Opto-Electronic Advances

Original Article

2020, Vol. 3, No. 9

DOI: 10.29026/oea.2020.190018

Visual tracking based on transfer learning of deep salience information

Haorui Zuo[1,2,3]*, Zhiyong Xu[1,2,3]*, Jianlin Zhang[1,2] and Ge Jia[1,2]

In this paper, we propose a new visual tracking method in light of salience information and deep learning. Salience detection is used to exploit features with salient information of the image. Complicated representations of image features can be gained by the function of every layer in convolution neural network (CNN). The characteristic of biology vision in attention-based salience is similar to the neuroscience features of convolution neural network. This motivates us to improve the representation ability of CNN with functions of salience detection. We adopt the fully-convolution networks (FCNs) to perform salience detection. We take parts of the network structure to perform salience extraction, which promotes the classification ability of the model. The network we propose shows great performance in tracking with the salient information. Compared with other excellent algorithms, our algorithm can track the target better in the open tracking datasets. We realize the 0.5592 accuracy on visual object tracking 2015 (VOT15） dataset. For unmanned aerial vehicle 123 (UAV123) dataset, the precision and success rate of our tracker is 0.710 and 0.429.

Keywords: convolution neural network; transfer learning; salience detection; visual tracking

Zuo H R, Xu Z Y, Zhang J L, Jia G. Visual tracking based on transfer learning of deep salience information. *Opto-Electron Adv* **3**, 190018 (2020).

Introduction

Visual tracking is a fundamental problem in computer vision with wide-spread applications in many areas such as auto driving, trajectory guidance, robot navigation, surveillance systems and so on. With the development of artificial intelligence, there are urgent needs of highly efficient and robust visual tracking algorithms. Although much great progress in visual tracking has been made with a wide range of new methods like MD-CNNs (multi-domain convolution neural network)[1], considerable challenges still exist such as pose variation, severe occlusion, and background clutters. Existing appearance-based tracking methods adopt either generative models or discriminative models[2] to distinguish the foreground from the background. Generative methods generate objects representations with appearance models, and the object is the region in the image most similar to the appearance models. After static appearance models are applied to tracking, online appearance models updated frame-by-frame have also been presented. Nowadays, static appearance models are used for visual tracking mostly because the initial location and appearance of the object are given for tracking. Under most circumstances, objects change a lot compared with the initials after tracking for a long time. Consequently, drift occurs and the methods could not be kept stable. Correlation filter[3] is a kind of classic and effective generative method for tracking. These methods show great success in this field.

Aside from generative methods, discriminative methods track objects with a binary classification principle to distinguish the target from the background. These methods extract the typical and predefined features of both the target and background to exploit the information of the image. The traditional classifiers such as random forest[4] and support vector machine (SVM)[5] are often used in

[1]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China; [2]University of Chinese Academy of Sciences, Beijing 100049, China; [3]Key Laboratory of Optical Engineering, Chinese Academy of Sciences, Chengdu 610209, China.
*Correspondence: H R Zuo, E-mail: zuohaorui@sina.com; Z Y Xu, E-mail: xzy158@163.com
Received: 26 May 2019; Accepted: 7 September 2019; Published: 23 September 2020

discriminative methods for tracking. It's worth noting that the traditional classifiers lack the strong power to extract features. Features are important in the field of computer vision but the simple principles of these classifiers lead to a deficiency in representing the target. For visual tracking, the images can be sent into the network directly. Compared with these classifiers, it's apt to perform end-to-end tracking with deep models.

Recently deep learning methods have boosted the overall performance greatly in computer vision tasks including visual tracking. These methods explore the usage of autoencoder[6], RNN[7] and other deep learning models in online tracking. CNN[8] has been used in various computer vision fields and shown state-of-the-art performances. Learning from the biological neuron, CNN's structure and principle are similar to the actual biological attributes of neural network, such as local sensing, parameter sharing and the way of activation. CNN has great potency in extracting features and can be trained on large-scale data[9] for different tasks. For visual tracking, it is truly challenging to learn a unified representation of the video sequences due to the lack of objects' prior knowledge. An important characteristic of human perception is that humans don't process all of the input immediately when there is input to perceive from the outside world. Instead, humans will focus on the interesting part that attracts them most. Attention mechanism is introduced into many fields and the salience detection[10] is a kind of attention model. The tracking objects are of certain salience considering the motion, colors, intensity, and some patterns while they are presented in certain contexts. It's noticeable that salience detection has been beneficial for a variety of computer vision applications. Considering these factors, we propose the MD-CNNs based on transfer learning of salience information. The salience information is mainly used to highlight the salience regions which are the candidate areas of objects. In our paper, visual salience detection[11] is used to provide more prior knowledge and the ability of MD-CNNs is exploited to detect the deep features. We fully take advantages of their similarity in the biological characteristics of vision. We propose a new method of generating samples and make the samples well-distributed to avoid the over-fitting. Our new network architecture shows decent representation power and a good ability to exploit salience information. The main contributions of our paper can be summarized as follow. 1) FCNs[12] are proposed to detect salience and the salience detection ability of the network is exploited. 2) Transfer learning is adopted to build MD-CNNs architecture from FCNs. 3) Extensive experiments are carried out to test our method on some benchmark datasets with large-scale sequences and they show that our method performs well against state-of-the-art trackers.

