International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.10, (2016), pp.195-204
http://dx.doi.org/10.14257/ijsip.2016.9.10.19

Robust Visual Tracking Integrating Spatio-Temporal Model

Min Jiang[1*], Jiao Wu[1], Jun Kong[1,2], Chenhua Liu[1] and Shengwei Tian[2]

[1]*Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China*
[2]*College of Electrical Engineering, Xinjiang University, Urumqi, 830047, China*
minjiang@jiangnan.edu.cn

Abstract

Recently, the compressive tracking (CT) method has attracted much attention due to its high efficiency. However, the CT extracts samples around the previous target region within a fixed search radius; the searching area is unsuitable when the target undergoes abrupt acceleration change. Meanwhile, the classifier learns the features of the target online without judgment even the target is fully occluded. Thus, the improper searching area and incorrectly updated features lead to a marked drop in precision of tracking. To solve this issue, a robust target tracking method integrating spatio-temporal model to constrain the searching area is proposed in this paper. Different from CT, the proposed method initially constructs the spatio-temporal model to calculate a confidence map between consecutive frames, and the region with high confidence suggests the high possibility that target exists. Thus the samples can be extracted in the high confidence area. Then, the optimal target location can be estimated with a naive Bayes classifier using sparse coding features. Experiments show that the proposed method outperforms several competing methods in efficiency and robustness.

Keywords: *Spatio-Temporal model, Confidence map, Searching constraints, Robust target tracking*

1. Introduction

Visual tracking becomes one of the hot issues in the computer vision field due to the wide applications including motion classification, visual surveillance, and activity analysis, etc. The challenge for visual tracking is the difficulty to deal with complex appearance changes of the target object [1-2] including occlusion, illumination changes, shape deformation and camera motion. To overcome these difficulties, numerous algorithms have been proposed. Most of them pay close attention to build an effective appearance models [3-5], but ignore the relationship between the target and its surroundings context information [6].

In the real world, the target is barely independent to other objects in the entire scene. For example, nose and ears keep fixed space position in the face region. So the spatial relationship between the target and its surrounding context information is important to locate the target [6]. Meanwhile, the appearance of the target changes gradually in continuous frames and the historical appearances information usually influence the next appearance. So it is necessary to take the advantage of the temporal context in an efficient way to predict the next state of the target [7]. In this paper, a novel tracking framework is proposed based on the spatio-temporal model to precisely locate the target, which is expected to be much more robust than the compressive tracking (CT) [8] method and other state-of-art methods.

* Corresponding Author

ISSN: 2005-4254 IJSIP

2. Related Work

According to the learning strategy, tracking methods can be divided into two classes: generative learning and discriminative learning. Generative tracking [3-4] methods typically learn an effective appearance model to represent the target, and then search for the target region with minimal error. However, generative models do not take advantage of surrounding visual context and discard useful information which can be exploited to better distinguish the target from the background. Discriminative tracking [9-10] methods often treat the tracking problem as a binary classification problem, which separate the target from its local background. However, discriminative trackers need to learn numerous discriminative features to separate the target from the background, which increases the computational burden.

Recently, the compressive tracking (CT) [8] method has attracted much attention due to its high efficiency. A very sparse measurement matrix is constructed to extract the data-independent features for the appearance model. However, the CT extracts samples around the previous target region within a fixed search radius and the searching area is unsuitable when the target undergoes abrupt acceleration changes. Meanwhile, the classifier learns the features of the target online without judgment even the target is fully occluded. In this way, the searching area will be farther away from the target location once the target is lost and it is hard to locate the target precisely. The improper searching area and incorrectly updated features lead to a marked drop in precision of tracking.

Motived by the effectiveness of the contex for the tracking task [6].In this paper, we present a robust and real-time tracking algorithm. First we learn a spatio-temporal model between the continuous frames. Then, a confidence map is calculated based on the model, and the region with high confidence is regarded as the searching area of the target. Similar to CT, a sliding window is used to detect samples in the searching area. The samples are represented by compressive features using a sparse measurement matrix and classified with naive Bayes. Finally the location of the sample with the maximum likelihood will be chosen as the new location of the target.

The key contributions of the proposed method are summarized as follows:

1 The spatio-temporal model is introduced in our paper to capture the spatio-temporal information, which can be futher used to prevent the target from drifting to the background.

