第50卷第8期 中国科学技术大学学报 Vol. 50, No. 8
2020年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2020
文章编号:0253-2778(2020)08-1162-08

UAV target tracking based on visual attention mechanism

LI Peng[1], ZHENG Yu[2], ZHANG Tangui[3]
(1. *AVIC Aero space System CO., LTD, Beijing* 100028, *China*;
2. *Beijing Institute of Nearspace Vehicle's Sysetems Engineering, Beijing* 100076, *China*;
3. *The Center of Nuclear and Safety of Gansu Province, Hefei* 230027, *China*)

Abstract: In recent years, the demand for small Unmanned Aerial Vehicles (UAV) in GPS-denied environment is increasingly strong. To solve the problem of multi-target recognition, we study the multi moving target recognition and location technology based on the platform of the small multi-rotor UAV. We used a method to quickly locate the region of interest based on the visual attention mechanism, and then used the machine learning algorithm to classify the region of interest to obtain the target accurately. Our method can track the specified target in the image and locate the target in real time, which the algorithm delay is about 50ms and the location error is less than 15 cm. Our solution can effectively reduce the influence of light variation, motion blur, the color analogue interference and complex background. The ground robot is used as the tracking target to test and verify the algorithm, which can achieve a better tracking effect.

Key words: unmanned Aerial Vehicle; target tracking; visual attention mechanism; machine learning
CLC number: TP18 **Document code**: A doi:10.3969/j.issn.0253-2778.2020.08.017
2010 Mathematics Subject Classification: 94B15

Citation: LI Peng, ZHENG Yu, ZHANG Tangui. UAV target tracking based on visual attention mechanism [J]. Journal of University of Science and Technology of China, 2020,50(8):1162-1169.
李 鹏,郑 宇,张谈贵. 基于视觉显著性的无人机目标跟踪[J]. 中国科学技术大学学报,2020,50(8):1162-1169.

基于视觉显著性的无人机目标跟踪

李 鹏[1],郑 宇[2],张谈贵[3]
(1.中航机载系统有限公司,北京 100028;2.北京临近空间飞行器系统工程研究所,北京 100076
3.甘肃省核与辐射安全中心,安徽合肥 230027)

摘要:近年来,小型无人机在无卫星导航条件下的使用需求日益强烈.针对多目标识别问题,提出一种基于小型多旋翼无人机平台的多运动目标识别定位技术,首先采用一种先快速定位感兴趣区域基于视觉注意机制,再用机器学习算法分析感兴趣区域获取目标的方法.实现了对图像中的指定目标进行追踪,同时还实现对目标的定位,其定位误差小于15 cm.该方法有效降低了光照变化、运动模糊、颜色类似物干扰及复杂背景等因素的影响.以地面机器人作为追踪目标进行算法测试验证,在目标消失时间较短的情况下,能够达到较好的追踪效果.

关键词:无人机;目标检测;视觉注意机制;机器学习

0 Introduction

With the rapid development of military revolution and the UAV market, the UAV target tracking technology has been closely related to national security interests and citizens' applied to work and life. UAV can not only observe the scene from the human perspective, but also make use of its flight advantages to break through the terrain restrictions to observe the target in a large range, for a long time and from multiple angles. Therefore, it can obtain reliable information in a large range at a lower cost. Moreover, due to its unique perspective, it is often used in important

Received: 2020-06-23; **Revised**: 2020-08-26
Biography: LI Peng, male, born in 1981, Master candidate. Research field: Graph processing, target locating and tracking. E-mail: lipeng201905@126.com
Corresponding author: ZHANG Tangui, Advanced Engineer. E-mail: lptangui@163.com

reference of multi-dimensional information fusion. At present, the target tracking technology of UAV mainly includes the following three methods:

0.1 Based on image segmentation and feature extraction

Image segmentation is a process of dividing an image into different regions according to different pixel attributes[1], such as color, texture, brightness and so on. The thresholding method is simple and efficient, but it is not suitable for the slightly complex scene. Feature spatial cluster, such as mean shift[2], can achieve image segmentation by grouping similar pixels, without considering image space information. The methods of region growing and region splitting and merging[3-4] divide the attributes together as a segmentation region, but the time and memory consumption of the algorithm are relatively large. The Markov random field (MRF) model can synthesize the texture, context, spatial features and prior knowledge of the image. Its main disadvantages are the complexity of mathematical expressions and high time complexity. On the basis of image segmentation, to achieve target recognition, we need to use different feature operators for different target types. For example, Fourier descriptors[5] and Hough transform features[6], etc. However, this scheme is difficult to deal with the target recognition problem with geometric changes, the complex background and partial occlusion.

