Robust Visual Tracking Combining Correlative Features and Maximum Overlap

Yinghong Xie, Xiaowei Han, Jie Shen and Chengdong Wu

***Abstract*—In tracking applications, the target may suffer drastic deformation in pose and viewpoint, which brings more difficulty for predicting the bounding box. In this paper, we break the traditional classifier based tracking method and divide the tracking method into two branches: the semantic information branch which is used to describe the semantic features of the target, and the IoU (intersection-over-union) branch which is used to describe the space state information of the target. For the feature extracted from the deepest layer containing most semantic information, firstly, the convolutional features extracted from the deepest layer are used to design semantic model, and they are input into the correlative filter to get the correlative scores of each pair of image patch. According to the scores, the coarse locations of the target are gained. Secondly, we estimate the location state of the target using the features from the first and second convolutional layers, because the lower the features are extracted the more the location state information is contained. By calculating the weighted IoU scores for each the bounding box gained from the semantic branch, we gain the final tracking bounding box. And the two branches are trained separately. Extensive experiments results illustrates that our tracker out-performs all the compared trackers.**

***Index Terms*—Object tracking, convolutional neural network, correlative features, maximum overlap, IoU.**

I. INTRODUCTION

THE main task of video object tracking is to estimate the target location in all the continuous frames accurately. It has a great number of applications in surveillance, augmented reality and autonomous system. Many traditional algorithms have made great progress， such as IVT[1], SCM[2],TLD[3], STRUCK[4], MIL[5], APGL1[6]. However, due to the extremely complex application environment, the performance of these algorithms can＇t meet the requirments of various applications.

It is meaningful to distinguish object classification and object estimation in the research field of visual tracking. Object classification is basically to determine the existence of the target in an image location. However, only part of the information about the object status is obtained, for example, the image coordinates. The object estimation aims to find the full state which is not only including the coordinates but also including a bounding box[7],[8],[9],[10]of the target. The aim of tracking is to compute and gain the bounding box that fits the object best. For a rigid object with simple movement, the object classification algorithm is often used to complete the tracking task. But in tracking applications, the object may suffer drastic deformation in pose and viewpoint, which brings more difficulty for predicting the bounding box.

In fact, the two important sub-tasks for designing a robust tracker are semantic representation modeling task and space information representation modeling task. A strong expressive semantic model can describe the high-level semantic information of the target stably, even when the target appearance changes obviously. While a robust spatial model can locate the target more accurately even in the complex environment. In this paper, we break the traditional classifier based tracking method and divide the tracking method into two branches: the semantic information branch which is used to describe the semantic features of the target, and the IoU branch which is used to describe the space state information of the target.

For the feature map from a deeper convolutional layer that corresponds to a larger receptive field, it can be understood that CNN does feature extraction of the image from a more global perspective. Therefore, the outputs of the last convolutional layer encode the highest semantic information of the target and such representations are robust to significant appearance variations[11]. Meanwhile, correlative filters have been widely applied for visual tracking algorithm because of its high computational efficiency with Fourier transformation. Based on the above, we design the semantic branch of the tracker, in which the convolutional features extracted from the deepest layer are used to design semantic model, and they are input into the correlative filter to get the correlative scores of each pair of image patch. According to scores, the coarse locations of the target are gained.

For IoU-Net (intersection-over-union) estimation, paper[12] designed an accurate object detection method by applying IoU-Net learning to predict the IoU between each detecting bounding box and the ground-truth, and paper[10] built a

This work is supported by the National Science Foundation of China under Grant [No.61503274], [No. 71601126].

Yinghong Xie is with College of Information Science and Engineering， Shenyang University, Shenyang 110044, China(e-mail: xieyinghong@ 163.com).
Xiaowei Han is with the Institute of scientific and technological innovation, he the corresponding author(hxw69@163.com)
Jie Shen is with the College of Engineering & Computer Science, University of Michigan Dearborn, MI, 48128, USA. (shen@umich.edu)
Chengdong Wu is with College of faculty of Robot science and engineer Northeastern University, Shenyang, 110819, China(wuchengdong@mail.neu.edu.cn)

tracking network consisting of estimation branch and classification branch, and use the IoU-Net as the part of estimation Network. Being inspirited by the two methods, our tracker design an IoU-Net to refine the final tracking result. We estimate the location state of the target using the features from the first and second convolutional layers, because the lower the features are extracted from, the more the location state information is contained. By calculating the weighted IoU scores for each the bounding box gained from the semantic branch, we gain the final tracking bounding box.

The main contributions of the proposed tracker include:

(1) Highest convolutional features is applied as the semantic branch input, which is robust to appearance change of the target.

(2) Unlike the traditional classification based tracking methods, we use the IoU network to refine bounding box, which contains the more space state information of the target.

(3) A hybrid strategy offers complementary benefits. Semantic branch gained the coarse location of the target, and the Iou-Network branch refine the bounding box.

(4) The two branches are trained off line, which prompt the tracking speed..

II. Related Work

CNN Tracking: In recent years, CNNs are outstanding in solving the problem of target recognition. Paper [13] learns a deep compact representation for visual tracking. Paper [14] adapts in two layers of deep features learning module to include the appearance information of specific target. Paper [15] utilizes a large image set with ground-truth values to extract appearance representation from CNN. Paper [16] represents the appearance by CNNs and manages the appearance models in a tree for tracking. And trackers [17][18] utilize CNNs as classifiers and take advantage of end-to-end training. Paper [19] uses the implicit interpolation model to solve the learning problem in the continuous space domain. The formula can effectively integrate multi-resolution deep features map. And papers [20][21][22] integrate deep features into traditional tracking algorithms, being benefit from the expression ability of CNN features. As it is shown in [23], the features from deepest convolutional layer have more semantics information and less space information, we make full use of the CNN features of the deepest layer to describe semantic information.