2 The Fast Fourier Transform is adopted to learn the spatio-temporal model and a very sparse measurement matrix is adopted to efficiently extract the features to represent the samples which are detected in the searching area. Therefore, our algorithm can meet the requirements of the real time tracking.

3. Compressive Tracking

3.1. Compressive Sensing

In the compressive tracking, samples are drawn around the former target region and each sample is represented by a high-dimensional vector via convolving each patch with some rectangle filters. CT employs a very sparse random matrix $q\hat{I}\ ¡^{b´d}$ to project the high-dimensional image feature $I\hat{I}\ ¡^{d´1}$ onto a lower-dimensional vector $E\hat{I}\ ¡^{b´1}$

$$E = qI \tag{1}$$

where $b = d$. The elements of q satisfied

$$q_{i,j} = \sqrt{s} \times \begin{cases} 1 & \text{with probability } 1/2s \\ 0 & \text{with probability } 1 - 1/2s \\ -1 & \text{with probability } 1/2s \end{cases} \quad (2)$$

where $s = 2$ or 3.

3.2. Online Classifier Update

Assumed all elements in E are independently distributed. The naive Bayes classifier is adopted

$$H(e) = \log \frac{\prod_{i=1}^{b} p(e_i|y=1)p(y=1)}{\prod_{i=1}^{b} p(e_i|y=0)p(y=0)} = \sum_{i=1}^{b} \log \frac{\prod_{i=1}^{b} p(e_i|y=1)}{\prod_{i=1}^{b} p(e_i|y=0)} \quad (3)$$

where $p(y=1)=p(y=0)$ and $y \in \{0,1\}$ represents the sample label.

The conditional distributions $p(e_i \mid y=1)$ and $p(e_i \mid y=0)$ follow to the Gaussian distributed with four parameters $(\mu_i^+, \sigma_i^+, \mu_i^-, \sigma_i^-)$

$$p(e_i \mid y=1) : N(\mu_i^+, \sigma_i^+), \; p(e_i \mid y=0) : N(\mu_i^-, \sigma_i^-) \quad (4)$$

where (μ_i^+, σ_i^+) and (μ_i^-, σ_i^-) are mean and standard deviation of the positive and negative respectively. The parameters in μ_i^+ and σ_i^+ are incrementally updated by

$$\mu_i^+ \leftarrow \lambda \mu_i^+ + (1-\lambda)\mu^+, \sigma_i^+ \leftarrow \sqrt{\lambda(\sigma_i^+)^2 + (1-\lambda)(\sigma^+)^2 + \lambda(1-\lambda)(\mu_i^+ - \mu^+)^2} \quad (5)$$

where λ is a learning parameter, $\mu^+ = \frac{1}{b}\sum_{k=0|y=1}^{b-1} e_i(k)$ and $\sigma^+ = \sqrt{\frac{1}{b}\sum_{k=0|y=1}^{b-1}(e_i(k)-\mu^+)^2}$. Parameters (μ_i^-, σ_i^-) are updated with similar rules. Finally, the region of the sample with the maximal classification response is selected as the new target region.

4. Proposed Algorithm

In this section, we present our tracking integrating spatio-temporal model in details.

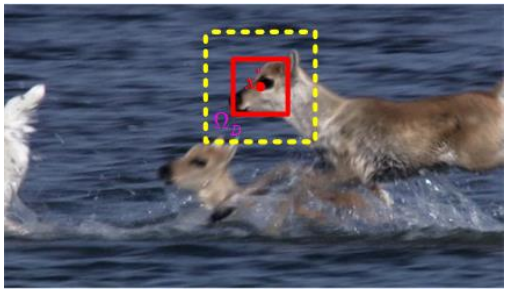

Figure 1. The Spatial Region $\Omega_D(x^*)$ is Inside the Yellow Rectangle which Includes Target Region Surrounding by the Red Rectangle Centering at the Location x^*

4.1. Spatial Relationship Model

The tracking problem is formulated to calculate a confidence map for obtaining the probability distribution of the target location:

$$G(X)=P(X|o) \tag{6}$$

where $X \hat{I}$ ¡ ^{2} is coordinates of the target location and o is the target in current scene. Then, the spatio-temporal model will be learned to estimate (6). Figure 1, illustrates the spatial relationship between the target and its surrounding context.