0.2 Based on knowledge and model

According to the prior knowledge, the feature extraction and abstract modeling of the target are carried out, and the constraints of target recognition are constituted by them. The corresponding feature extraction and target hypothesis are carried out for the actual image, and the hypothesis is verified by the constraints of the model, and the recognition results are obtained. For example, Kalman filter[7] is applied to real-time target tracking of UAV. It is robust to target occlusion, but only suitable for linear systems. Some researchers[8-10] use particle filters to realize UAV target tracking. Compared with Kalman filter, particle filter can be used in the nonlinear system and non-Gaussian random system, but it has high complexity and large amount of calculation. Camshift algorithm[11-12] can also be used, which is simple to implement and fast in operation. However, the obtained template features are single and not suitable for a complex background. This kind of algorithm has a small amount of calculation and a fast processing speed, and is widely used in the UAV platform, but its performance needs to be improved.

0.3 Based on machine learning

The method of machine learning is to get a classifier by learning and analyzing a large number of target samples, so as to realize the correct target recognition of the input image. Chang et al[13] proposes a target recognition algorithm based on the Bayesian framework. Firstly, SIFT features are extracted from image targets and classified by clustering, then the targets are classified by the Bayesian network. In recent years, the methods of target recognition based on the convolutional neural network have been developed rapidly[14-15], such as R-CNN based on the proposed region[16] and the end-to-end Yolo (you only look once)[17] algorithms. Their recognition effect is good, but the time complexity of the algorithm is large, so it is difficult to realize real-time recognition. Dalal et al[18] proposed a target recognition algorithm based on the support vector machine. By extracting histogram of oriented gradient (HOG) features of the training samples, the support vector machine (SVM) was used to train models, achieving a good recognition effect. The above algorithms also have the problem of time complexity.

At present, the main problem of UAV target tracking is how to extract the features of the target and effective tracking in the presence of the complex background, lighting change, similar color objects and occlusion. The classical algorithm has the advantages of small amount of computation and real-time performance, but it is sensitive to noise. The recognition effect of the method based on machine learning is good, but the time complexity is large. In view of the above problems, we propose a method based on the combination of visual attention mechanism and machine learning avoiding the problem of using the sliding window to scan the whole image, which takes into account the detection effect and time cost. In addition, using the brightness and color features of the image to establish a saliency map, and then extract the region of interest can adapt to the change of illumination to a certain extent.

1 Region of interest acquisition

The target tracking technology of UAV needs to detect the target in real time under the complex and changeable background, but the image processing algorithm which can achieve high accuracy is time-consuming. We synthesize a variety of target recognition methods, and meet the requirements of real-time and accuracy at the same time. Firstly, the algorithm of visual attention mechanism with less time complexity is used to process the whole image to obtain the region of interest, and then machine learning is

used to detect the region of interest to extract the target accurately in real time.

1.1 Modeling of visual attention mechanism

The important theoretical basis of the experimental method in this paper is the visual attention mechanism, which decomposes and simplifies the external complex visual scene, selects the necessary information in a scene and the areas that the human eye will be attracted to, so that the important information can be given priority in the next processing[19]. In the process of image analysis, it can extract a few areas of interest from a complex scene, and prioritize the analysis and processing of these areas, which can greatly improve the processing efficiency. In recent years, great progress has been made in the modeling and calculation of visual attention mechanism, which is widely used in target recognition, video quality assessment and other fields.

To determine the suitable model, we analyze the principle of several commonly used calculation models of the saliency map and compare them with experiments. As shown in Fig. 1, from top to bottom, there are four original pictures, as well as saliency maps calculated by the Itti model, SR model, FT model, CA model, SLIC model and GBVS model. In the saliency map, the brightness of the color corresponds to the magnitude of the saliency. The Itti model[20] is simple and fast to implement the model, but the effect of the saliency map is not good, and there is a situation of missing judgment for some salient areas. SR (spectral residual saliency detection) model[21] has made full use of the features of the image in the frequency domain, paying attention to the boundary information of the salient region, but the global information of the image is lost seriously, and the whole target cannot be marked evenly. FT (frequency-tuned salient region detection) model[22] has used the method of global comparison to calculate the saliency map. However, the proportion of misjudgment in the salient area is high. CA (context-aware saliency detection) model[23] synthetically extracts the visual features of color and location, retains the contour of the more complete salient regions, but loses the information in the salient region. It does not highlight the most significant part, and the calculation process is complex and time-consuming. The SLIC (superpixel-based saliency detection) model[24] can extract the boundary of the salient region clearly, and the effect is the best, but the algorithm is too complex. The GBVS (graph-based visual saliency) model[25] weakens the extraction of the object contour, and the model is relatively simple. At the same time, through the extraction of color, brightness and direction features, we can accurately obtain salient areas, and the contrast between the salient area and the background is very obvious. So we use the GBVS model to calculate the saliency map.