The remaining of the paper is organized as below: In Section 2, the related work of visual tracking and salience detection is introduced. In Section 3 and Section 4, we give the details of the salience detection model and the new MD-CNN model. We perform the corresponding experiments to validate our methods in Section 5. In Section 6, our contributions are summarized and the conclusions are drawn.

Related work

With deep learning pervading many fields of computer vision, it is also applied in the area of salience detection. Many methods are designed for visual salience such as absorbing Markov chain[13], graph cuts[14], sparse representation[15] before deep learning. After a variety of measures to collect large scales of training data are used and some famous datasets for salience detection have been released, deep learning methods begin to be applied in salience detection. The methods[16] train CNN with available datasets to learn to detect salient objects. Because FCNs can generate pixel-wise maps corresponding to original images, the methods such as Ref.[17–19] generate salience detection result maps via FCNS in an end-to-end manner.

As deep learning shows excellent performance in object detection, researchers start spending efforts in exploring the combination of deep learning and visual tracking due to the similarity between object detection and tracking. Ref.[6] is the first to train deep learning models for visual tracking. The process of online tracking uses a classification neural network. It adopts the structure of the encoder part from the auto-encoder to extract features and an additional layer to do the classification. Ref.[20] recommends CNN in visual tracking. It is faced with the same problem of lack of labeled data. It gains the rich feature hierarchies from an offline trained convolutional neural network and utilizes these data for online tracking to address this issue. The CNN is fine-tuned online to adapt to tracking target specified at the beginning of tracking. MD-CNNs are a convolutional neural network for visual tracking and won the VOT2015 challenge. Its samples are generated from bounding box regression, which is used in Regions-with-CNN[11]. Hundreds of positive and negative samples can be generated this way at

one time randomly. Its convolutional neural network is shallow and adopts three layers from VGG-Net[21] to avoid pre-training to save time. SiameseFC[44], a novel network, is designed to search the objects based on the instance image. For its advantages, the Siamese structure is also used in other computer vision applications. In Ref.[45] Danelljan et al. use the continuous convolution operator tracker (C-COT)[22] to combine shallow-layer feature maps and deep-layer feature maps of CNN to improve time efficiency and space efficiency. The efficient convolution operators (ECO)[23] tracker is the improved version of C-COT and shows better performance. In Ref.[46] a multi-functional inverse sensing approach was used for some specific environments. In Ref.[47] the spatial information is used to leverage the strength of learning the deformation and scale variation information.

Salience detection model

In the human visual system, the visual attention mechanism plays a very important role as the key to ensure the human perception of the environment in time and efficiently. In the field of computer science, the research of visual attention mechanism is to process the given image and highlight the areas of interest. This process is the visual salience detection, visual salience detection is used to identify the most distinct region in a complex scene, which attracts ordinary creature's attention at the first sight. To reduce the complexity of image analysis and exclude useless information, visual salience detection is widely applied to tasks of computer vision including image compression, image segmentation, object recognition, visual tracking and so on. Visual attention prediction[24] and general salient object detection are two typical categories of the salience detection module. The former one is used to predict the specific regions where a human observer may focus on. The aim of salient object detection is to detect specific targets with salience information, which is obvious in the background. According to the different ways of information processing, visual salience detection can be classified into two categories: top-down model and bottom-up model[25]. Top-down model is more complex and aimed at specific tasks. In most current researches, the bottom-up model is preferred since it uses various features to generate the salience measurement. The deep learning methods are applied in visual salience detection after their good performances have been shown in other computer vision fields. In this paper, we make a new model with FCNs shown in Fig. 1 to perform salience detection.

We propose a model adopting FCNs for pixel-wise salience detection. We investigate the deep learning methods for end-to-end training and pixel-wise salience predictions. Due to the lack of enough labeled salience detection training data, we adopt some famous video object segmentation datasets such as FBMS[26], SegTrackV2[27] to provide enough training data. The input of the network is a single frame image and multi-layer convolution networks are applied to transforms the image to deep feature representation. Deconvolution networks are used to up-sample the feature representation inversely. Making use of the fully-convolution network with the 1×1 kernel and the sigmoid activation function, we obtain the output as a probability map of the same size as the input. In the map, higher value means higher salience in the original image. We train our network in an end-to-end manner. We initialize the weights of the first five convolutional layers with VGG blocks and other layers randomly. The way to train the network is stochastic gradient descent (SGD)[28]. We use the overall clustering error and average clustering error as the evaluation metrics. The overall clustering error is the percent of negative labels in the total labels on the basis of the pixel. The average clustering error is different from the clustering overall error in averaging through regions after calculating errors for

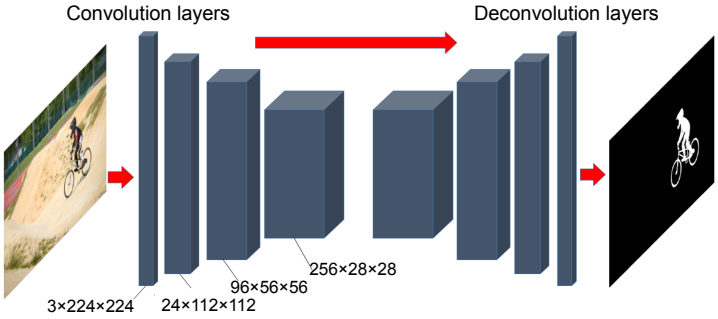

Fig. 1 | The FCNs for salience detection. Our network is designed for static salience detection. It takes a single frame as input and outputs the estimation of the static salience prediction of the image.

every region separately. On FBMS and SegTrackV2 dataset, Global contrast methods[15] attain 37.5% and 45.8% in overall error and average error. The overall error and average error of DHSNet[18] are 34.9% and 40.4%, respectively. Compared with these methods our network gets the results of 34.1% overall error and 39.5% average error. We realize the acceptable salience detection performance and this shows the salience feature extraction ability of our network. As can be seen from the Fig. 2, it gets good results when we test our proposed algorithm on exemplary videos.