Correlative filtering. Correlation filters have been widely applied for visual tracking algorithm because of its high computational efficiency with Fourier transformation. Tracking algorithms using correlation filters[24] don't need hard-threshold samples of target appearance because they regress all the circular-shifted versions of input features to a Gaussian function. Correlation filters[25][26] also have been developed. For example, in paper [1], it designs a minimum output sum of squared error filter for the target appearance for fast visual tracking. The context learning method [27] describes the spatial-temporal relationship between the tracking objects and their local dense context in the Bayesian framework, and adopts Fast Fourier Transform (FFT) to adjust the target scale whenever it changes. The paper [28] models a scale estimation filter to estimates the target scale by learning discriminative correlation filters using a scale pyramid representation. The paper [29] designs kernelized correlation filters for training and detection by using circulant matrices. Paper [31] represents a long-term correlation tracking, and an online random classifier is trained for objects redetection. Papers [30][31] utilize multiple dimensional features for tracking. And paper [32] introduces a spatial regularization component to penalize correlation filter coefficients based on their spatial location.

III. The Proposed Tracking Algorithm

The proposed network architecture consists of two parts, the semantic features solution part and the space features solution part. For the solution of semantic features, we treat it as a similarity problem using discriminative correlation filter to get the output response, and then the location of three maximum response patches are gained. For the solution of space feature part, we gain the tracking result by computing the maximize the IoU value of the bounding box between the reference frame and the current frame using IoU network.

A. The Network Architecture

Figure1 shows the network architecture of our tracker. The inputs of the network are two image patches cropped from the previous frame and the current tracking frame of the video sequence. Let T denotes the exact target region, R denotes the target with surrounding context, S denotes the search region set. R and S have the same size $W_S \times H_S \times 3$, and the size of T is$W_T \times H_T \times 3$. W_S and $H_S$denote the width and height of the region R and S, $W_T$and $H_T$denote and region T respectively. Because the exact target region T are smaller than the target with surrounding context R, the relationship of the above variations are yand $H_S > H_T$. The search set $S = \{s_i\}$ $(i = 1..n)$, where $s_i$is the ith candidate search patch, and n is the total number of candidate search patch. The candidate search patch $s_i$has the same size as the exact target region T.The proposed network consists of two branches, the semantic branch and the location branch. The semantic branch is based on CNN network and correlative filter. And the dotted box in Figure 1 shows the semantic branch. The output of the semantic branch is part of the input of the location branch which is implemented by the IoU architecture.

The main workflow of the tracker is as follows. Firstly, the target with surrounding context R and the search region set S are input into the semantic branch, the score of correlation coefficient is computed for each. And the output of the branch is the three candidate regions with the highest scores. Then, the three candidate regions and the exact target region T are input into the location branch, for each candidate region and T, the IoU values are gained via IoU network. After refining the bounding box of each pair of input, the output is the region having the maximum IoU value. It's worth noting that the two branches are trained separately.

B. The Semantic Branch

We utilize a CNN pertained in the image classification task and keep all the parameters unchanged. And input (R, S) to this

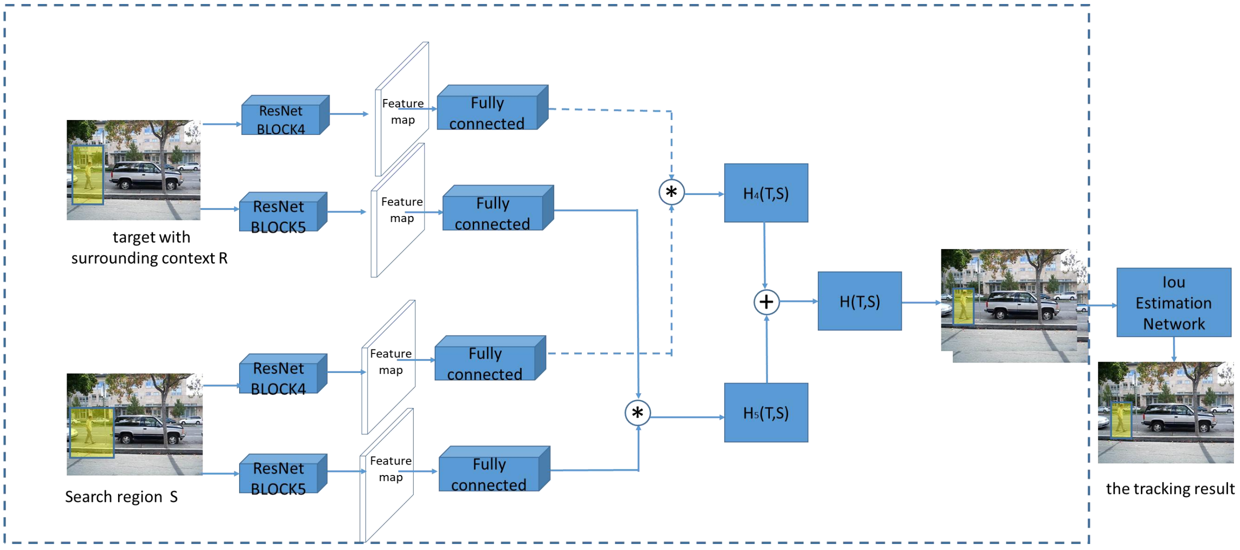

Fig 1. Schematic overview of the proposed framework. It consists of the following four stages: (1) extracting features of higher layers （2）computing correlation coefficients (3) inputting into IoU estimation Network (4) gaining the tracking result

CNN and extract the features of the last two convolutional layers as the semantic features of each image patch. Let the two features from CNN are denoted by $\boldsymbol{f_{r1}}(\cdot)$ and $\boldsymbol{f_{r2}}(\cdot)$. The response maps of the correlative filter for the last two convolutional layers are

$$h_1(R,S)=cor(f_{r1}(R),f_{r1}(S)) \tag{1}$$

and

$$h_2(R,S)=cor(f_{r2}(R),f_{r2}(S)) \tag{2}$$

where $\boldsymbol{cor}(\cdot)$ is the correlation operation. And the response map for the semantic branch is denoted as

$$h(R,S)=\lambda h_1(R,S)+(1-\lambda)h_2(R,S) \tag{3}$$

where λ is the weighting parameter used to balance the influence of two addend items on the result. The training method of all the parameters are the same as similarity learning problem. For each data pair $(\boldsymbol{R_i},\boldsymbol{s_i})$, let Y_i denotes the ground truth feature map of the corresponding search region.
The semantic branch is optimized by minimizing the logistic loss function $\boldsymbol{Los}(\cdot)$as follows:

$$\arg\min_{w_i}\frac{1}{n}\sum_{i=1}^{n}\{Los(h(R_i,S_i;\lambda;w_i),Y_i)\} \tag{4}$$

where w_i is the parameter in the network, and n is the number of training sample. Table 1 illustrates the workflow of semantic branch.