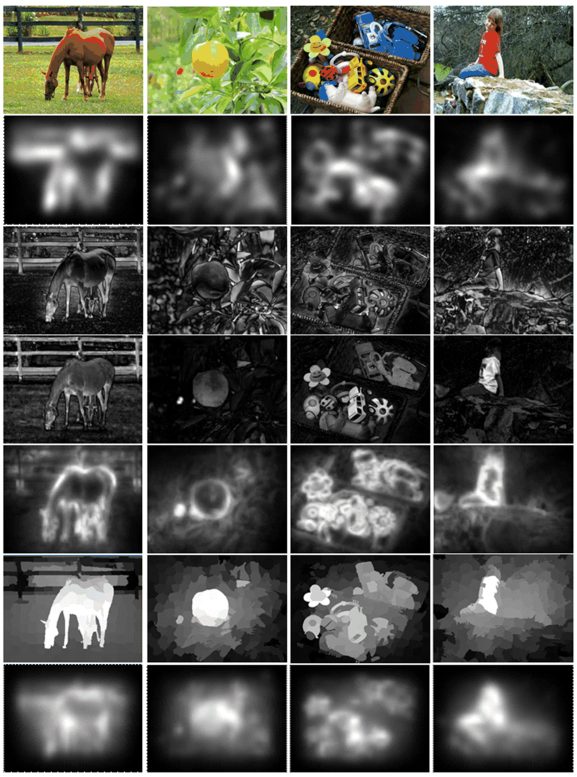

Fig. 1 Comparison of saliency maps calculated by different models

1.2 Graph-based visual saliency model

GBVS model is based on the traditional Itti model, and then simulates the visual resource allocation characteristics of human vision system in scene search scanning to generate a saliency map. The core idea is to use the ergodic probability distribution of Markov Random Chain, the principle of the model is shown in Fig. 2. The first step is to obtain the color, direction and brightness of the original image at different scales. Then, the Markov stationary distribution of each visual feature in different scales is calculated, and the results are superposed in each feature channel and normalized. Finally, the saliency map is obtained by linear superposition of these three kinds of graphs.

Firstly, the input image is preprocessed with 1/2 down-sampling and Gaussian low-pass filter to generate nine scale pyramids, each layer of which is actually a two-dimensional low-pass filter, that is

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (1)$$

In formula (1), the coordinates of each pixel

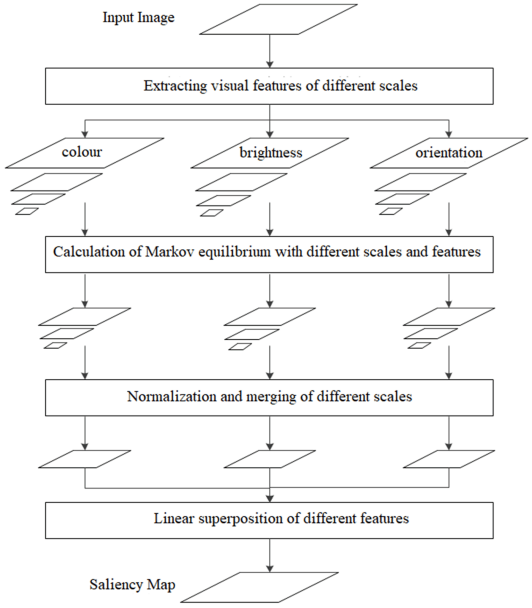

Fig. 2　The schematic diagram of GBVS model

in the image are represented by (x,y), and the scale factor is represented by σ. The larger the σ, the smoother the image. The results of filtering, and down-sampling in the previous layer will be output to the next layer.

Before extracting primary visual features on different scales of the original image, the color in the input image must be transformed into biologically appropriate color and brightness information. Then the Gabor filter is used to obtain the information of the image direction feature, which generally includes four directions: 0 °, 45 °, 90 ° and 145 °. Visual features are represented as one brightness channel, four color channels and four direction channels. Therefore, nine feature pyramids are formed, expressed as $F=\{I,G,R,B,Y,O_1,O_2,O_3,O_4\}$, each feature pyramid in the set has nine layers, that is, nine scales.