Our model

The deep learning methods require a large amount of labeled training data for supervised learning. Prior knowledge is not sufficient in the field of tracking in most cases. To improve the tracking performance by deep learning approaches, we design a new deep learning architecture for tracking by integrating the result of salience detection as prior knowledge. One of the advantages of the architecture is that it makes the target with salience easier to be detected. The deep convolution neural network can extract distinct features of the salient target, which can be exploited to track the target stably.

Basic model structure

The original MD-CNNs can learn domain-independent information from capturing domain-specific representations[29]. On the purpose of discriminating the target from the background, the architecture of the MD-CNN is simpler than those for other tasks such as image classification, object detection. We train the network offline, and fine-tune the networks online with the ground truth in the first frame. The FCNs for salience detection is based on receptive field and appropriate to set up the model with the MD-CNNs. In the following, we incorporate the structure of the network for object detection and the part of the original MD-CNNs. We applied multi-domain-learning to train the CNN, and our training data are obtained from different domains. To make a CNN distinguish the target from the background in any domain is the goal of our learning algorithm. The network should be in small size because it should meet the demand for real-time tracking. In Fig. 3 we adopt the network structure similar to the structure of the network in Ref.[1].

The input of MD-CNNs is color image patch of size 224×224. For tracking, we only need to classify the image

Fig. 2 | Salience sketch image from videos. The 1st row images are from skier video. The 2nd row images are the salience maps of the skier; the 3rd row images are from leopard video; the 4th row images are the salience maps of leopard.

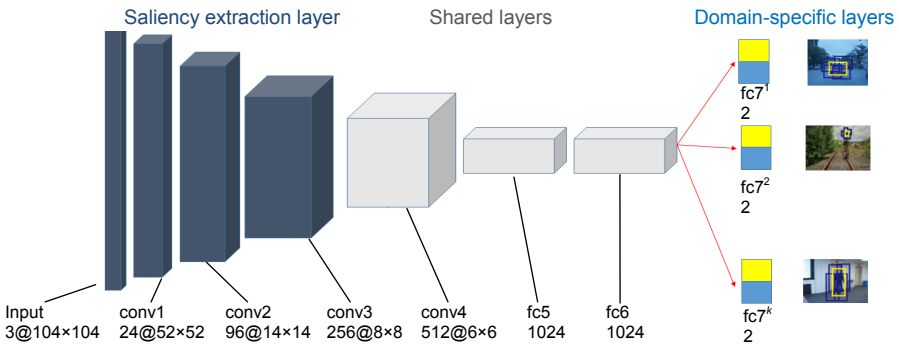

Fig. 3 | The architecture of new multi-domain network.

into the target and background. It is simply a binary classification problem. Let target be denoted as label "1" for output, and background as "0". There are 7 shared layers and the former 4 layers are used for salience detection. Since we have different targets in different tracking sequence, we train the last fully-connected layer as the domain-specific layer.

Samples generating methods

We have analyzed the tracking sequences of the benchmark dataset OTB100 and VOT15 and find that the object and background are updated slowly in a wide range of scenes. They remarkably resemble each other if there is a short length of sequence between them. The changes such as deformation, occlusion and scale variation, just occur occasionally. It can be seen from the Fig. 4.

Due to the bounding-box-regression[11], the method may obtain many similar samples, even identical samples, which easily cause over-fitting on the test set. We propose a new method for generating positive and negative samples, which are sufficient and diverse. The samples are appropriate for training the network. We divide the tracking sequence into different groups because of their difference in target and background. Many factors lead to the various changes between the frames. Illumination variation, scale variation, occlusion, deformation, rotation, background clutters are all the attributes. There are N images in the tracking sequence, and we may divide it into M groups according to their similarities of these factors. It will be 50 N/M positive samples and 200 N/M negative samples totally in this way. Every group is designed to contain a certain number of positive and negative samples, which are proportional to the number of images in the group. If the group consists of T images, it gives out $250T$ samples totally. The sequential images of every group change gradually so the samples we collect

Fig. 4 | Representative images selected from the Car2 sequence of VOT15. As can be seen from the three similar segment sequences, many images show little difference compared to others. When we go through at least 20 frames generally, scale variation, illumination variation, background clutter and so on can be found.