C. The IoU Estimation Branch

Because the lower layer of CNN contains more spatial information, we use the feature maps of lower CNN layer to estimate the spatial similarity between the exact target region and each candidate region gained from the semantic branch. The IoU network architecture are built based on the modulation-based network[10]. But it is different from[10] as the estimation branch has two sub-branches, which are reference sub-branch and test sub-branch.

For test sub-branch the input are the search region set $S\ =\{S_i\}\quad i=1,2,3$ and its bounding box set

For the reference sub-branch, the input are the ground truth region T and its boundingbox B_0 .The feature map of layer 1 and layer 2 are extracted from the pre-trained ResNet, then followed by Prpool and a fully connected layer. For features from layer 1, it returns an adjustment vector $a_1(T,B_0)$ ith positive coefficients of size $1\times1\times D$, and for features from layer 2 , it returns an adjustment vector $a_2(T,B_0)$ with positive coefficients of size $1\times1\times D$.

TABLE 1
The Workflow of The Semantic Branch

Input: the target with surrounding context R, the search region set S
Step1: Input R to VGG-16 network to extract feature map $f_{r1}(R)$ of the fourth convolutional layer. After fully connected layer, we get the feature map $f_{r1}(R)$. And extract feature map of the fifth convolutional layer. After fully connected layer, we get the feature map $f_{r2}(R)$. **Step 2**: The same process as step 1 is done on S. And we get the feature maps $f_{r1}(S)$ and $f_{r2}(S)$. **Step 3**. Compute the response map $h_1(R,S)$ using formula (1). **Step 4**: Compute the response map $h_2(R,S)$ using formula (2). **Step 5**: Compute the response map $h(R,S)$ for semantic branch using formula (3).
Output: The first three regions with the highest score of correlation coefficient

$B\ =\{B_i\}\quad i=1,2,3$ .For each element in S, the feature map of layer 1 and layer 2 are extracted seperately from the pre-trained ResNet, and followed by a Prpool with the bounding box estimate B. As the test sub-branch extracts general features for IoU prediction, which is a more complicate task, it uses more convolutional layers and higher pooling resolution, which is shown in figure 2. It results vectors $v_1(S_i, B_i)$ $i = 1,2,3$ with size $K_1 \times K_1 \times D$ and $v_2(S_i, B_i)$ $i = 1,2,3$ with size $K_2 \times K_2 \times D$ for layer 1 and layer 2 respectively, where K_1 and K_2 are the spatial size for the pool layer. Then the two result vectors are adjusted by the reference coefficient output vectors a_1 and a_2 via a channel-wise multiplication. Then the predicted IoU of bounding box B_i is formulated as

$$IoU(B_i) = \eta IoU_1(B_i) + (1 - \eta) IoU_2(B_i) \quad (5)$$

$$IoU_1(B_i) = g(a_1(T,B_0) \cdot v_1(S_i,B_i)) \quad (6)$$

$$IoU_2(B_i) = g(a_2(T,B_0) \cdot v_1(S_i,B_i)) \quad (7)$$

where g is the IoU predictor module consisting three fully connected layers. And η is the weighting parameter used to balance the influence of two layers on the result.

For training, we minimize the prediction error of the equation (5), and for estimating the target region, we maximize the equation (5).

And the tracking result B_{TRUTH} is defined as

$$B_{TRUTH} = \arg \max_{i=1}^{3} (IoU(B_i)) \quad (8)$$

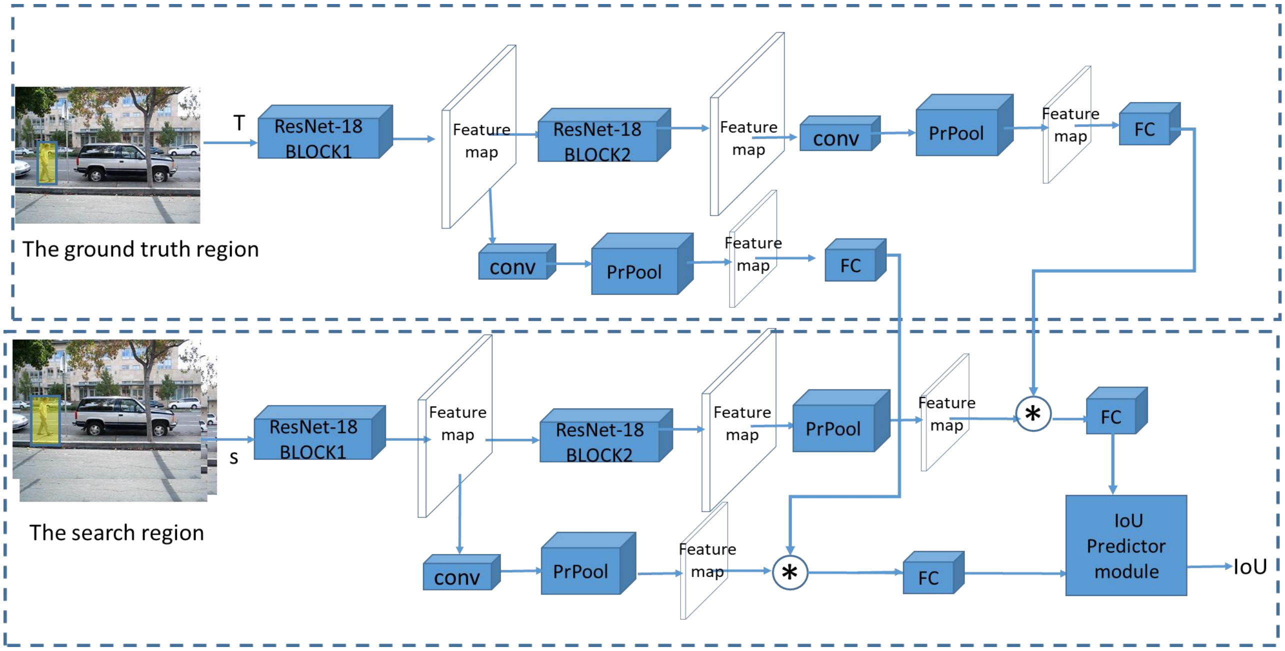

Fig. 2. The IoU estimation.