The Markov stationary distribution is calculated for the filtering results of each scale in 9 channels. The specific method is to find out the difference value between any two pixels in the image, and assign the corresponding weight value to the two points by the distance between the two points and the difference of the gray level, so as to get a full connection graph (Fig. 3). Next, according to the principle of Random Walk, we can find the nodes with less probability of being visited, and then assign a larger salience value to the corresponding pixel points. The process of finding the stationary state is actually the process of finding the eigenvector corresponding to the maximum eigenvalue in the corresponding weight matrix of the graph.

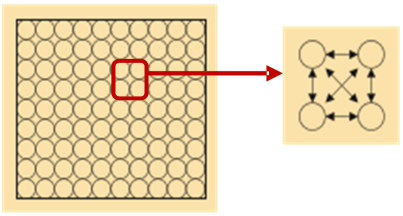

Fig. 3　Full connection graph[25]

Next, the results of each channel are superposed and normalized. The small-scale image is expanded and then overlapped with the large-scale image in the same channel. For example, the direction channel can obtain 4 sub direction maps, and then merge them into a saliency map. Finally, three saliency images are added linearly, and a saliency image of the same size as the original image can be generated by normalization.

1.3　Region of interest description

HOG[26] is a gradient eigenvector descriptor, which can effectively represent the density distribution of the edge direction of the target. Define the gradient of pixel (x,y) in the image as

$$\left.\begin{aligned} G_x(x,y)&=H(x+1,y)-H(x-1,y)\\ G_y(x,y)&=H(x,y+1)-H(x,y-1)\end{aligned}\right\} \quad (2)$$

where $G_x(x,y)$, $G_y(x,y)$, $H(x,y)$ are the gradient in the horizontal direction, the gradient in the vertical direction and the pixel value at the point (x,y) in the input image respectively. The gradient amplitude and direction at the point (x,y) are

$$\begin{aligned} G(x,y)&=\sqrt{G_x(x,y)^2+G_y(x,y)^2}\\ \theta(x,y)&=\tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right)\end{aligned} \quad (3)$$

The extraction and calculation of HOG feature are mainly divided into the following steps:

(Ⅰ) Adjust the sample image size to 64×64 pixels.

(Ⅱ) The image is divided into 8×8 cells, the gradient direction range $[-\pi/2 \quad \pi/2]$ is divided into 9 intervals evenly, and the gradient values of the pixel points in each interval are counted to the gradient histogram, each interval represents one dimension, and each cell is represented by a 9-dimensional feature vector.

(Ⅲ) A block is composed of four adjacent cells, so that a block can be described by a 36-dimensional feature vector.

(Ⅳ) The whole image is scanned in blocks, and the scanning step is one cell.

(Ⅴ) All 36-dimensional block eigenvectors are combined into a 1764-dimensional HOG eigenvector.

2　Build classifier

We can use the saliency map modeling to extract multi-target rough, and then use classifier

based on machine learning to identify target fine, which can greatly improve the speed of the algorithm. In this paper, we use SVM[27] to build the target classifier. SVM is based on the principle of maximizing the geometric interval, that is, the classification hyper plane with the largest geometric interval is the optimal classification surface.

There are countless hyperplanes to classify a linearly separable dataset, but only one with the largest geometric interval. The purpose of maximizing the interval is to classify the training data with sufficient credibility. Suppose x represents the data points of an n-dimensional vector, y represents two different types of each data point, usually -1 and 1. The SVM classifier is trained by using N data point test sets to get the optimal hyperplane, which can be expressed as

$$w^{\mathrm{T}}x+b=0 \tag{4}$$

where x is the point on the hyperplane and w is the vector perpendicular to the hyperplane. Taking the two-dimensional space shown in Fig. 4 as an example, circle and cross are used to represent positive and negative training samples, $H: w^{\mathrm{T}}x+b=0$, $H_1: w^{\mathrm{T}}x+b=1$, $H_2: w^{\mathrm{T}}x+b=-1$, which are parallel to each other. These sample points on H_1 and H_2 together form the support vector, and the distance from these points to the optimal interface H is the same and the closest to the interface H .