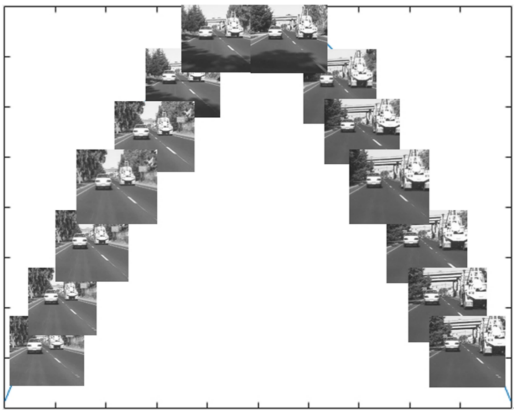

Fig. 5 | The weights for the images distributed by Gaussian distribution to generate certain numbers of samples. As can be seen from the figure, the frames at the beginning and the end of the group will have less weight than those in the middle. The frames located closer to the center of the group are designed to generate more samples because they could make greater distinctions.

from the middle images are more than images at the beginning and at the end of the sequence. We employ the Gaussian distribution $X\sim N(\mu,\ \sigma_2)$ to distribute the weights to every image as shown in Fig. 5. The specified weight represents the number of samples. We set T as $\mu = 0.1\ N/M$, $\sigma = 0.1\ N/M$. In one group the generating sample number of the j-th image can be calculated as

$$Num_j = 250\frac{N}{M} \times X_j \times \left(\sum_{i=1}^{N} X_i\right)^{-1}. \tag{1}$$

We assign the number of samples generated from every image in the group. Hard-negative-mining[1] is adopted to improve the robustness of our system.

Experiments details

Model setting

For deep learning, it requires a large number of labeled data to train. In this paper, we use the salience information to optimize the supervised learning, and test the network in the task of single object tracking. The salience information can be used as guidance for training, which explicitly shows the relationship of the object between the source image and the salience information. According to the results of salience detection, we find that the salience image is capable of providing some information such as the location of the object and partial contour. These are distinct features or individual characteristics in the original image. There is a strong correlation between the salience information and source image. From some perspectives, salience information is the lightweight feature of the image. The performance shows us that although we train our network with a less scale dataset it can get good effects to a certain degree.

Our learning algorithm aims to train a multi-domain CNN to discriminate the target and the background in any different domains. It is not straightforward because in our training data some objects will be regarded as backgrounds from different domains and vice versa. When we set up the model, it needs to process the original data and recognize the tracking objects given in the first frame. We would set this model in this way rather than consider the ground truth of the salience information as the inputs. In our view, the supervised learning with labeled data can make convolution neural network extract the features automatically. We ensure that all operations are performed on a unified scale in the phase of the network training. The salience information detection as an additional procedure of feature extraction actually strengthens the ability of the networks. It can make the network "see" like a creature observer. With the experiments we perform we can make out whether the performances of the model mainly result from the salience information or the expansion of the model.

In the phase of offline training, we use the front part of the FCNs for salience detection to accomplish the initialization of parameter. Our network is trained with the SGD method. Next, we add the samples generated with ground truth to the network training. We set the learning rate as 0.001 and the reduction ratio as 10 after 10k iterations. The momentum is set as 0.9 and the decay is set as 0.0005.

When finishing the training and coming to online tracking, we use a single layer (fc6) to replace the multiple layers domain-specific of ($fc6^1$–$fc6^k$). We pre-train the network with the first frame of the tracking sequence and randomly initialize the last layer's weight and bias. We update the weights of the layers when estimating the position.

Experiment results

We test our algorithm on OTB50, OTB100 and UAV123 Benchmark. One pass evaluation (OPE)[30] is a classic tracker evaluation method and applied in our experiment. It initializes an algorithm with the ground-truth object state in the first frame and reports the average precision or success rate of all the results. The precision plot represents the percentages of frames whose estimated locations are within the range of the threshold to the ground-truth centers. If intersection over union (IOU) between the ground truth and the tracking result get closer to the maximum 1 in our experiment, the algorithm tracks better. In a frame the IOU larger than a threshold is termed as a successful frame. As the thresholds ranged from 0 to 1 the ratios of successful frames are plotted in success plots. The precision and success plots are both intuitive tools to present the performance of the tracker on the test dataset. We think the success plot a better tracking evaluation index than the precision plot because the precision plot hardly reflects the predicted shape similarity. In the precision plot the distance between the center points show that the predicted area is close to the ground truth, but some important factors such as scale, shape, deformation are ignored. Therefore, the precision plot is supplementary to evaluate the tracker.

The legend shows the area under the curve (AUC) score for each tracker. OTB is the open challenging

long-term dataset. OTB100 is the overall data that OTB50 is selected from. The test sequences are manually tagged with 9 attributes such as illumination, scale variation, and occlusion. They represent the challenging problems awaiting solutions in visual tracking. Despite of its simplicity, our method outperforms most of the other trackers, ranking among the top. We change the amount of the training data to find out if the scale of the training data has some influence. We decrease and increase the size of the training data by 1/4 and a similar result is obtained. It's beneficial to train the MD-CNN by transfer learning with salience information. To verify the effectiveness of salience detection layers we can directly compare our results with the MD-CNN. For the optimum effectiveness of the salience detection layers, we conduct the ablation experiments on the salience extraction layers. We change the layers transferred from the FCNs to optimize the tracking results. In order to eliminate the interference of the network structure, we keep the amount and the type of the new MD-CNNs' layers constant. The test shown in Fig. 6 on OTB50 suggests that we use 3 convolutional layers of FCNs for optimal performance.

