Dynamic Background Modeling Based on SIFT Feature Matching

Xiaodong Hu
Beijing Institute of Tracking and Telecommunications Technology, Beijing, China
Email: huxiaodong2037@163.com

Zhenghao Zhang and Chong Li
Beijing Institute of Tracking and Telecommunications Technology, Beijing, China
Email: 2842249840@qq.com, lichongpku@126.com

***Abstract*—In recent years, the detection of moving objects in surveillance video has become a hot topic in the field of computer vision, which has a wide range of applications. In order to extract the foreground object in moving surveillance video, a motion compensation model combining adaptive Gaussian mixture model and SIFT feature is proposed in this paper. The method compensates for camera motion to offset the effects of background image motion. The simulation results show that the dynamic background modeling method based on SIFT feature matching proposed in this paper has excellent performance. The method can extract the foreground targets of surveillance video in different scenes.**

***Index Terms*— SIFT, Gaussian mixture model, object extraction, background modeling**

I. Introduction

Video surveillance is the most important means of information acquisition in China's "security industry." With the smooth development of the "Safe City" construction, surveillance cameras are widely installed throughout the country [1]. People use information from a wide range of surveillance video to deal with problems in areas such as security. In recent years, the number of cameras in counties and townships in various provinces, cities and provinces in China has been increasing. A large number of companies and departments have even achieved full coverage of surveillance video [2]. For example, the distribution density of surveillance cameras in Beijing, Shanghai, and Hangzhou is approximately 71, 158, and 130 per square kilometer. The number of cameras reached 1,150,000, 1 million and 400,000 respectively, providing us with abundant and massive monitoring video information.

At present, the automatic processing and forecasting of surveillance video information has received great attention in many fields such as information science, computer vision, machine learning and pattern recognition [3]. How to effectively and quickly extract the foreground target information in the surveillance video is a very important and fundamental problem [4-6]. The difficulty of this problem lies in the fact that videos that require effective separation of moving foreground objects often have complex, varied, and dynamic backgrounds [7, 8]. This technique can often provide effective assistance for general video processing tasks. Take the example of screening and tracking criminals at night. If we can extract foreground objects in advance, and determine which videos do not contain moving foreground objects. For the video that contains the remaining moving targets, police officers need to identify the pure prospects of eliminating background interference. Therefore, this technology has been widely used in video target tracking, urban traffic detection, long-term scene monitoring, video motion capture, video compression and other applications.

II. Extraction Model of Foreground Target

The flow chart of the research method proposed in this paper is shown in Fig. 1. Firstly, the initial background is obtained by statistically averaging the image sequence over a period of time (5 frames are selected in this paper) and stored in the cache. After denoising the current frame and differentiating the background, the Gaussian model filter is used to update the background. Then use the updated background to replace the background image in the cache and perform the difference with the newly acquired current frame image. Finally, according to the characteristics that the gray value of the change region corresponding to the current frame in the difference image is obviously larger than the gray value of the background region, the separation of the foreground and the background is accomplished by the threshold segmentation technique.

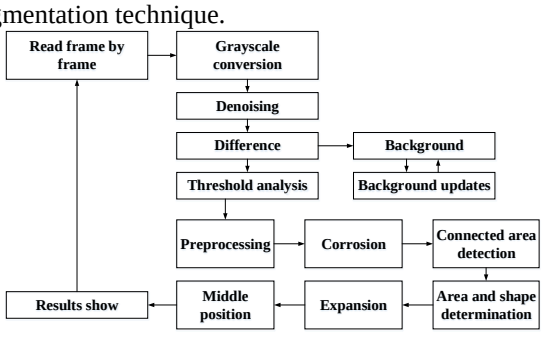

Figure 1. Flow chart of foreground object extraction

Based on the inhomogeneity of the model distribution in the scene, this paper proposes an adaptive hybrid Gaussian model method. Compared with the hybrid Gaussian model method, this method enhances the adaptability of the model and greatly improves the speed of the algorithm. For the problem of ghosting in the detection of moving objects, this paper conducts in-depth research, proposes a corresponding solution by introducing self-learning, and verifies the effectiveness of the method through experiments.