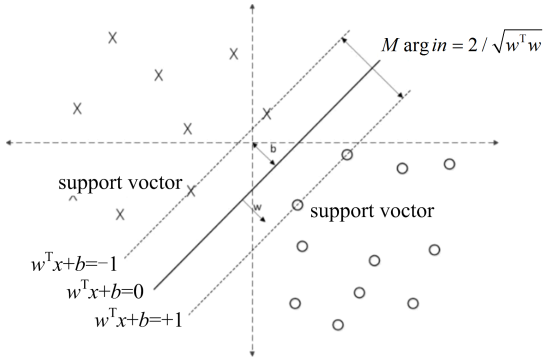

Fig. 4 Schematic diagram of two-dimensional linear separable optimal interface

The distance between the two hyperplanes can be solved by mathematical geometry. The distance is $\frac{2}{\|w\|}$. it can be seen that the smaller $\|w\|$ is, the greater the distance between them. In addition, it is necessary to ensure that all sample points cannot be within the interval between two hyperplanes, so the following conditions should be met for all sample points, we have

$$y(w^{\mathrm{T}}x+b)\geqslant 1 \tag{5}$$

It can be seen that the problem of finding the optimal hyperplane by SVM is an actual optimization problem, that is, finding the minimum $\|w\|$ under the condition of Eq. (5). By solving the quadratic programming problem, we can find the maximum interval classifier. For the linear non separable problem, the kernel function can be transformed into the linear classification problem, and the kernel function is the Gaussian kernel function[28], that is

$$K(x,z)=\varphi(x)\cdot\varphi(z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) \tag{6}$$

where $z=\varphi(x)$ is the mapping from the input space to the feature space. Let $\frac{1}{2\sigma^2}=0.09$, and finally get the classification decision function as

$$f(x)=\mathrm{sign}\left(\sum_{i=1}^{N_s}a_i^* y_i\exp(-0.09\|x-z\|^2)+b^*\right) \tag{7}$$

We use the support vector mechanism to build target classifier. Firstly, the kernel function is selected according to the data characteristics, and the penalty factor C and the insensitive parameter γ are selected. We selected 1000 positive samples (target) and 1000 negative samples (non-target), with the ratio of training and testing samples was 4∶1. Each sample is represented by a 1764-dimensional HOG feature vector, with positive samples labeled 1 and negative samples labeled -1. We normalize the sample data and save it in a unified format. Input the training data to train the SVM model, and the classifier is used to classify the region of interest. After getting the model, test the model with test data. Fig. 5 is a relationship block diagram of the classifier design and recognition.

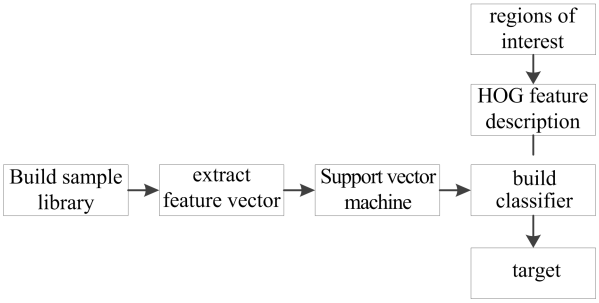

Fig. 5 Classifier training and recognition block diagram

3 Target localization

Firstly, we need to obtain the region of the target in the image, and then the location of the target is realized by using the imaging principle of the camera. Generally, we will simplify the camera model to the pinhole model. Considering the problem of distortion, using the pinhole model to calculate the target position will cause large

errors. Fig. 6 is a single camera imaging model considering the first-order radial distortion, in which $O-X_wY_wZ_w$ is the world coordinate system, $P(X_w,Y_w,Z_w)$ is any point in the world coordinate system, $p(u,v)$ is the ideal imaging point, p' is the actual imaging point.

Let the camera optical axis be perpendicular to the ground without considering the distortion temporarily. Take the intersection of the camera optical axis extension line and the ground as the origin of the world coordinate system, as shown in Fig. 6. If the pixel value of the target in the image is $p(u,v)$, and the $p(u,v)$, $p(u,v)$, $p(u,v)$, $p(u,v)$ of the camera is known, and the height of the camera from the ground is $p(u,v)$ (i. e., $p(u,v)$), then according to the geometric principle, we can get: $p(u,v)$. Finally, the position information of the target relative to the camera is obtained.

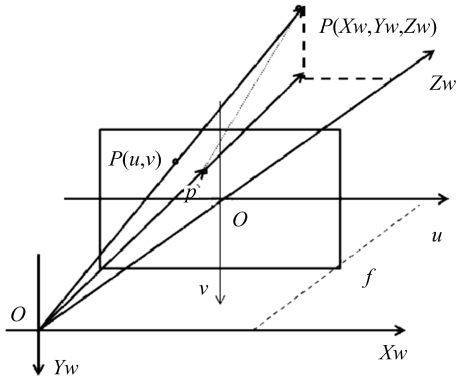

Fig. 6　A single camera pinhole imaging diagram considering radial distortion

Let the camera optical axis be perpendicular to the ground without considering the distortion temporarily. Take the intersection of the camera optical axis extension line and the ground as the origin of the world coordinate system, as shown in Fig. 7. If the pixel value of the target in the image is $p(u,v)$, and the f_x, f_y, c_x, c_y of the camera is known, and the height of the camera from the ground is H (i. e., z_w), then according to the geometric principle, we can get: $\frac{u-c_x}{f_x}=\frac{x}{H}$, $\frac{v-c_y}{f_y}=\frac{y}{H}$. Finally, the position information of the target relative to the camera is obtained.

