, that is, how to generate many candidate samples. The speed and quality of generating candidate samples directly determine the performance of the tracking system. There are two commonly used methods: particle filter and sliding window. Particle filter is a sequential Bayesian inference method, which

infers the hidden state of the object recursively. Sliding window is an exhaustive search method, which lists all possible samples near the object as candidate samples.

Next is feature extraction, which is how to use the feature to represent the object. Discriminative feature representation is one of the keys to object tracking. Commonly used features are classified into two types: hand design features and depth features. The commonly used manual design features are grayscale features, directional gradient histogram, scale invariant features, etc. Unlike the features of artificial design, depth features are learned from a large number of training samples, which are more discriminates than those of manual design. Therefore, the depth feature tracking method is usually very easy to get a good result.

Next is the observation model, that is, how to score for many candidate samples. Most of the tracking methods are mainly focused on the design of this block. According to different ideas, observation models can be divided into two types: generative model and discriminates model. Generative models usually look for candidates that are most similar to object templates as tracking results. This process can be regarded as template matching. The discriminates model trains a classifier to distinguish the object from the background, and chooses the candidate sample with the highest confidence as the prediction result. Discriminates method has become the mainstream method in object tracking, because there are a lot of machine learning methods available.

Next is the updating of the model, that is, how to update the observation model to adapt to the change of the object. The updating of the model is mainly to update the observation model to adapt to the change of the object appearance and prevent the drift of the tracking process. There is no uniform standard for model updating. It is generally considered that the appearance of the object changes continuously, so the model is often updated once per frame. However, some people think that the past appearance of the object is very important for tracking. Continuous updating may lose the past appearance information and introduce too much noise. Therefore, a combination of long-term and short-term updates is used to solve this problem.

Finally, the integration method is to study how to integrate multiple decisions to obtain a better decision result. The integration method is helpful to improve the prediction accuracy of the model, and is often regarded as an effective means to improve the tracking accuracy. The ensemble method can be roughly divided into two categories: selecting the best one among multiple prediction results, or using the weighted average of all predictions [Ning, Yang, Jiang et al. (2016); Wang, Ouyang, Wang et al. (2016); Tao, Gavves and Smeulders (2016); Graves (2013)].

For the SiamFC network, there are many follow-up papers in just one year. It can be said that another direction of object tracking has been created. From the results of VOT2017, the SiamFC series is one of the few surviving end-to-end offline. The training tracker is currently the only direction that can compete with the relevant filtering, and is the most promising direction that can benefit from big data and deep learning.

For the tracking problem of any object, the traditional method is basically to use the separate video itself as the training set, and then learn the apparent model of an object. Although this method is very successful, this online learning method limits the richness of the model. Recently, many people have tried to explore the powerful expressive power of

deep convolution networks. However, since the object in the object tracking is not known in advance, a random gradient descent algorithm needs to be performed online to adjust the network weight each time, which seriously affects the speed of the tracking system.

The SiamFC network actually solves the problem of similarity measure. The similarity between the template image and the position of the image to be detected is calculated by template matching. The point with the highest similarity is the object, network structure. As follows, z is a template picture, and x is a picture to be detected. Both go through the same feature extraction function respectively and finally converge to obtain a final similarity matrix.

4 Conclusion

For tracking arbitrary objects, the traditional method basically uses a separate video itself as a training set, and then learns the apparent model of an object. Although this approach is successful, this online learning approach limits the richness of models that can be learned. Recently, many people have tried to explore the powerful expressive power of deep convolution networks. However, because the object is unknown in advance, a random gradient descent algorithm is needed to adjust the weight of the network every time, which seriously affects the speed of the tracking system.

The problems in video tracking are often caused by many interference factors. The fundamental reasons for the failure of SiamFC model are as follows: Firstly, the target features are not specific, prominent and comprehensive; Then, the space and motion information are not utilized. The last point is the limitations of the search domain method, etc. So we can propose the following solutions:

1. Join the online update strategy

Some methods choose to abandon online updates in order to increase speed, which greatly wastes the target information in the video sequence. However, online update must face two problems, one being how to update. It will affect the speed and effect; the other is the model drift caused by the update, that is, the accumulation of target information error.

2. Processing the first frame label image

The first frame label is the only absolutely trusted target information. In the SiamFC model, the rectangular exemplar image is used to calculate the cross-correlation between the search region and the search region. If the interference of the background information in the exemplar image cannot be reduced, the result will be affected. Therefore, the target image should be further extracted and the background information should be suppressed.