In addition, in view of the problem of noise in the detection results, this paper proposes a method to detect moving objects in images based on adaptive Gaussian mixture model. After the background difference and the threshold segmented image are morphologically filtered and connectivity is detected, some foreground regions are obtained. The segmented foreground region is expressed and described by using the region area and the circumscribed bounding box rectangle, and then according to the criterion, the human body motion can be found. In the foreground area of the target feature, the pedestrian target is divided, and its spatial position in each frame of the video sequence is obtained to provide a data source for the intelligent monitoring system.

A. Denoising and smoothing

Many noises are generated during image acquisition. These noises mainly include random noise generated during the acquisition process and interference caused by external environmental influences. These noises affect the accuracy of target detection. Therefore, removing unnecessary noise is a very important task in the image processing.

From a statistical point of view, most of the neighboring pixels in the same image have small differences in gradation and are highly correlated. This characteristic determines that most of the energy in the image converges in the low frequency region. In contrast, the high-frequency region mainly concentrates on the details of the image and the noise and other information and energy generated during image conversion and transmission. This feature provides us with ideas for removing image noise. The high-frequency component is attenuated and the low-frequency component is enhanced. This is the smooth denoising of the image. The image smoothing process also attenuates the boundary information during the process of suppressing and eliminating the outside noise. As a result, the sharpness of the image has to be reduced to different extents. Therefore, both noise and detail must be taken into consideration when processing the image.

We use a median filter denoising method. The basic principle of median filtering is to replace the value of a point in a digital image with the median value of each point in a neighborhood of the point, eliminating isolated noise points. The specific implementation method is to use a structured two-dimensional sliding template to sort the pixels in the board according to the size of the pixel value, and generate a monotonically rising (or falling) two-dimensional data sequence. Two-dimensional median filter output is as follows:

$$g(\mathrm{x},\mathrm{y}) = \mathrm{med}\{ f(\mathrm{x}-\mathrm{k},\mathrm{y}-\mathrm{l}),(\mathrm{k},\mathrm{l}\in \mathrm{W})\} \quad (1)$$

Among them, *f(x, y)* and *g(x, y)* are the original image and the processed image respectively.

B. Binarization

The purpose of binarization is to extract the target object from the image. The specific extraction process can be described as follows: For the characteristics of the image f(x), a certain threshold T is selected (the threshold selection principle depends on the specific situation), then two sets will appear. One is a set that is larger than the threshold, and the other is a set that is smaller than the threshold. For the set larger than the threshold, we set the gray value of the pixel to 255 and the pixel value of the other set to 0. At this point, the image is converted to a black and white image. The specific threshold transformation formula is as follows:

$$f(\mathrm{x},\mathrm{y}) = \begin{cases} 0, & x < T \\ 255, & x > T \end{cases} \quad (2)$$

Where T is the specified threshold and x is the pixel grayscale value at the midpoint of the image.

C. Adaptive Gaussian mixture model

The Gaussian mixture model is proposed by Stauffer et al. The basic idea is to define K (basically 3-5) states to define the characteristics of each pixel. By training the video images, a corresponding Gaussian model is established to form a corresponding background model. And according to the corresponding formula, It is judged whether the pixel belongs to the background model and the former point of view is extracted. If the pixel satisfies the Gaussian model, this pixel is marked as a background point, otherwise it is marked as a front sight.

Due to the influence of environmental factors, there are usually some disturbing factors, such as leaf shaking, water waves, and light changes. For background instability, using a single Gaussian model to model the background is not satisfactory. In the case of background object interference, we need a Gaussian model to describe this pixel value when a background object exists. When the background object leaves, we need another Gaussian model to describe it. Therefore, a background Gaussian model is used to describe the background model.