To obtain the position of the target relative to the UAV, the camera's internal parameters are obtained by checkerboard plane calibration method[29]. Since the wide-angle camera is used, distortion needs to be considered. The parameters obtained by calibration are: linear model parameters $\alpha_u,\alpha_v,u_0,v_0$, and nonlinear distortion parameters k_1,k_2. The rotation matrix R of camera

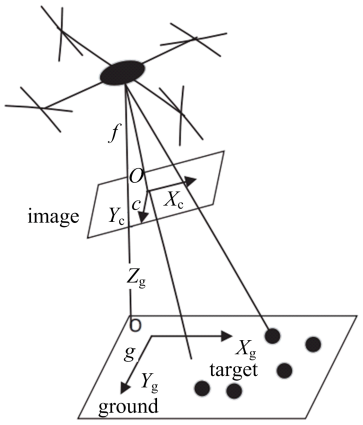

Fig. 7　Target position calculation of monocular camera

external parameters can be calculated by the attitude angle given by IMU, and the translation vector t is $\vec{0}$, the aircraft height z_g is obtained by the ultrasonic sensor.

The corresponding relationship between the points in the image and the points in the three-dimensional coordinate system is as

$$z_c\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}1/dx & s & u_0\\0 & 1/dy & v_0\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}f & 0 & 0 & 0\\0 & f & 0 & 0\\0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}R & t\\0^{\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}x_g\\y_g\\z_g\\1\end{bmatrix}=$$
$$\begin{bmatrix}\alpha_u & s & u_0\\0 & \alpha_v & v_0\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}R & t\\0^{\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}x_g\\y_g\\z_g\\1\end{bmatrix}=M_2M_1\begin{bmatrix}x_g\\y_g\\z_g\\1\end{bmatrix}=M\begin{bmatrix}x_g\\y_g\\z_g\\1\end{bmatrix}\tag{8}$$

Taking the midpoint of each target area as the coordinates (u,v) of the target in the image, the coordinates of the target in the world coordinate system can be obtained by Eq. (8), that is, the relative position of the target and the aircraft.

4　Experimental results and discussion

The UAV used in our project is DJI M100. The processing system is the DJI Manifold system, which is an embedded high-performance airborne computer specially designed for flight platform. Target tracking is carried out by using Zenmuse X3 camera.

When the UAV and the target are stationary, the target can be located in the visible range at different altitudes. The relative position measured by the hand-held laser rangefinder is taken as the reference value and compared with the positioning result. Fig. 8 shows the positioning results of our method. The data in Tab. 1 is the comparison between the calculated results of positioning algorithm and the measured data of the rangefinder.

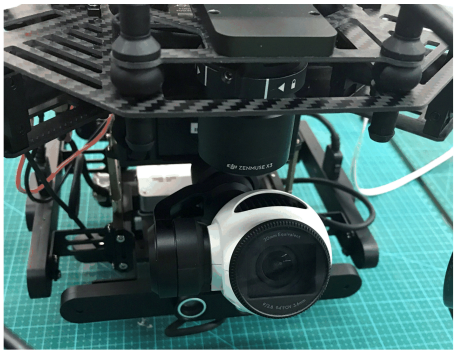

Fig. 8 Experimental platform

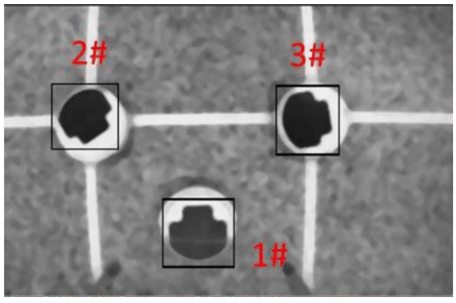

Fig. 9 Target identified when UAV is 1.2 m high

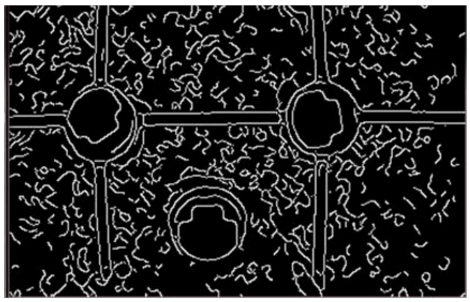

Fig. 10 Canny's edge image

Tab. 1 Comparison between the calculation results of location algorithm and the measured data

number	measured data: x/cm	calculated data: x/cm	measured data: y/cm	calculated data: y/cm
1#	−20	−21.527	4	−6.132
2#	27	29.214	−45	−49.528
3#	25	28.716	50	51.914