IV. DETAILS AND EXPREIMENTAL EVALUATION

A. *Implementation Details*

CNN Training: We randomly select 300 classes of ImageNet[40] as training dataset, and divide them into 80% training data set, 15% validating data set and 15% testing data set. We adopt the ResNet [57] to extract features for semantic branch and IoU estimation branch. For semantic training, we choose randomly a pair of images, and crop target with surrounding region R from one image and crop searching region S for the other image with the ground truth target in the center. For IoU estimation branch training, we use the ground truth region T for the reference image and we sample the ground truth region with some perturbation in the position and scale to simulate the tracking scenario as testing images, the sampled regions are then unified to the same size. For each target region, we crop 32 sample regions by adding Gaussian noise to the ground truth coordinates, that is, 32 candidate bounding boxes, while ensuring the minimum IoU is 0.1. Furthermore, the training epochs is set 100, and the learning rate is 0.01.

Correlation Filters Training: The regularization parameter of equation (6) is set to $\lambda = 10^{-4}$, and the kernel width is 0.1 to generate the Gaussian labels. The learning rate δ in equation (7) and (8) is set to 0.01.

We implement our tracker in TensorFlow 2.0 framework on a computer with a single NVIDA GTX 1080, an Intel Core i7 at 4.0 GHz CPU and 256GB memory. Furthermore, the parameters for each of the compared methods are set in accordance with the original definition of the respective method.

B. *Experiments on OTB Benchmarks and VOT Benchmarks*

We use the benchmark datasets OTB2015[43] and VOT2016[45] to evaluate our tracker's performance.

On OTB2015[43] benchmark, our tracker is compared with other 6 well-known trackers as SRCDFdecon[47], Staple[49], DSST[28], Struck[4], MIL[5], and CT[53]. Figure 3 illustrates the overlap success plots and distance precision plots using success of spatial robustness evaluation (SRE), success of temporal robustness evaluation (TRE) and success of one-pass evaluation (OPE). The legend of overlap success contains the area-under-curve (AUC) score while the legend of distance precision contains threshold scores at 20 pixels for each tracker. Moreover, Figure 4 shows the Distance precision plots over 9 tracking challenges of illumination variation, out-of-plane rotation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, and background clutter. And Figure 5 shows the tracking results under 4 different video sequences with various challenging attributes. All these results indicate that our tracker has better performance.

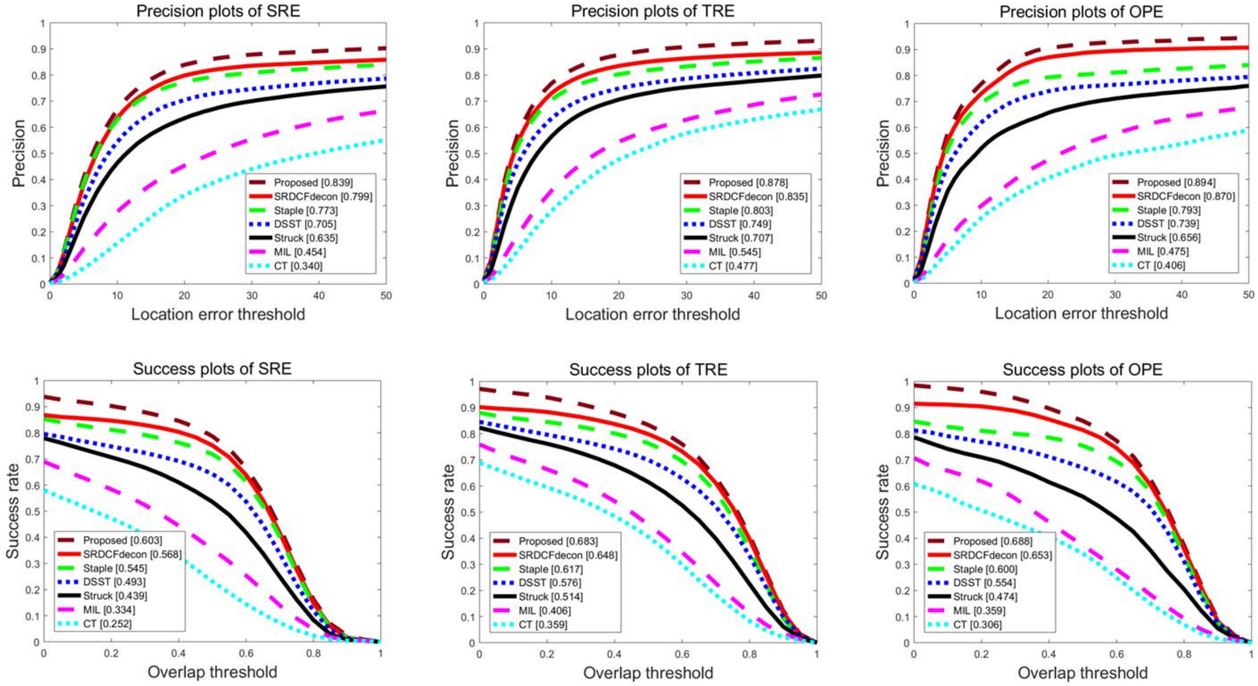

Fig. 3. Overlap success plots and distance precision plots using SRE, TRE and OPE. The legend of overlap success contains AUC score while the legend of distance precision contains threshold scores at 20 pixels for each tracker.