Assume that at time t, the pixel value of each pixel of the image can be described by K Gaussian distributions. The probability of the pixel observations in the current frame is as follows:

$$P(\mathrm{I}_t(\mathrm{x},\mathrm{y})) = \sum_{i=1}^{K} w_t^i(\mathrm{x},\mathrm{y}) * \mathrm{N}(\mathrm{I}_t(\mathrm{x},\mathrm{y}), \mu_t^i(\mathrm{x},\mathrm{y}), \delta_t^i(\mathrm{x},\mathrm{y})) \quad (3)$$

The K Gaussian distributions are arranged in descending order. Since the variance and weight of the matching Gaussian distribution will change, the purpose of the ordering is to ensure that the Gaussian distribution with the most possible background is at the front. After sorting, only the first B Gaussian distributions are used to represent the background distribution, ie

$$B = \arg\min(\sum_{i=1}^{b} w_{i,t} > T_H) \tag{4}$$

Where T_H is the background threshold. After the model is established, the i-th model is matched according to the following rules:

$$O_t^i(\mathrm{x},\mathrm{y}) = \begin{cases} 1, |\mathrm{I_t}(\mathrm{x},\mathrm{y}) - \mu_t^i(\mathrm{x},\mathrm{y})| < \mathrm{D}\,\delta_t^i(\mathrm{x},\mathrm{y}) \\ 0, \text{otherwise} \end{cases} \tag{5}$$

Among them, D is a confidence parameter, and the value is generally between 2 and 3. If the result is 1, the pixel value conforms to the background model and can be determined as a background point, and the corresponding parameter in the model is updated using the pixel value. If it is 0, then the pixel is considered not to conform to the Gaussian model, and it is judged as a front sight, and the model parameters are not updated.

In the process of target detection, the pixel values of each pixel in the video image are matched with a Gaussian mixture model to determine whether the pixel satisfies a certain model already existing in the mixed Gaussian model. When a pixel value meets a certain Gaussian distribution, the weight of the Gaussian distribution is updated:

$$O_t^i(\mathrm{x},\mathrm{y}) = \begin{cases} (1-\alpha)\omega_t^i(\mathrm{x},\mathrm{y}) + \alpha, i = m \\ (1-\alpha)\omega_t^i(\mathrm{x},\mathrm{y}), \text{otherwise} \end{cases} \tag{6}$$

In the formula, α is the background learning rate, which determines the rate of change of distribution weights; ω is the distribution weight. After updating, we need to normalize Gaussian distribution weights at the same location. In the update algorithm, the average update rate is larger, and the variance update rate is smaller. However, a small variance update rate will cause the variance to converge slowly, so the update rate of the variance should be taken to a large value during the initial period of Gaussian distribution training, and then the update rate of the variance should be kept stable and small. Based on this, the update rates of the mean and variance should be determined separately. The update rate of mean and variance in this experiment is shown in the following formula.

$$\begin{aligned} p_\mu(t) &= \alpha / \omega_t^i \\ p_{\sigma^2}(t) &= 1/\mathrm{t}^{3/2} + 0.001 \end{aligned} \tag{7}$$

Among them, the average update rate is fast and accurate. The variance update rate has a relatively large value at the beginning of the training of the model. The convergence of the model is good. Only a small sample can be used to establish a good model. The value of update rate after the establishment of the model is relatively small, and it will stabilize at a smaller value, so that the model has better stability. This can not only guarantee the rapid convergence of the model, but also make the model have better stability.

For a fixed scene in video surveillance, if the pixels in the scene are from the surface of an object under fixed lighting conditions, it is reasonable to describe the pixel process and image noise with a Gaussian model. However, in the actual monitored scene, often some regions will be in a situation where the surfaces and edges of multiple objects alternate. For these regions, multiple Gaussian models need to be used to approximate the description. When the light of the scene changes slowly, it is reasonable to use a parameter-adaptive Gaussian model description. However, after analysis, it is found that the background model of most areas in the video surveillance scene is accurate with the single Gaussian model. Only a few areas need to be approximated by multiple Gaussian models, and the number of models often needs to change according to the change of the scene. Especially when the light changes rapidly or the scene changes rapidly, the parameters do not have time to adapt to rapid changes. In this case, the old model needs to be deleted and a new model is created to adapt to the scene change.