To test the positioning error when the target moves, we use the positioning data of the motion capture system OptiTrack as the benchmark to measure the accuracy of the algorithm in this paper. The Figs. 10-12 show the comparison results of the positioning obtained by the algorithm and the motion capture system. It can be seen that our algorithm has a delay of about 50ms and a positioning error of about 14cm, which meets the flight requirements of small UAV in the indoor environment.

The sources of errors may be: ① the small change of aircraft attitude angles will affect the camera angle and cause errors. ② Camera calibration errors. ③ A long time use of camera makes internal parameters change.

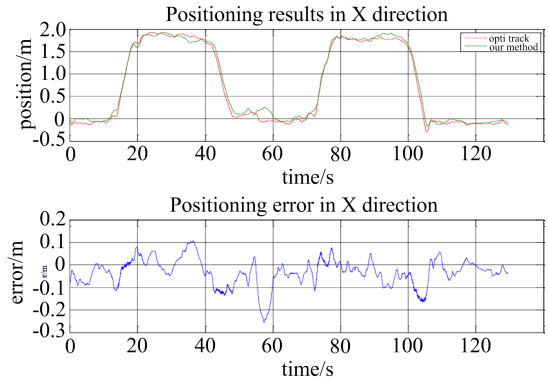

Fig. 11 *X*-direction positioning results and errors

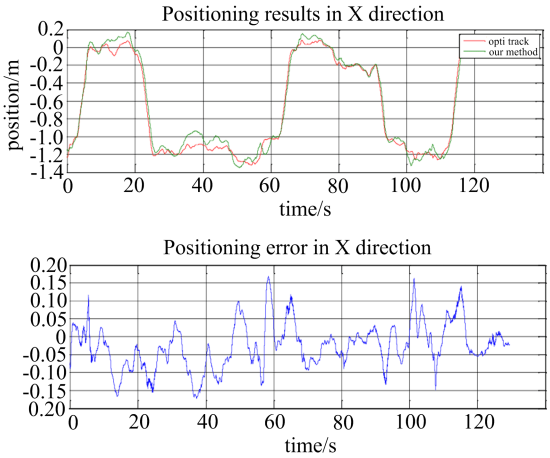

Fig. 12 *Y*-direction positioning results and errors

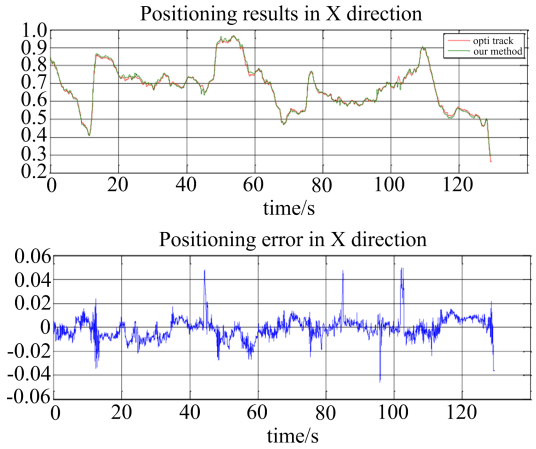

Fig. 13 *Z*-direction positioning results and errors

Compared with the ground robot, target tracking based on UAV platform has the following challenges:

(Ⅰ) UAV can move in a three-dimensional environment, and the captured images are complex and noisy.

(Ⅱ) UAV is a relatively unstable platform, which requires fast update frequency of data and a higher speed of image processing.

(Ⅲ) The target information identified will directly affect the control of UAV, so the algorithm should have higher robustness and

accuracy.

Our approach is applicable to systems with these problems, not just the UAV platform. The target detection algorithm in this paper improves the accuracy on the premise of ensuring the detection speed.

At present, the condition of the UAV flight experiment is ideal, but in the actual flight of UAV, there will be more complex situations. Therefore, we will design more target tracking experiments in different scenarios to verify and improve it.