Assumed that the center of the target x^* is given in Figure 1, the spatial region $\mathbb{W}_D(x^*)$ is defined as twice the size of the target region centering at the location x^*, which includes the target and its local context. The spatial context feature are defined as $X^D = \{D(m)=(C(m),m)|m$? $_D(x^*)\}$, where $C(m)$ denotes image intensity at the location m in $\mathbb{W}_D(x^*)$.

According to Bayes' rule, the function in (6) can be represented by

$$\begin{aligned} G(X) &= P(X \mid o) \\ &= \sum_{D(m)\hat{I} X^D} P(X,D(m)|o) \\ &= \sum_{D(m)\hat{I} X^D} P(X|D(m),o)P(D(m)|o) \end{aligned} \tag{7}$$

Where $P(X, D(m)|o)$ is the conditional probability which models the spatial relationship between the target position and its surrounding context information. The spatial context prior probability $P(D(m)|o)$ models appearance of the spatial context.

The spatial context prior probability $P(D(m)|o)$ is defined as

$$P(D(m)|o)=C(m)\omega_\sigma(m\text{-}x^*) \tag{8}$$

and $W_S(m - x^*)$ is the Gaussian function which is defined as

$$W_S(m - x^*) = ke^{-\left|\frac{m-x^*}{S}\right|^2} \tag{9}$$

where k is a normalized constant, s is a scale parameter. The conditional probability $P(X,D(m)|o)$ represents the spatial relationship between the target position and its surrounding context information. It is modeled as

$$P(X,D(m)|o)=h^{sm}(X\text{-}m) \tag{10}$$

where $h^{sm}(X - m)$ is a function that denotes the spatial relationship between target position X and its surrounding context location m in terms of distance and direction.

In order to account the prior information of the target position, the confidence map is defined as:

$$G(X)=P(X|o) = te^{-\left|\frac{X-x^*}{a}\right|^b} \tag{11}$$

where t is a normalized constant, the scale parameter a and the shape parameter b are used to present the range of the target.

The spatial relationship model will be learned based on the spatial context prior function (8) and the confidence map function. Putting the formulas (7), (8), (10) and (11) together, we can get:

$$\begin{aligned} G(X) &= te^{-\left|\frac{X-x^*}{a}\right|^b} \\ &= \sum_{m \in \Omega_D(x^*)} h^{sm}(X-m)C(m)w_\sigma(m-x^*) \\ &= h^{sm}(X) \otimes (C(X)w_\sigma(X-x^*)) \end{aligned} \tag{12}$$

where $\otimes$ represents the convolution operator. According to the Fast Fourier Transform method (FFT), the formulas (12) can be computed in the frequency domain. So

$$\mathcal{F}(te^{-\left|\frac{X-x^*}{a}\right|^b}) = \mathcal{F}(h^{sm}(X)) \odot \mathcal{F}(C(X)w_\sigma(X-x^*)) \tag{13}$$

where $\mathcal{F}$ represents the FFT function, and $\odot$ is the element-wise product. Therefore, we learn a spatial relationship model

$$h^{sm}(X) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(te^{-\left|\frac{X-x^*}{a}\right|^b})}{\mathcal{F}(C(X)w_\sigma(X-x^*))}\right) \tag{14}$$

where $\mathcal{F}^{-1}$ represents the inverse FFT function.

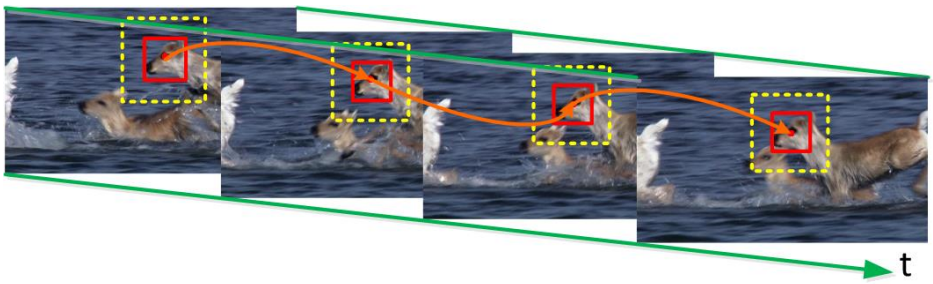

Figure 2.The Target Appearance Changes Gradually and Most of the Context Information Remain Similar in Continuous Frames. The Orange Curve Describes the Temporal Property of the Interesting Target