D. Detection of connectivity blocks

The adaptive hybrid Gaussian model is a conventional method for segmenting camera motion targets. Because it has the advantages of simple principle, good real-time performance, and rich target information contained in the segmentation results, it has become a widely used method for moving target detection. However, in more complex scenarios, it is difficult to obtain satisfactory results. Because of the complexity of the background and human motion, such as the scene light changes and the closeness of the human body's gray and background, it will affect the image segmentation, making it difficult to accurately reflect the target contour and the uniform connected area.

Since the above processing is only based on low-level image processing, several connected foreground regions included in the obtained binary image need to be further processed and discriminated so as to eliminate the foreground region generated due to light variation and other noise interference. The morphological erosion treatment removes objects smaller than the structural elements, so the image undergoes a slight morphological corrosion operation (in this paper, the radius is 3). Corrosion of the small structure elements in the image is to destroy the connectivity of the foreground connected areas generated by the change of the scene light, and also to remove noise interference. Then, the image is scanned, each connected region is marked by the clustering principle of neighborhood grayscale homogeneity, and the region area descriptor is used to describe it. Compared with the area threshold (selected as 100 pixels in this paper), the reserved area is greater than or equal to the connected area of the threshold, and the other foreground area gray value is set to 0.

In order to highlight the reserved foreground area, the morphological expansion operation of the obtained image is performed after the connectivity detection ends to enhance its connectivity and counteract the negative effects of the corrosion operation. This completes the extraction of the foreground region of the corresponding moving object in the current frame.

III. MODELING OF DYNAMIC BACKGROUND

Research on moving target detection in the context of sports cameras began in the 1990s. The ideas for solving

this problem can be roughly divided into two categories: (1) Solving the image transformation model or the camera motion parameters. In addition, the motion of the camera is compensated to cancel out the influence caused by the motion of the background image, thereby extracting the moving object. (2) Suppose there is a big difference between the background motion and the target motion, and directly solve the motion characteristics between the pixel or image feature blocks. Image segmentation is then performed to obtain areas with similar motion characteristics.

In surveillance video, when the surveillance camera shakes or shifts, the video will also experience temporary jitter. This kind of video transformation can be regarded as a kind of linear affine transformation in a short time, such as rotation, translation, scale change and so on. In this paper, based on the background motion compensation of feature matching in Scale Invariant Feature Transform, the Gaussian background difference is used to detect the target. The flow chart of the research method for the foreground object extraction model under the sports camera is shown in Fig. 2.

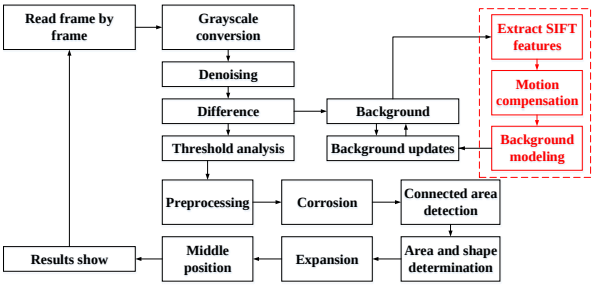

Figure 2. Dynamic background modeling flowchart based on SIFT feature matching

A. Extracting SIFT Features

SIFT is called Scale Invariant Feature Transform. The SIFT feature is invariant to rotation, scaling, and brightness change of the image, and has good stability to the change of viewing angle, affine transformation, and noise. The main steps of the SIFT feature calculation are: detecting scale spatial extreme points; accurately locating extreme points; assigning direction parameters for each key point; and generating key-point descriptors. After solving the SIFT feature vectors of two adjacent images, KD-TREE is used to organize the feature data. The Euclidean distance is used as a similarity criterion to match, and the matching feature point of the current frame feature point in the previous frame is found.

When the SIFT feature extraction generates a keypoint descriptor, this paper takes a keypoint as the center and takes a 16x16 window, and computes 8 directions of gradient direction histogram on the 4x4 small block. A seed point is formed by calculating the cumulative value of the gradient direction. In this way, each key point has 4x4 total 16 seed points, each seed point has 8 directions of information, and the feature vector of each feature point is 128-dimensional.

Fig. 3 shows the extracted SIFT feature points in two adjacent frames.