5 Conclusion

In view of the high requirement of real-time and robustness of UAV target recognition and tracking algorithm, as well as the noise caused by the change of illumination and camera motion, our proposed method is to extract the region of interest roughly based on the visual attention mechanism, and then obtain the target by the fine identifying of the region of interest. We use the HOG feature to describe the extracted region of interest. Finally, we use SVM to train the target classifier, use the machine learning method to identify the target from the region of interest, and then carry out the target location calculation. The experimental results show that our algorithm is of great significance to the target tracking of small UAV in the indoor environment.

References

[1] FORSYTH D A, PONCE J. Computer Vision: A Modern Approach [M]. Second Edition. New Jersey: Prentice Hall, 2012: 255-261.

[2] COMANICIU D, MEER P. Mean shift: A robust approach toward feature space analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619.

[3] NING J F, ZHANG L, ZHANG D, et al. Interactive image Segmentation by maximal similarity based region merging[J]. Pattern Recognition, 2010, 43(2): 445-456.

[4] MEDEIROS F N S, CARVALHO E A, USHIZIMA D M, et al. SAR imagery segmentation by statistical region growing and hierarchical merging [J]. Digital Signal Processing, 2010, 20(5): 1365-1378.

[5] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters [C]// Computer Vision and Pattern Recognition. IEEE, 2010:2544-2550.

[6] UGARRIZA L G, SABER E, VANTARAM S R, et al. Automatic image segmentation by dynamic region growth and multiresolution merging [J]. IEEE Transactions on Image Processing, 2009, 18(10): 2275-2288.

[7] ZHAI W X, CHENG C Q. A camshift motion tracking algorithm based on kalman filter [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2015, 51(5): 799-804.

[8] YU W. Object tracking with particle filter in UAV video [J]. Proceedings of SPIE-The International Society for Optical Engineering, 2013, 8918(48):10.

[9] LI Z H, WANG L, CUI J G. Weak aerial target tracking algorithm based on Camshift and Particle Filter [J]. Computer Engineering and Applications, 2011, 47(9): 192-195.

[10] ZHAO P, SHEN T, SHAN B. An object tracking algorithm for TV guiding systemof UAV based on particle filter [J]. Optics and Precision Engineering, 2008, 1: 026.

[11] ZHAO P, SHEN T, SHAN B. An object tracking algorithm for TV guiding system of UAV based on particle filter [J]. Optics and Precision Engineering, 2008, 1: 026.

[12] WANG W, MENG Z H. Target tracking base on improved Camshift algorithm [J]. Information Technology, 2015(1):85-88.

[13] CHANG L, DUARTE M M, SUCAR L E, et al. A Bayesian approach for object classification based on clusters of SIFT local features [J]. Expert Systems with Applications, 2012, 39(2): 1679-1686.

[14] YU G, YING X. Architecture design of deep convolutional neural network for SAR target recognition[J]. Journal of Image and Graphics, 2018.

[15] RUSSAKOVSKY O, DENG J. Imagenet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3):211-252.

[16] UIJLINGS J R R, SANDE K E A, GEVERS T. Selective search for object recognition[J]. International Journal of Computer Vision, 2013.

[17] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525.

[18] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2005, 1: 886-893.

[19] ZHAO B, WU X, FENG J, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.

[20] ITTI L, KOCH C. Computational modelling of visual attention[J]. Nature Reviews Neuroence, 2001, 2(3): 194-203.

[21] HOU X, ZHANG L. Saliency detection: A spectral residual approach [C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007:11-19.

[22] HEMAMI S, ESTRADA F, SUSSTRUNK S. Frequency-tuned salient region detection [C]// IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009). IEEE, 2009:1597-1604.

[23] GOFERMAN S, ZELNIKMANOR L, TAL A. Context-aware saliency detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(10):1915-26.

[24] LIU Z, LE M, LUO S. Superpixel-based saliency detection[J]. International Workshop on Image Analysis for Multimedia Interactive Services, 2013, 8215(2):1-4.

[25] HAREL J, KOCH C, PERONA P. Graph-based visual saliency[C]. Proceedings of Neural Information Processing Systems (NIPS), 2007.

[26] KLETTE R. Concise Computer Vision-An Introduction into Theory and Algorithms[M]. Springer London, 2014.

[27] Chang C C, Lin C J. LIBSVM: A library for support vector machines [J]. ACM Transactions on Intelligent Systems & Technology, 2011, 2(3):389-396.

[28] HOFMANN T, SCHÖLKOPF B, SMOLA A J. Kernel methods in machine learning [J]. Annals of Statistics, 2008, 36(3):1171-1220.

[29] ZHANG Z. A flexible new technique for camera calibration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330-1334.