The experiment of the proposed algorithm shows that salience detection improves the representing ability for the feature of MD-CNNs in visual tracking. We compare our algorithm to other classic state-of-the-art tracking algorithms such as TLD[31], Struck[32], DSST[33] on OTB50. We compare our algorithm with other trackers: ECO[23], CFNet[34], DCFNet[35], MEEM[36], STAPLE[37], LCT[38], KCF[39], and SANet[40] on OTB100 Afterwards. On UAV123 ADNet[41] and DCF[42] are tested for comparison additionally. The comparisons with other trackers are shown in Figs. 7–9. The straightforward comparison with MD-CNN shows a definite improvement by salience detection layers. Our algorithm successfully tracks the target stably even with some difficulties that make other trackers fail, like fast motion and occlusion. In the experiments, some occlusions make other trackers lose the target meanwhile ours keep tracking in the video. The stability is also guaranteed when the tracker is faced with fast motion of the target. Salience detection makes the tracker able to tackle with some adverse conditions and re-detect the target blocked by obstacles. In some cases, salience information plays an important role to prevent interference from the background. In the results of OTB100, the dominant tracker is ECO and it takes advantage of the correlation filter and deep learning. The deep learning is used to extract features, and the filter picks up the relatively effective features in classification. The performance of SANet surpasses our tracker in success plots, which

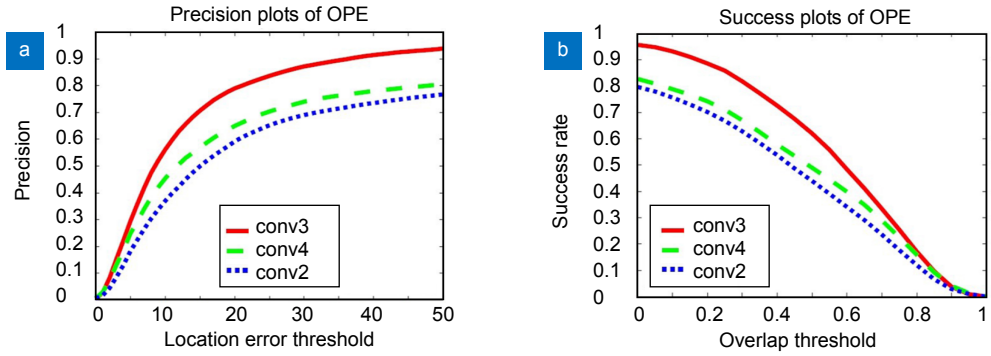

Fig. 6 | The Precision plots and the success plots on the OTB50 dataset. We compare the results with the change of the salience extraction layers and find three layers best in them.

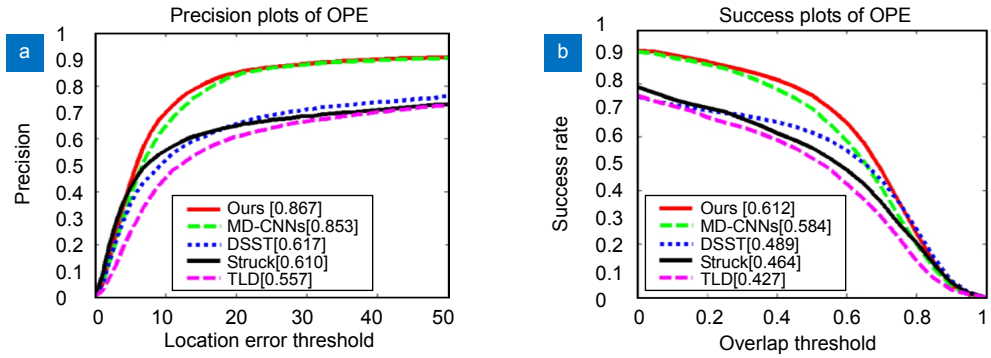

Fig. 7 | The Precision plots and the success plots on the OTB50 dataset.

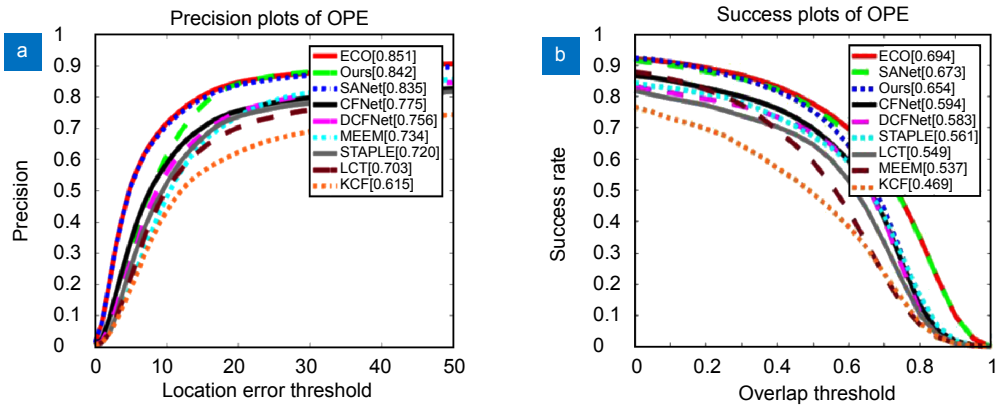

Fig. 8 | Comparison with the state-of-the-art methods on the OTB100 dataset.