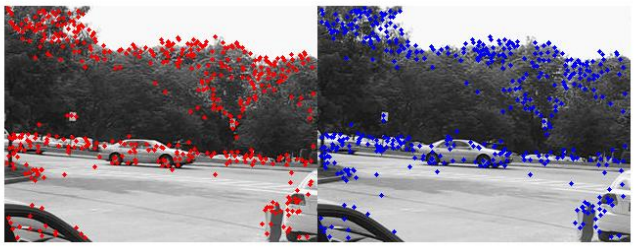

Figure 3. SIFT feature extraction results

B. Motion Compensation

Because there are many points of mis-matching in SIFT feature matching based on simple Euclidean distance measure, these mis-matching points will seriously affect the solution of the parameters of the subsequent transformation model. Therefore, we need to use robust estimation methods to remove these mismatched points. The matching features in the two graphs, one part on the background and the other on the moving target. Relative to the background, the area of the foreground part of the target is relatively small, and the number of corresponding matching features is also generally less, so the wild point in the matching feature point can be eliminated by the RANSAC method.

After matlabR2013a version, RANSAC fine matching and transformation model parameter estimation can be implemented by the function estimate Geometric Transform at the same time.

[T,iloc2_m,iloc1_m]=estimate Geometric Transform(loc2_m,loc1_m,'affine')

Among them, the input value loc2_m is the coordinate of the rough matching point of the image to be registered, the loc1_m is the coordinate of the coarse matching point of the reference image, the affine is the transformation model, and the output value T is the transformation model parameter matrix. The iloc2_m is the coordinates of the fine matching point after the image to be registered passes RANSAC, and iloc1_m is the coordinates of the fine matching point of the reference image.

Fig. 4 shows the matched SIFT feature points in Fig. 3. The left and right graphs successfully match 362 feature points.

Figure 4. SIFT feature matching results

Fig. 5 shows the features matched by RANSAC in Fig. 3. The left and right graphs successfully match 317 feature points. This shows that the algorithm can effectively remove noise.

Figure 5. RANSAC precise matching results

Next, this paper solves the affine transformation model parameters between two adjacent images. When the camera is far away from the scene, the distance between the camera and the scene is far greater than the camera focal length. The 6-parameter affine model can better approximate the changes between the images. The affine transformation of the image is as follows:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} a_0 x_{t-1} + a_1 y_{t-1} + a_2 \\ a_3 x_{t-1} + a_4 y_{t-1} + a_5 \end{bmatrix} \tag{8}$$

Where x_t and y_t are the x- coordinates and y-coordinates of the image at time t; parameters a0, al, a3, a4 describe the rotation and scaling of the image motion; a2, a5 describe the translation of the image.

The following figure shows the affine transformation results of two adjacent frames in Fig. 3. Among them, the affine transformation matrix is as follows:

T.T = [0.9997 0.0004 0
-0.0006 1.0001 0
1.2724 0.0245 1.0000]

That is, the formula for image affine transformation is:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0.9997 x_{t-1} - 0.0006 y_{t-1} + 1.2724 \\ 0.0004 x_{t-1} + 1.0001 y_{t-1} + 0.0245 \end{bmatrix} \tag{9}$$

The left picture in Fig. 6 is the original picture of the 15th frame of the video, and the right picture is the result of the affine transformation of the 15th frame. When you look closely at the right image, you can find discontinuities on the right and bottom sides of the image.

Figure 6. The results of affine transformation

In Fig. 7, the left picture shows the original image of the 14th frame of the video, the middle picture shows the result after the affine transformation of the 15th frame, and the right figure shows the registration of the original image of the 14th frame and the affine transformation result of the 15th frame.

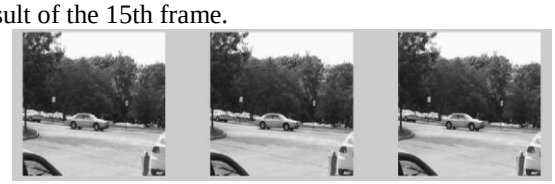

Figure 7. Image overlay results after affine transformation

C. Background Modeling

Unlike the background modeling of a stationary camera, in the case of a sports camera, the background in the previous frame of image needs to be motion compensated, and then the current background model is updated. The basic principle of background updating is that the area detected as the foreground is updated with the affine transformed background model, and the area detected as the background is updated with the current frame image.