C. *Experiments on OTB Benchmarks and VOT Benchmarks*

We use the benchmark datasets OTB2015[43] and VOT2016[45] to evaluate our tracker's performance.

On OTB2015[43] benchmark, our tracker is compared with other 6 well-known trackers as SRCDFdecon[47], Staple[49], DSST[28], Struck[4], MIL[5], and CT[53]. Figure 3 illustrates the overlap success plots and distance precision plots using success of spatial robustness evaluation (SRE), success of temporal robustness evaluation (TRE) and success of one-pass evaluation (OPE). The legend of overlap success contains the

area-under-curve (AUC) score while the legend of distance precision contains threshold scores at 20 pixels for each tracker. Moreover, Figure 4 shows the Distance precision plots over 9 tracking challenges of illumination variation, out-of-plane rotation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, and background clutter. And Figure 5 shows the tracking results under 4 different video sequences with various challenging attributes. All these results indicate that our tracker has better performance.

On VOT2016 [45] benchmark**,** we evaluate the tracking performance by comparing with 6 popular trackers including SRDCF[47], HCF[23], EBT[50], SKCF[54], SCT4[51] and Staple[49]. Table 2 shows the comparison results under the metrics standards of expected average overlap (EAO), accuracy (A) and robustness (R). A better tracker has higher A and EAO scores and lower R scores.And Table 3 illustrates the accuracy values (A) under different challenging sequences, where the red, blue and green fonts indicate the first, second and third place respectively. On average, our tracker ranks first, which is also verified in Figure 5 and Figure 6 that is accuracy-robustness plot with best trackers closer to the top right corner. From all of the above, we can conclude that our tracker has much better performance than the compared trackers. The reason is that we design the semantic information model and spatial information model separately, with the output from the deepest CNN layers as semantic model and the result of IoU estimation as spacial model, which forms complementary advantages. Furthermore, the raw FPS is shown in table 1, the Raw FPS of our tracker is 42.3487 under speed report for experiment baseline. It is much slower than HCF tracker and SRDCF tracker, because the features extraction for two branches take up most of the running time, but the computation efficiency of correlative filter makes the tracker much faster than other compared trackers.

V. CONCLUDING REMARKS

The target may undergo dramatic changes in appearance and shape during tracking, which brings more difficulty for predicting the bounding box. In this paper, taking the advantage of the feature extracted from the deeper layer containing more semantic information, we propose a novel tracking method combining correlative features and maximum overlap. And the location of the target is determined from coarse to fine. In detail, the tracking method is divided into two branches: the semantic information branch and the IoU branch. And they are trained separately. For the semantic information branch, the feature from the deepest layer is extracted to represent the semantic feature of the target, then they are input into the correlative filter to get the correlative scores of each pair of image patch. According to the scores, some coarse locations of the target are gained. For the IoU estimation branch, the feature from the first two layers is extracted to describe the space information of the target. By calculating the weighted IoU scores for each bounding box gained from the semantic branch, we gain the final tracking bounding box exactly.

In the method, a hybrid strategy offers complementary benefits that the semantic branch gained the coarse location of the target and the Iou-Network branch refine the bounding box. Extensive experiments results on OBT benchmarks and VOT benchmarks show that our tracker out-performs all the compared trackers.