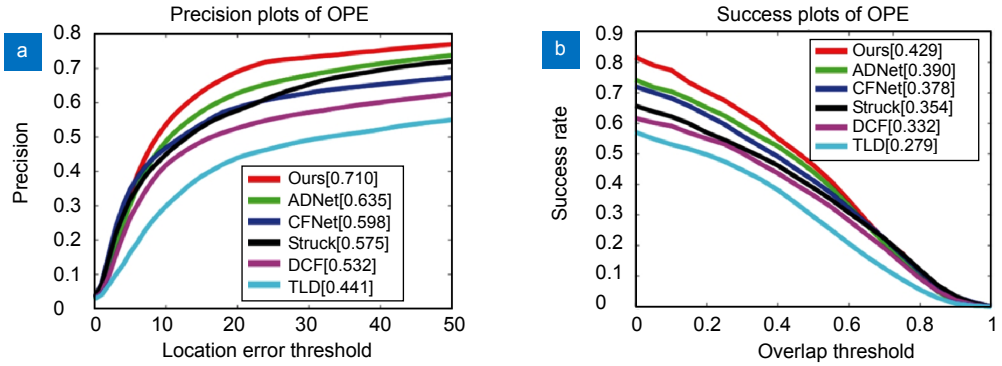

Fig. 9 | Comparison among the proposed method and several deep-learning methods and traditional methods on UAV123.

shows the strong ability in exploiting the temporal information. The SANet is made up of CNN and RNN layers and adopts many techniques in MD-CNNs. It is also a combined model like ours, which seems to have become a trend in tracking.

We also compare the accuracy and the speed of our proposed method with that of other typical trackers on VOT15 datasets. For VOT15, three main measures used to analyze tracking performance are accuracy, failures and the overlap. The overlap is the proportion of the overlapping area in both object area and predicted area. If the overlap is 0, the sequence is treated as a failure. The accuracy is used to evaluate the success of the trackers. Additionally, we evaluate their speed to see if they meet real-time requirements. The structure with shallow layers takes effects on its accuracy and saves time for forward and backward propagation of the network during training. We test these algorithms under the same hardware environment. We compare our tracker with other trackers like SO-DLT[20] and DeepSRDCF[43] which adopt deep learning methods with convolution neural networks to obtain features. The tracking results are shown in the Table 1. It demonstrates that our performance is decent among all trackers.

In Fig. 10 we show some practical tracking examples where our tracker generates more accurate positions than others. Though there are severe inferences in the tracking environments, our tracker shows enough robustness to tackle with the disadvantages that other trackers are vulnerable to. It is demonstrated that our tracker is able to

Table 1 | The testing results of our proposed method and some typical trackers on the VOT15 challenge.

Tracker	Accuracy	Failures	Overlap	Speed (fps)
Struck[32]	0.4129	103	0.2014	2
DLT[6]	0.4345	113	0.2152	0.5
SO-DLT[20]	0.5086	117	0.2006	6
DeepSRDCF[43]	0.5216	64	0.2931	0.3
SiameseFC[44]	0.4931	95	0.2307	35
MD-CNNs[1]	0.5543	49	0.3488	1
Ours	0.5592	55	0.3694	1

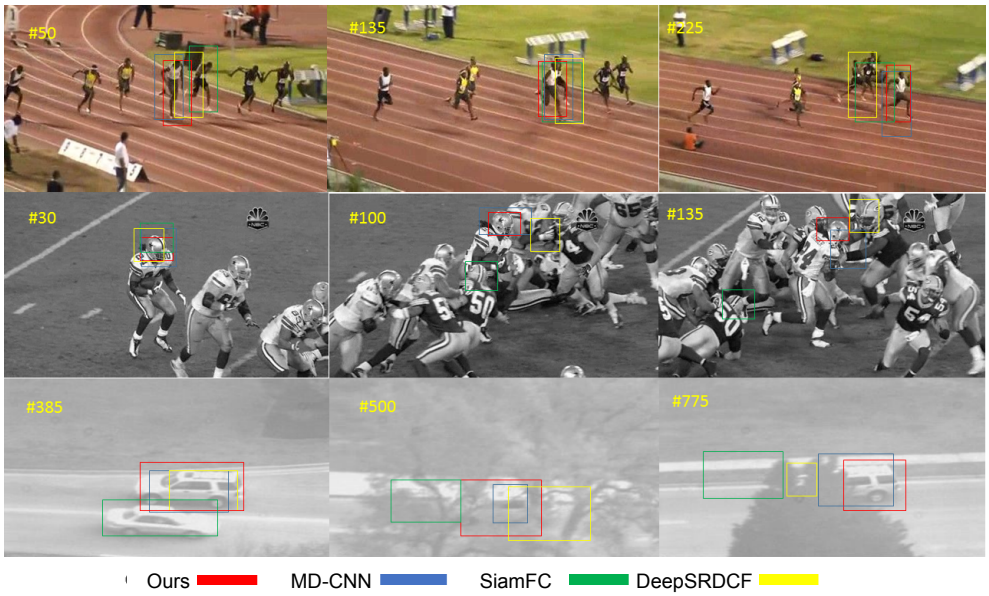

Fig. 10 | The tracking examples where our proposed algorithm is compared with other trackers. As illustrated by the challenging videos, other trackers are vulnerable and sensitive to inferences caused by generic objects and background clutter while our tracker recognizes the target in the most difficult cases.

extract useful features for intra-class classifications and exclude the background inferences despite of the exploitation of salience information. After all, the transfer learning of salience information is conducted in the initial phase of the network. Although we put emphasis on the salience information, it's not the leading factor of our tracking method.