4.2. Spatio-Temporal Model

Due to the influence between the sequential frames (see Figure 2), the temporal context is introduced to construct the spatio-temporal model.

$$H_{t+1}^{stm} = (1-\eta)H_t^{stm} + \eta h_t^{sm} \tag{15}$$

where η is a update parameter. At the t^{th} frame the spatial context model h_t^{sm} is computed as formulas (14), which is used to update the spatio-temporal model H_{t+1}^{stm}. H_t^{stm} and H_{t+1}^{stm} are the spatio-temporal model at the t^{th} frame and the $(t+1)^{th}$ frame respectively ,which $H_2^{stm} = h_1^{sm}$.

When the $(t+1)^{th}$ frame arrives, we build the spatial region $\Omega_D(x_t^*)$ based on the location x_t^* at the $(t+1)^{th}$ frame and construct the spatial context feature set $X_{t+1}^D = \{D(m)=(C_{t+1}(m),m) | m \in \Omega_D(x_t^*)\}$.

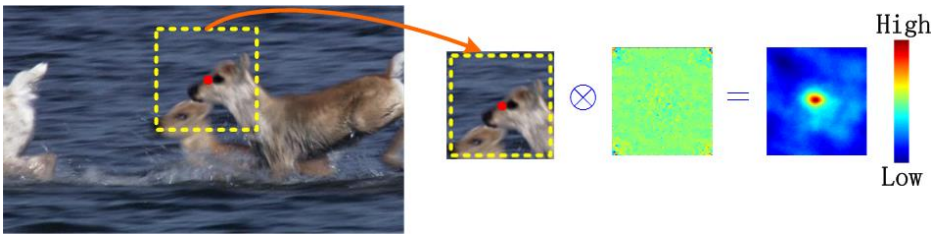

Figure 3. The Confidence Map is Constructed by the Spatio-Temporal Model. Left: The Spatial Region at the $t+1$ Frame. Middle: the Spatio-Temporal Model. Right: The Confidence Map

4.3. Update Searching Area

The confidence map is constructed by the spatio-temporal model at the $(t+1)^{th}$ frame shown in Figure 3.

$$G_{t+1}(X) = H_{t+1}^{stm}(X) \ddot{A} (C_{t+1}(X) W_{\varsigma}(X\text{-}x^{*})) \tag{16}$$

$G_{t+1}(X)$ represents the confidence value at location X in the confidence map, which indicates that the probability distribution of the target location in the spatial region.

Therefore, the region $\mathbb{W}_s$ with high confidence above threshold j_{t+1} is determined as the searching area of the target (see Figure 4).

$$\mathbb{W}_s(X) = \{X \mid G_{t+1}(X) > j_{t+1}\} \tag{17}$$

where $\varphi_{t+1} = \dfrac{3[G_{t+1}(X)]_{\max}}{4}$.

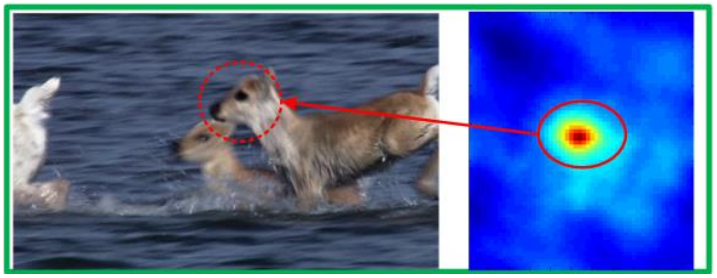

Figure 4. The Region with Confidence of High Value is Seem as the Searching Area of the Target. The Samples are Detected in the Searching Area, which will be Classified using Naive Bayes Classifier

Therefore, a sliding window is used to detect samples in the searching area. The samples are represented by compressive features using a sparse measurement matrix and send to the naive Bayes classifier. Finally the location of the sample with the maximum classifier score will be chosen as the new location of the target.

$$\mathrm{H}(*) = \underset{e? \; s}{max}\, \mathrm{H}(e) \tag{18}$$

Considering that the target may undergoes heavy occlusion in the tracking, we assume that the classifier will not be updated if maximum classifier response $\mathrm{H}(*) < \mathcal{G}$.