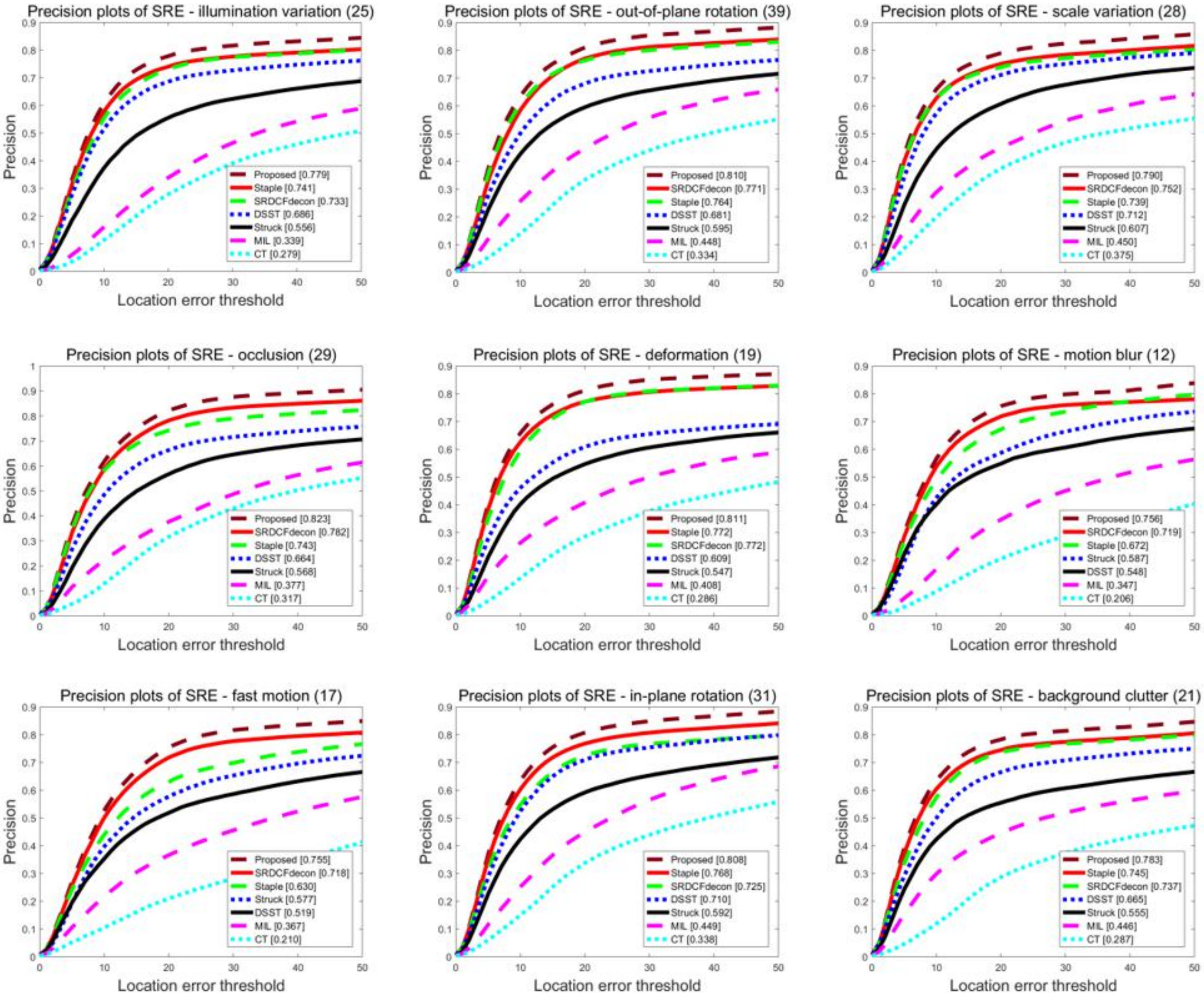

Fig. 4. Distance precision plots over 9 tracking challenges of illumination variation, out-of-plane rotation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, and background clutter.

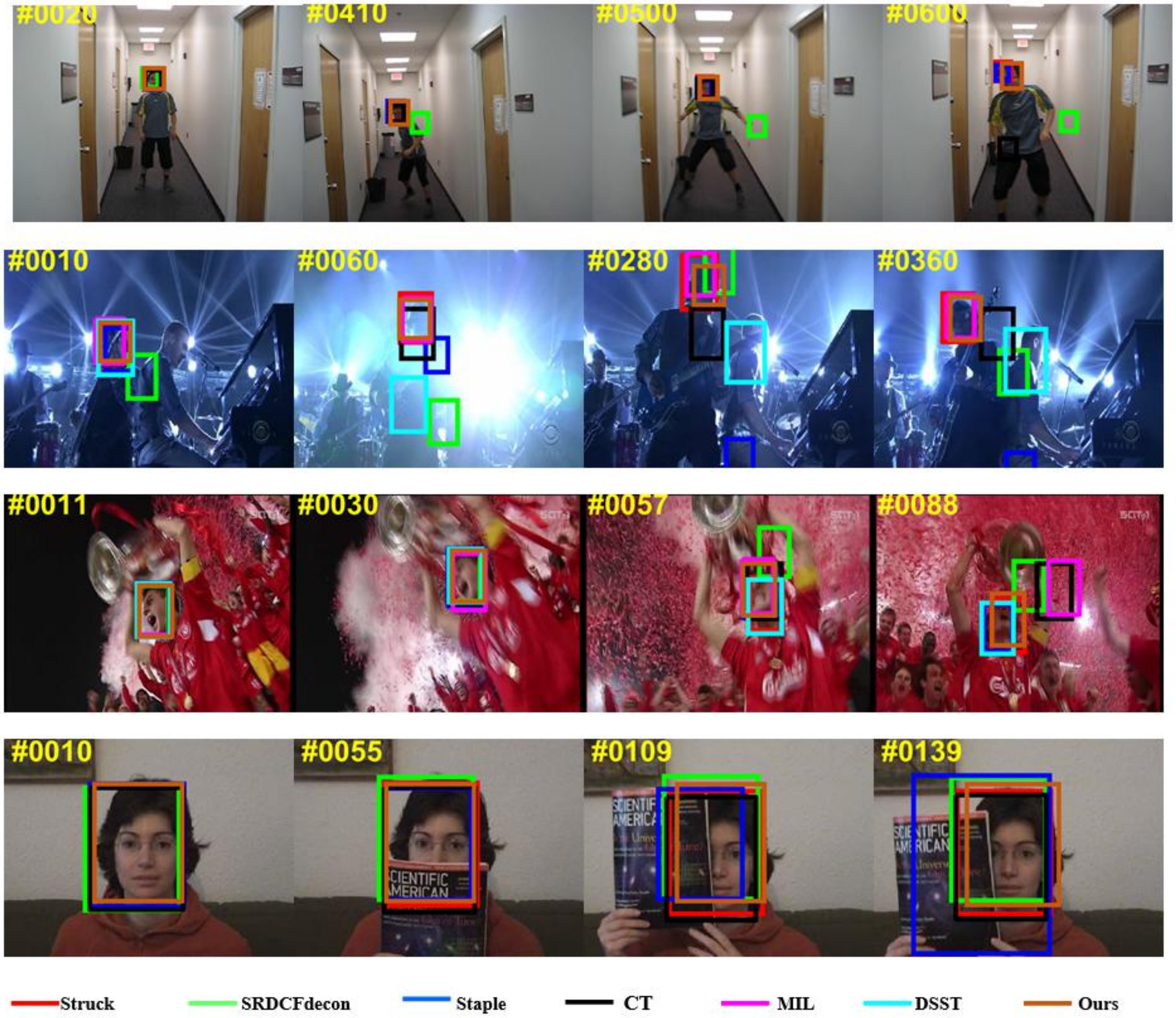

Fig. 5. Bounding box results for the proposed algorithm and the compared algorithms

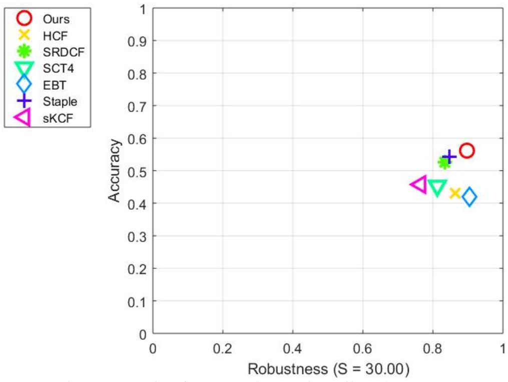

Fig. 6.AR plot for experiment baseline (mean)