Conclusions

In this paper, an effective algorithm for visual object tracking is proposed and it further enhances the feature exploiting ability of CNN. In consideration of the problems in visual tracking, we make the trackers use the prior information by transfer learning in the sequence to distinguish the target. We combine the salience detection and deep learning in the light of their biological characteristics. We make the data processing of CNN match the information processing by human visual perception better in our methods. We change the structure of the original MD-CNNs and propose a new model. In certain aspects, we improve the original MD-CNNs and make it robust in spite of interference under certain circumstances. The final results manifest that our network provides rich features, and allows for object tracking in practical situations. We believe that our algorithm is complementary to tracking methodologies based on neuroscience principles of deep learning. Attention mechanism will present great potential by the investigation of biometric vision characteristics. Further study by integrating more information from vision attention mechanism will make the approach straightforward and better.

References

1. Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* 4293–4302 (IEEE, 2016); http://doi.org/10.1109/CVPR.2016.465.
2. Yang M H, Lin R S, Lim J, Ross D. Adaptive discriminative generative model and application to visual tracking: US, 7369682. 2008.
3. Liu S, Zhang T Z, Cao X C, Xu C S. Structural correlation filter for robust visual tracking. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* 4312–4320 (IEEE, 2016); http://doi.org/10.1109/CVPR.2016.467.
4. Ko B C, Kwak J Y, Nam J Y. Human tracking in thermal images using adaptive particle filters with online random forest learning. *Opt Eng* **52**, 113105 (2013).
5. Li X, Dick A, Wang H Z, Shen C H, Van Der Hengel A. Graph mode-based contextual kernels for robust SVM tracking. In *Proceedings of 2011 International Conference on Computer Vision* 1156–1163 (IEEE, 2011); http://doi.org/10.1109/ICCV.2011.6126364.
6. Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* 809–817 (Curran Associates Inc, 2013).
7. Cui Z, Xiao S T, Feng J S, Yan S C. Recurrently target-attending tracking. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* 1449–1458 (IEEE, 2016); http://doi.org/10.1109/CVPR.2016.161.
8. Wang L J, Ouyang W L, Wang X G, Lu H C. STCT: sequentially

training convolutional networks for visual tracking. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* 1373–1381 (IEEE, 2016); http://doi.org/10.1109/CVPR.2016.153.

9. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems* 1097–1105 (Curran Associates Inc, 2012).
10. Hou X D, Zhang L Q. Saliency detection: a spectral residual approach. In *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition* 1–8 (IEEE, 2007); http://doi.org/10.1109/CVPR.2007.383267.
11. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition* 580–587 (IEEE, 2014); http://doi.org/10.1109/CVPR.2014.81.
12. Yi Y, Su L, Huang Q M, Wu Z, Wang C F. Saliency detection with two-level fully convolutional networks. In *Proceedings of 2017 IEEE International Conference on Multimedia and Expo* 271–276 (IEEE, 2017); http://doi.org/10.1109/ICME.2017.8019309.
13. Zhang L H, Ai J W, Jiang B W, Lu H C, Li X K. Saliency Detection via Absorbing Markov Chain with Learnt Transition Probability. *IEEE Transactions on image processing*: a Publication of the IEEE Signal Processing Society. 27 (**2**), 987-998 (IEEE, 2018)
14. Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned salient region detection. In *Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition* 1597–1604 (IEEE, 2009); http://doi.org/10.1109/CVPR.2009.5206596.
15. Cheng M M, Zhang G X, Mitra N J, Huang X L, Hu S M, Global contrast based salient region detection. In *Proceedings of CVPR 2011* 409–416 (IEEE, 2011); http://doi.org/10.1109/CVPR.2011.5995344.
16. Zhao R, Ouyang W L, Li H S, Wang X G. Saliency detection by multi-context deep learning. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition* 1265–1274 (IEEE, 2015); http://doi.org/10.1109/CVPR.2015.7298731.
17. Wang L Z, Wang L J, Lu H C, Zhang P P, Ruan X. Saliency detection with recurrent fully convolutional networks. In *Proceedings of the 14th European Conference on Computer Vision* 825–841 (Springer, 2016); http://doi.org/10.1007/978-3-319-46493-0_50.
18. Liu N, Han J W. DHSnet: deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 678–686 (IEEE, 2016); http://doi.org/10.1109/CVPR.2016.80.
19. Li X, Zhao L M, Wei L N, Yang M H, Wu F *et al*. DeepSaliency: multi-task deep neural network model for salient object detection. *IEEE Trans Image Process* **25**, 3919–3930 (2016).
20. Wang N Y, Li S Y, Gupta A, Yeung D Y. Transferring rich feature hierarchies for robust visual tracking. In *Proceedings of 2015 Conference on Computer Vision and Pattern Recognition* (2015).
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014).
22. Danelljan M, Robinson A, Khan F S, Felsberg M. Beyond correlation filters: learning continuous convolution operators for visual tracking. In *Proceedings of the 14th European Conference on Computer Vision 2016* 472–488 (Springer, 2016); http://doi.org/10.1007/978-3-319-46454-1_29.
23. Danelljan M, Bhat G, Khan F S, Felsberg M. ECO: efficient convolution operators for tracking. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition* 6931–6939 (IEEE, 2016); http://doi.org/10.1109/CVPR.2017.733.
24. Jian M W, Lam K M, Dong J Y, Shen L L. Visual-patch-attention-aware saliency detection. *IEEE Trans Cybern* **45**, 1575–1586 (2015).
25. Fang Y M, Lin W S, Lau C T, Lee B S. A visual attention model combining top-down and bottom-up mechanisms for salient object detection. In *Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing* 1293–1296 (IEEE, 2011); http://doi.org/10.1109/ICASSP.2011.5946648.
26. Ochs P, Malik J, Brox T. Segmentation of moving objects by long term video analysis. *IEEE Trans Pattern Anal Mach Intell* **36**, 1187–1200 (2014).
27. Li F X, Kim T, Humayun A, Tsai D, Rehg J M. Video segmentation by tracking many figure-ground segments. In *Proceedings of 2013 IEEE International Conference on Computer Vision* 2192–2199 (IEEE, 2013); http://doi.org/10.1109/ICCV.2013.273.
28. Bottou L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* 177–186 (Springer, 2010); http://doi.org/10.1007/978-3-7908-2604-3_16.
29. Hoffman J, Kulis B, Darrell T, Saenko K. Discovering latent domains for multisource domain adaptation. In *Proceedings of the 12th European Conference on Computer Vision* 702–715 (Springer, 2012); http://doi.org/10.1007/978-3-642-33709-3_50
30. Wu Y, Lim J, Yang M H. Online object tracking: a benchmark. In *Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition* 2411–2418 (IEEE, 2013); http://doi.org/10.1109/CVPR.2013.312.
31. Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. *IEEE Trans Pattern Anal Mach Intell* **34**, 1409–1422 (2012).
32. Hare S, Saffari A, Torr P H S. Struck: structured output tracking with kernels. In *Proceedings of 2011 International Conference on Computer Vision* 263–270 (IEEE, 2011); http://doi.org/10.1109/ICCV.2011.6126251.
33. Danelljan M, Häger G, Khan F S, Felsberg M. Accurate scale estimation for robust visual tracking. In *Proceedings of British Machine Vision Conference* (BMVA Press, 2014).
34. Valmadre J, Bertinetto L, Henriques J, Vedaldi A, Torr P H S. End-to-end representation learning for correlation filter based tracking. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition* 5000–5008 (IEEE, 2017); http://doi.org/10.1109/CVPR.2017.531.
35. Wang Q, Gao J, Xing J L, Zhang M D, Hu W M. DCFNet: discriminant correlation filters network for visual tracking. arXiv:1704.04057 (2017).
36. Zhang J M, Ma S G, Sclaroff S. MEEM: robust tracking via multiple experts using entropy minimization. In *Proceedings of the 13th European Conference on Computer Vision* 188–203 (Springer, 2014); http://doi.org/10.1007/978-3-319-10599-4_13.
37. Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr P H S. Staple: complementary learners for real-time tracking. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* 1404–1409 (IEEE, 2016);