The key factor affecting the speed of the correlation filter tracking algorithm is the three-dimensional feature. For the three-dimensional feature of $m*n*d$, the complexity of the algorithm is $O(d*m*n*\log(m*n))$, $m*n$ being the resolution of the feature, d being the number of feature channels, One of the effective methods of algorithm acceleration is to reduce the three-dimensional feature, but reducing the three-dimensional will greatly affect the performance of the algorithm, so the difficulty is how to reduce the dimension without compromising the performance of the algorithm; One method is the simplification or optimization of the algorithm, which is suitable for the more complex algorithm of SRDCF.

To build a depth network suitable for video object tracking, we need to balance the ability of object representation and real-time performance. We should not only maintain the advantages of deep learning feature learning, but also take into account the requirements of high real-time tracking. At the same time, losing spatial information in such operations as down-sampling in convolution neural network is an obstacle to the application of tracking task. Therefore, necessary improvements are needed to make deep network truly applicable to tracking problem.

Acknowledgement: The research is supported by National Natural Science Foundation of China (Grant No. 51874300), the National Natural Science Foundation of China and Shanxi Provincial People's Government Jointly Funded Project of China for Coal Base and Low Carbon (Grant No. U1510115), National Natural Science Foundation of China (51104157), the Qing Lan Project, the China Postdoctoral Science Foundation (Grant No. 2013T60574). the Scientific Instrument Developing Project of the Chinese Academy of Sciences (Grant No. YJKYYQ20170074).

References

Babenko, B.; Yang, M. H.; Belongie, S. (2011): Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619-1632.

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A. et al. (2016): Fully-convolutional Siamese networks for object tracking. *European Conference on Computer Vision*, pp. 850-865.

Choi, J.; Chang, H. J.; Yun, S.; Fischer, T.; Choi, J. Y. (2017): Attentional correlation filter network for adaptive visual tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4828-4837.

Cui, Q.; McIntosh, S.; Sun, H. Y. (2018): Identifying materials of photographic images and photorealistic computer generated graphics based on deep CNNs. *Computers, Materials & Continua*, vol. 55, no. 2, pp. 229-241.

Cui, Z.; Xiao, S.; Feng, J.; Yan, S. (2016): Recurrently target-attending tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1449-1458.

Dai, J.; He, K.; Li, Y.; Ren, S.; Sun, J. (2016): Instance-sensitive fully convolutional networks. *European Conference on Computer Vision*, pp. 534-549.

Dai, J.; Li, Y.; He, K.; Sun, J. (2016): R-FCN: object detection via region-based fully convolutional networks. arXiv:1605.06409v2.

Danelljan, M.; Bhat, G.; Khan, F. S.; Felsberg, M. (2016): Eco: efficient convolution operators for tracking. *Computer Vision and Pattern Recognition*, pp. 6931-6939.

Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K. et al. (2009): ImageNet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255.

Galoogahi, H. K.; Fagg, A.; Lucey, S. (2017): Learning background-aware correlation filters for visual tracking. arXiv:1703.04590v2.

Gao, J. Y.; Yang, X. S.; Zhang, T. Z.; Xu, C. S. (2016): Robust visual tracking method via deep learning. *Chinese Journal of Computers*, vol. 39, pp. 1419-1434.

Graves, A. (2012): Long short-term memory. *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 37-45.

Graves, A.; Mohamed, A. R.; Hinton, G. (2013): Speech recognition with deep recurrent neural networks. *International Conference on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 6645-6649.

Guan, H.; Xue, X. Y.; An, Z. Y. (2016): Advances on application of deep learning for video object tracking. *Acta Automatica Sinica*, vol. 42, no. 6, pp. 834-847.

Huang, K. Q.; Chen, X. T.; Kang, Y. F.; Tan, T. N. (2015): Intelligent visual surveillance: a review. *Chinese Journal of Computers*, vol. 20, pp. 1093-1118.

Kalal, Z.; Mikolajczyk, K.; Matas, J. (2012): Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409-1422.

Kristan, M.; Eldesokey, A.; Xing, Y.; Fan, Y.; Zhu, Z. et al. (2017): The visual object tracking VOT2017 challenge results. *IEEE International Conference on Computer Vision Workshop*, pp. 1949-1972.

Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R. et al. (2016): The visual object tracking VOT2016 challenge results. *Computer Vision-ECCV 2016 Workshops, PT II*, vol. 9914, pp. 777-823.

Kristan, M.; Pflugfelder, R.; Leonardis, A.; Matas, J.; Porikli, F. et al. (2014): The visual object tracking VOT2014 challenge results. *Computer Vision-ECCV 2014 Workshops, PT II*, vol. 8926, pp. 191-217.

Kristan, M.; Pflugfelder, R.; Leonardis, A.; Matas, J.; Porikli, F. et al. (2014): The visual object tracking VOT2013 challenge results. *IEEE International Conference on Computer Vision Workshops*, pp. 98-111.

Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. (2016): Deeper depth prediction with fully convolutional residual networks. *International Conference on 3D Vision*, pp. 239-248.

Li, H. Y.; Bi, D. Y.; Yang, Y.; Zha, Y. F.; Qin, B. et al. (2015): Research on visual tracking algorithms based on deep feature expression and learning. *Journal of Electronic and Information Science*, vol. 37, pp. 2033-2039.

Li, H. Y.; Bi, D. Y.; Zha, Y. F.; Yang, Y. (2016): An easily initialized visual tracking algorithm based on similar structure for convolutional neural network. *Journal of Electronics and Information Science*, vol. 38, pp. 1-7.

Long, J.; Shelhamer, E.; Darrell, T. (2014): Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, pp. 640-651.

Ma, C.; Huang, J. B.; Yang, X.; Yang, M. H. (2015): Hierarchical convolutional features for visual tracking. *IEEE International Conference on Computer Vision*, pp. 3074-3082.

Milletari, F.; Navab, N.; Ahmadi, S. A. (2016): V-net: fully convolutional neural networks for volumetric medical image segmentation. *Fourth International Conference on 3D Vision*, pp. 565-571.

Nam, H.; Han, B. (2016): Learning multi-domain convolutional neural networks for visual tracking. *Computer Vision and Pattern Recognition*, pp. 4293-4302.

Ning, J.; Yang, J.; Jiang, S.; Zhang, L.; Yang, M. H. (2016): Object tracking via dual linear structured SVM and explicit feature map. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4266-4274.

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. *Computer Science*, pp. 1-14.

Son, J.; Jung, I.; Park, K.; Han, B. (2016): Tracking-by-segmentation with online gradient boosting decision tree. *IEEE International Conference on Computer Visio*n, pp. 3056-3064.

Tao, R.; Gavves, E.; Smeulders, A. W. M. (2016): Siamese instance search for tracking. *Computer Vision and Pattern Recognition*, pp. 1420-1429.

Tu, Y.; Lin, Y.; Wang, J.; Kim, J. U. (2018): Semi-supervised learning with generative adversarial networks on digital signal modulation classification. *Computers, Materials & Continua*, vol. 55, no. 2, pp. 243-254.

Wang, L.; Ouyang, W.; Wang, X.; Lu, H. (2016): Visual tracking with fully convolutional networks. *IEEE International Conference on Computer Vision*, pp. 3119-3127.

Wang, L.; Wang, L.; Lu, H.; Zhang, P.; Xiang, R. (2016): Saliency detection with recurrent fully convolutional networks. *European Conference on Computer Vision*, vol. 2, pp. 825-841.

Wang, N.; Li, S.; Gupta, A.; Yeung, D. Y. (2015): Transferring rich feature hierarchies for robust visual tracking. *Computer Science*, pp. 1-9.

Wang, N.; Yeung, D. Y. (2013): Learning a deep compact image representation for visual tracking. *International Conference on Neural Information Processing Systems*, pp. 809-817.

Xie, Y.; Zhang, W.; Qu, Y.; Zhang, Y. (2014): Discriminative subspace learning with sparse representation view-based model for robust visual tracking. *Pattern Recognition*, vol. 47, no. 3, pp. 1383-1394.

Xiong, Z.; Shen, Q. Q.; Wang, Y. J.; Zhu, C. Y. (2018): Paragraph vector representation based on word to vector and CNN learning. *Computers, Materials & Continua*, vol. 55, no. 2, pp. 213-227.

Yin, H. P.; Chen, B.; Chai, Y.; Liu, Z. D. (2016): Vision-based object detection and tracking: a review. *Acta Automatica Sinica*, vol. 42, pp. 1466-1489.