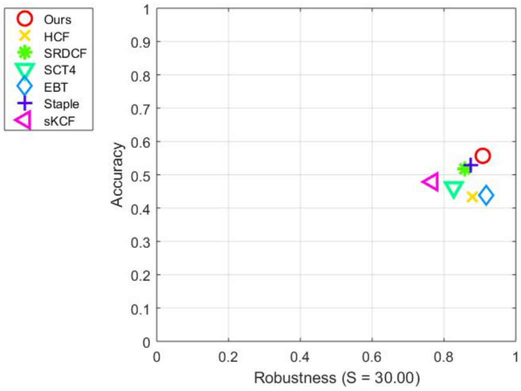

Fig. 7.AR plot for experiment baseline (weighted_mean)

TABLE 2
PERFORMANCE COMPARISON OF STATE-OF-THE-ART TRACKERS ON VOT-2016 BENCHMARKS

	Ours	HCF	SRDCF	SCT4	Staple	SKCF	EBT
EAO	0.453	0.220	0.247	0.188	0.295	0.153	0.291
A	0.5441	0.450	0.535	0.462	0.544	0.485	0.465
R	0.169	0.396	0.419	0.545	0.378	0.816	0.252
Raw FPS	42.3487	328.7264	503.1796	10.0178	14.4329	118.6575	2.8697

TABLE 3
ACCURACY VALUES UNDER DIFFERENT CHALLENGING SEQUENCES

	tag_camera_motion	tag_empty	tag_illum_change	Tag-Motion_change	tag_occlusion	tag_size_change	Mean	Weighted mean
Ours	0.5428	0.5927	0.6256	0.5021	0.4876	0.5135	0.5441	0.5414
HCF	0.4383	0.4928	0.4497	0.4255	0.4337	0.3458	0.4310	0.4336
SRDCF	0.5306	0.5745	0.6891	0.4798	0.4153	0.4662	0.5259	0.5176
SCT4	0.4748	0.5331	0.4591	0.4411	0.4451	0.3675	0.4535	0.4619
EBT	0.4767	0.4869	0.4007	0.4275	0.3777	0.3465	0.4193	0.4374
Staple	0.5284	0.5741	0.7200	0.4989	0.4311	0.5037	0.5427	0.5290
sKCF	0.5151	0.5722	0.4290	0.4332	0.4343	0.3606	0.4574	0.4966

REFERENCES

[1] D.A. Ross , J. Lim , R. Lin , M. Yang , Incremental learning for robust visual track- ing, Int. J. Comput. Vis. 77 (1–3) (2008) 125–141

[2] W. Zhong, H. Lu, and M.-H. Yang. Robust object tracking via sparse collaborative appearance model. TIP, 23(5):2356– 2368, 2014.

[3] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning detection. TPAMI, 34(7):1409–1422, 2012.

[4] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In ICCV, 2011.

[5] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. TPAMI, 2011, 33(8), 1-8

[6] Alper Yilmaz, Omar Javed, Mubarak Shah. Object tracking: A survey[J]. ACM Computing Surveys 2006. 38(4):1-45

[7] H. K. Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey. Need for speed: A benchmark for higher frame rate object tracking. In ICCV, 2017. 2, 6, 7, 8

[8] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. TPAMI, 37(9):1834–1848, 2015. 2

[9] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pfugfelder, L. C. Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey, G. Fernandez, and et al. The sixth visual object tracking vot2018 challenge results. In ECCV workshop, 2018. 1, 2, 6, 8

[10] M. Kristan, Goutam Bhat, Fahad Shahbaz Khan, Michael Felsberg. ATOM: Accurate Tracking by Overlap Maximization CVPR, 2019. 4660-4669.

[11] C. Ma, J. B. Huang, X. Yang, M. H. Yang. ‘‘Hierarchical convolutional features for visual tracking,’’ *in Proc. CVPR*, 2015, pp. 3074–3082.

[12] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang. Acquisition of localization confidence for accurate object detection. In ECCV, 2018. 2, 3, 5

[13] N. Wang and D. Yeung. Learning a deep compact image representation for visual tracking. In NIPS, 2013. 1- 8

[14] L. Wang, T. Liu, G. Wang, K. L. Chan, and Q. Yang. Video tracking using learned hierarchical features. TIP, 24(4):1424–1435, 2015. 1, 2, 3

[15] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4293–4302, 2016. 2, 7

[16] H. Nam, M. Baek, and B. Han. Modeling and propagating cnns in a tree structure for visual tracking. arXiv preprint arXiv:1608.07242, 2016. 2

[17] B. Han, J. Sim, and H. Adam. Branchout: Regularization for online ensemble tracking with convolutional neural networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017. 2, 3

[18] L. Wang, W. Ouyang, X. Wang, and H. Lu. Stct: Sequentially training convolutional networks for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1373–1381, 2016. 2

[19] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In European Conference on Computer Vision, pages 472–488. Springer, 2016. 2

[20] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Convolutional features for correlation filter based visual tracking. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 58–66, 2015. 2, 7

[21] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang. Hedged deep tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4303–4311, 2016. 2

[22] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017. 2, 7, 8

[23] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, pages 3074–3082, 2015. 2

[24] X. K. Lu, W.G. Wang, C. Ma, J. B. Shen, L. Shao, F. Porikli.. “See more, know more: Unsupervised video object segmentation with co-attention siamese networks,” *in Proc. CVPR,* 2019, pp. 3623-3632.

[25] X. K. Lu, C. Ma, B. B. Ni, X. K. Yang. “Adaptive Region Proposal with

Channel Regularization for Robust Object Tracking," in Proc. ECCV, 2018, pp. 1-17. *IEEE Trans. on Circuits and Systems for Video Technology.* pp.1-14, 2019.