http://doi.org/10.1109/CVPR.2016.156.
38. Ma C, Yang X K, Zhang C Y, Yang M H. Long-term correlation tracking. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition* 5388–5396 (IEEE, 2015); http://doi.org/10.1109/CVPR.2015.7299177.
39. Henriques J F, Caseiro R, Martins P, Batista J. High-speed tracking with kernelized correlation filters. *IEEE Trans Pattern Anal Mach Intell* **37**, 583–596 (2015).
40. Fan H, Ling H B. SANet: structure-aware network for visual tracking. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops* 2217–2224 (IEEE, 2017); http://doi.org/10.1109/CVPRW.2017.275.
41. Yun S, Choi J, Yoo Y, Yun K, Choi J Y. Action-decision networks for visual tracking with deep reinforcement learning. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition* 1349–1358 (IEEE, 2017); http://doi.org/10.1109/CVPR.2017.148.
42. Danelljan M, Häger G, Khan F S, Felsberg M. Learning spatially regularized correlation filters for visual tracking. In *Proceedings of 2015 IEEE International Conference on Computer Vision* 4310–4318 (IEEE, 2015); http://doi.org/10.1109/ICCV.2015.490.
43. Danelljan M, Häger G, Khan F S, Felsberg M. Convolutional features for correlation filter based visual tracking. In *Proceedings of 2015 IEEE International Conference on Computer Vision Workshop* 621–629 (IEEE, 2015); http://doi.org/10.1109/ICCVW.2015.84.
44. Bertinetto L, Valmadre J, Henriques J F, Vedaldi A, Torr P H S. Fully-convolutional Siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision* 850–865 (Springer, 2016); http://doi.org/10.1007/978-3-319-48881-3_56.
45. Wu H R, Xu Z Y, Zhang J L, Yan W, Ma X. Face recognition based on convolution Siamese networks. In *Proceedings of 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics* 1–5 (IEEE, 2017); http://doi.org/10.1109/CISP-BMEI.2017.8302003.
46. Chen L W, Yin Y M, Li Y, Hong M H. Multifunctional inverse sensing by spatial distribution characterization of scattering photons. *Opto-Electron Adv* **2**, 190019 (2019).
47. Wu H R, Xu Z Y, Zhang J L, Jia G. Offset-adjustable deformable convolution and region proposal network for visual tracking. *IEEE Access* **7**, 85158–85168 (2019).

Acknowledgements

This work was supported by the West Light Foundation for Innovative Talents of the Chinese Academy of Sciences (CAS) (No.YA18K001). This work is done in the Signal Processing Laboratory, Institute of Optics and Electronics, CAS. We express our thanks for the experiment equipment provided by the lab. We appreciate the support of the relevant department.

Competing interests

The authors declare no competing financial interests.