5. Experiments

In this section, we compare the proposed tracking algorithm with five competing methods on five video sequences with several challenging factors including pose changes, heavy occlusion, background cluster and motion blur. The 5 trackers we compared with were the Fragment Tracking (Frag) [3], Compressive tracker (CT) [8], the tracking-learning-detection (TLD) [9], the MIL tracking algorithm (MIL) [11], the online AdaBoost method (OAB) [12]. All the sequences are available from Benchmark [2]. Implemented in MATLAB, our tracker runs at 95 frames per second (FPS) in five test sequences on a Pentium Dual-Core 2.90GHz CPU with 4GB RAM.

5.1. Experimental Setup

In our experiments, the parameter s is set to $s = (s_w + s_h)/2$, where s_w and s_h are width and height of the initial tracking rectangle respectively. The scale parameters a and the shape parameter b in *Eq* 11 are set to 2.25 and 1. The update parameter h in *Eq*.15 is set to 0.075. The dimensionality of projected space b is set to 50 and the parameter t is set to 0.85. When the maximum classifier response is less than $g = 0$, the classifier will stop update.

Table 1. Success Rate （SR）(%)

Sequence	Ours	CT	TLD	Frag	MIL	OAB
Deer	100	4	73	21	13	96
David	97	43	97	12	23	15
Shaking	86	81	16	7	23	1
David3	89	35	11	81	68	34
woman	95	16	17	18	19	61
Average SR	93	36	43	28	29	41

5.2. Experimental Results

Table 2. Center Location Error (CLE) (in Pixels)

Sequence	Ours	CT	TLD	Frag	MIL	OAB
Deer	5	246	125	105	101	7
David	9	11	8	82	17	22
Shaking	13	19	197	192	24	192
David3	8	89	185	14	30	83
woman	12	114	107	112	125	32
Average CLE	9	95	124	101	59	67

In this work, we use two metrics to evaluate the proposed algorithm with the five competing trackers. One widely used evaluation metric is the success rate, the overlap score is defined as $score = S(R_T \cap R_G) / S(R_T \cup R_G)$, where R_T is the tracking bounding box and R_G is the ground truth bounding box. If the score of overlap is larger than 0.5 in one frame, the tracking result is considered as a successful frame. Another evaluation metric is the center location error, which is defined as the average Euclidean distance between the center locations of the tracked targets and the manually labeled ground truth data. Then the average center location error over all the frames of test sequence is used to summarize the overall performance for that sequence. In Tables 1, and Tables 2, we displayed the quantitative results. The proposed algorithm achieved outstanding performance both in terms of center location error and success rate in five test sequences. Figure 5 illustrates some tracking results of different trackers.

Figure 5. Samples of Tracking Results

Performance in Occlusion. The target in the woman sequence undergoes occlusion at #104, #138, #202, #305 in Figure 5 (a). Due to the searching region drifting after suffering local occlusion at #104, the CT method performs poorly. In the david3 sequence shown in Figure 5(c), the appearance of the target suffers heavy occlusion. The CT method lost the target after #86. In our algorithm, we use the spatio-temporal context information to constrain a searching area and the classifier stops to update when the target

suffers heavy occlusion. The experiment results show that our algorithm is highly robust against occlusion.

Performance in pose variation and illumination. The appearance of the target in David sequence changes significantly due to the target rotates in the image plane (see #119, #169 of the David sequence in Figure 5 (b)). In the shaking sequence, the target suffers both the large pose variation and the drastic illumination changes in the whole process shown in Figure 5(e). Our method achieves highest success rate in Table 1 and lower center location error in Table 2. By integrating the spatio-temporal model, the proposed method is more robust in dealing with appearance variation.

Performance in background clutter and abrupt motion. The target in the deer sequence suffers fast movements in Figure 5(d). Especially, the texture of the target is very similar to that in the background at #55. All the trackers finally lost the target, except for the proposed algorithm that achieves 100% success rate.