[26] X. P. Dong, J. B. Shen, D. M. Wu, K. Guo, X. G. Jin, F. Porikli. "Uadruplet network with one-shot learning for fast visual online learning," *in Proc. ECCV, 2017,* pp. 1-12

[27] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang. Fast visual tracking via dense spatio-temporal context learning. In ECCV, 2014.target attending tracking. In CVPR, 2016.

[28] M. Danelljan, G. Hager, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In BMVC, 2014.

[29] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In ECCV, 2012.

[30] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Highspeed tracking with kernelized correlation filters. IEEE TPAMI, 2015.

[31] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer. Adaptive color attributes for real-time visual tracking. In CVPR, 2014.

[32] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In ICCV, 2015.

[33] T. Wakahara ; Y. Kimura ; A. Tomono Affine-invariant recognition of gray-scale characters using global affine transformation correlation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001,23(4): 84 – 395.

[34] YukihikoYamashita, ToruWakahara. Affine-transformation and 2D-projection invariant k-NN classification of handwritten characters via a new matching measure[J], Pattern Recognition. 2016,52(4): 459-470.

[35] Wu Y, Cheng J, Wang J Q. Real-time probabilistic covariance tracking with efficient model update[J]. IEEE Transactions on Image Processing, 2012,21(5): 2824-2837.

[36] Liu L, Jing D, Ding J. Adaptive Extraction of Fused Feature for Panoramic Visual Tracking[C].2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). IEEE, 2018: 21-25.

[37] Khan Z H, Gu I YH. Tracking visual and infrared objects using joint Riemannian manifold appearance and affine shape modeling[C]. IEEE Interational Conference on Computer Vision Workshops. Gothenburg, 2011:1847-1854.

[38] Khan Z H, Gu I YH. Bayesian online learning on Riemannian manifolds using a dual model with applications to video object tracking[C]. IEEE Interational Conference on Computer Vision Workshops. Gothenburg, 2011:1042-1409.

[39] B. C. Hall. Lie Groups, Lie algebras and Representations: An Elementary Introduction. Springer, 2003

[40] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 3, 5

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. ICLR, 2015. 3, 4, 5

[42] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In CVPR, 2013.

[43] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9):1834–1848, 2015.

[44] M. Kristan and et al. The visual object tracking vot2015 challenge results. In Proceedings of the IEEE international conference on computer vision workshops, 2015. 6

[45] M. Kristan and et al. The visual object tracking vot2016 challenge results. In European Conference on Computer Vision Workshop, 2016. 6, 7

[46] M. Kristan and et al. The visual object tracking vot2017 challenge results. In Proceedings of the IEEE international conference on computer vision workshops, 2017. 6

[47] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In CVPR, 2016.

[48] M. Wang, Y. Liu, and Z. Huang. Large margin object tracking with circulant feature maps. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017. 7

[49] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. Torr. Staple: Complementary learners for real-time tracking. In CVPR, 2016.

[50] G. Zhu, F. Porikli, and H. Li. Beyond local search: Tracking objects everywhere with instance-specific proposals. In CVPR, 2016.

[51] Jongwon Choi, Hyung Jin Chang, Jiyeoup Jeong, Yiannis Demiris, and Jin Young Choi. Visual Tracking Using Attention-Modulated Disintegration and Integration. CVPR (2016)

[52] V. N. Boddeti, T. Kanade, and B. V. K. V. Kumar. Correlation filters for object alignment. In CVPR, 2013. 4, 5

[53] K. Zhang, L. Zhang, and M.-H. Yang. Real-time Compressive Tracking. In ECCV, 2012.

[54] Montero A S , Lang J , Laganiere R . Scalable Kernel Correlation Filter with Sparse Feature Integration[C]// IEEE International Conference on Computer Vision Workshop. IEEE Computer Society, 2015.

[55] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In CVPR, 2010.

[56] C. Ma, X. Yang, C. Zhang, and M.-H. Yang. Long-term correlation tracking. In CVPR, 2015.

[57] kaiming He et al. Deep Residual Learning for Image Recognition, In CVPR, 2016.

Yinghong Xie received the Ph.D. degree in pattern recognition and artificial intelligence from Northeastern University, China, in 2014. Since2005, she has been with Shenyang University, where she is currently an Associate Professor with Information and Engineering Institute. From2014 to 2016, she was a Postdoctoral Researcher with Tianjin University. She was a Scholar with the University of Michigan–Dearborn, in 2017. She is the first author of more than 20 articles, and the Host Natural Science Foundation of China, in 2015. Her main research interests include artificial intelligence, video image processing, and pattern recognition.

Xiaowei Han received the Ph.D. degree in control theory and control engineering from Northeastern University, in 2005. He is currently a Professor and the President of Scientific and Technological Innovation Institute, Shenyang University. He has presided over or undertaken more than ten research projects supported by national, provincial and municipal funds, completed a number of horizontal engineering projects, compiled two monographs, published more than 40 articles, and obtained more than 50 invention patents, utility model patents. His current research interests include computer vision, artificial intelligence, and wireless sensor networks.

Jie Shen is the editor-in-chief of International Journal of Modelling and Simulation, which is an EI-indexed, peer-reviewed research journal in the field of modelling and simulation. I also served as an editorial board member for two international journals; an organizer for 8 international conferences; an associate editor of 2 international conference proceedings; a program committee member for 20 international conferences; a session chair for 13 international or national conferences; a board member for 3 international- or national-level technical committees; and a member for various committees at department and campus levels within the University of Michigan - Dearborn.

Dr. Shen Author's awards and honors include the Frew Fellowship (Australian Academy of Science), the I. I. Rabi Prize (APS), the European Frequency and Time Forum Award, the Carl Zeiss Research Award, the William F. Meggers Award and the Adolph Lomb Medal (OSA).

Chendong Wu is currently the Vice President of the Faculty of Robot Science and Engineering in Northeastern University, Shenyang, China, where he is also the Director of the Institute Artificial Intelligence, a Professor, and the Doctoral Tutor.

He has long been involved in automation engineering, artificial intelligence, and teaching and researching in robot navigation. He is an Expert of Chinese modern artificial intelligence and robot navigation, and he is also a Special Allowance of the State Council.