6. Conclusion

In this paper, a robust and fast visual tracking method is proposed based on the spatio-temporal model, which can effectively employed the context information to constrain a searching area. Moreover, a simple and effective update strategy is also introduced to the classifier when the target suffers heavy occlusion. Compared with five competing methods on five challenging sequences, the experiment results of our tracker present an excellent performance in terms of efficiency and robustness. Integrating with the technology of salient target detection, our method can be used in real life to construct an active tracking system, which can robustly and automatically detect and track the moving object in real time.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61362030, 61201429), the Project Funded by China Postdoctoral Science Foundation (2015M581720), the Science and Research Key Project of Xinjiang Uygur Autonomous Region University (XJEDU2012I08)

References

[1] S. Ojha and S. Sakhare, "Image processing techniques for object tracking in video surveillance: A survey", Proceeding of International Conference of Pervasive Computing, Pune,India, **(2015)**, pp. 1-6.
[2] W. Yi, L. Jongwoo and Y. Ming-Hsuan, "Online Object Tracking: A Benchmark", Proceeding of Computer Vision and Pattern Recognition , Portland, United states, **(2013)**, pp. 2411-2418.
[3] A. Adam, E. Rivlin and I. Shimshoni, "Robust Fragments-based Tracking using the Integral Histogram", Proceeding of Computer Vision and Pattern Recognition, New York, United states, **(2006)**, pp. 798-805.
[4] H. Zhao, X. Wang and M. Liu, "Robust Object Tracking with Occlusion Handling based on Local Sparse Representation," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 7, no. 3, **(2014)**, pp. 407-421.
[5] Y. B. Li, X. Z. Zhuang and Y. M. Liu, "UPF Tracking Method Based on Color and SIFT Features Adaptive Fusion," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 7, no. 6, **(2014)**, pp. 379-390.
[6] L. Y. Wen, C. Zhaowei, L. Zhen, Y. Dong and S. Z. Li, "Robust Online Learned Spatio-Temporal Context Model for Visual Tracking," Image Processing, IEEE Transactions on, vol. 23, no. 2, **(2014)**, pp. 785-796.
[7] K. H. Zhang, L. Zhang, Q. Liu, D. Zhang and M. H. Yang, "Fast Visual Tracking via Dense Spatio-temporal Context Learning," Proceeding of 13th European Conference on Computer Vision, Zurich, Switzerland, **(2014)**, pp. 127-141.
[8] K. H. Zhang, L. Zhang and M. H. Yang, "Real-time compressive tracking," Proceeding of the 12th European conference on Computer Vision, Florence, Italy, **(2012)**.
[9] Z. Kalal, K. Mikolajczyk and J. Matas, "Tracking-Learning-Detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 7, **(2012)**, pp. 1409-1422.

[10] D. X. Gao, J. T. Cao, Z. J. Ju and X. F. Ji, "Real Time Tracking with Sparse Prototypes," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 4, **(2015)**, pp. 279-296.
[11] B. Babenko, Y. M. Hsuan and S. Belongie, "Robust Object Tracking with Online Multiple Instance Learning," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 8, **(2011)**, pp. 1619-1632.
[12] H. Grabner, M. Grabner and H. Bischof, "Real-time Tracking via On-line Boosting," Proceeding of the British Machine Vision Conference , Edinburgh, United kingdom, **(2006)**.

Authors

Min Jiang, received her PH.D degree in computer science and technology from Institute of Plasma Physics, Chinese Academy of Sciences. She is an associate professor in the School of the Jiangnan University now. Her research is in the area of machine learning and computer vision. She has received support from the National Natural Science Foundation of China, the Technology Research Project of the Ministry of Public Security of China.

Jiao Wu, is currently a master student in the School of Internet of Things Engineering at the Jiangnan University. His primary research interest is in object tracking.

Jun Kong, received his M.S. degree in pattern recognition and intelligent system from Institute of Intelligent Machines, Chinese Academy of Sciences, and PH.D degree in electronic science and technology from Shanghai Institute of Technical Physics, Chinese Academy of Sciences. He is currently an associate professor. His research interests include computer vision, image processing, target tracking and human action recognition.

Chenhua Liu, is currently a master student in the School of Internet of Things Engineering at the Jiangnan University. His primary research interest is in object tracking.