The left figure of Fig. 8 shows the result of the registration of the original image of the 14th frame and the affine transformation result of the 15th frame. In order to better demonstrate the effect of background updating, this paper selects the original image of the first frame and the original image of the fifteenth frame to perform affine transformation. Firstly, the affine transformation matrix of the original image of the first frame and the original image of the fifteenth frame is obtained, and then the affine transformation of the fifteenth frame original image is performed. The figure on the right shows the registration result of the original image of frame 1 and the affine transformation image of frame 15. From the results, we can see that there is a clear dividing line on the left. This is because the background area in the judgment result is directly updated with the current image, and the foreground area in the judgment result is updated with the affine transformation of the background image. Due to the distortion of the affine transformed image, some margin will appear in the edge part. Therefore, after filling the blank area with the current image, there will be image discontinuity and obvious boundary lines.

Figure 8. Background modeling results after affine transformation

IV EXPERIMENTS

The upper left corner image (the first one) in FIG. 9 is the original video frame diagram, and the upper right corner (the second diagram) is the corresponding foreground image. The lower left corner (3rd map) is the foreground filtered image filtered by the connected graph, and the lower right corner (4th map) is the significant target block diagram, in which the green box is the detected significant target. It can be seen from the figure that there is little background noise and can effectively detect significant targets.

Figure 9. Extraction result of foreground target from outside sports camera

In the upper right corner of FIG. 10, the significant target detected in the 45th frame. Comparing Fig.1 and Fig.2, we can preliminarily think that the adaptive hybrid Gaussian model method can be applied to this scenario and can detect the motion foreground target more accurately. And it can solve the ghost problem well. However, due to factors such as indoor ambient lighting and reflection, there is a large amount of noise in the left and upper areas. Comparing Fig. 2 and Fig. 3, we can preliminarily assume that the connected block model can be applied to this scenario and can remove the above noise effects more accurately.

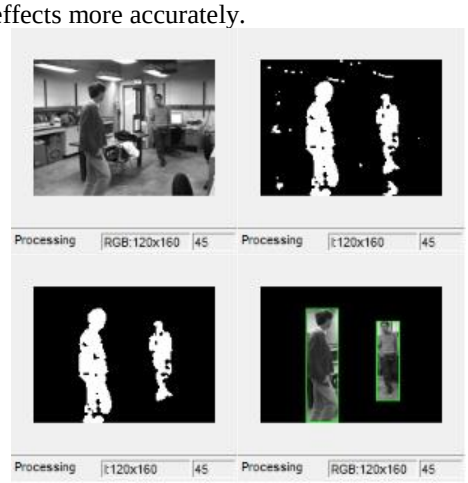

Figure 10. Extraction result of foreground target from inside sports camera

V CONCLUSION

In this paper, an adaptive hybrid Gaussian model is established, and a connected block detection method is adopted to solve the problem of extracting the foreground target from the surveillance video with dynamic background information and effectively removing the influence of ghosting. This paper proposes a motion compensation model that combines the adaptive Gaussian model with SIFT features to compensate for camera motion. The experimental results show that this method can cancel out the influence of background image motion and successfully extract the moving target.

REFERENCES

[1] Andrews Sobral & Antoine Vacavant, A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos, Computer Vision and Image Understanding, Volume 122, May 2014, Pages 4-21

[2] B. Lee and M. Hedley, “Background estimation for video surveillance,” IVCNZ02, pp. 315–320, 2002.

[3] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., vol. 2. IEEE, 1999.

[4] E. J. Cand`es, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” Journal of the ACM (JACM), vol. 58, no. 3, p. 11, 2011.

[5] D. Meng and F. De la Torre, “Robust matrix factorization with unknown noise,” in IEEE International Conference on Computer Vision, 2013, pp. 1337–1344.

[6] Q. Zhao, D. Meng, Z. Xu,W. Zuo, and L. Zhang, “Robust principal component analysis with complex noise,” in Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 55–63.

[7] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, “RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 11, pp. 2233–2246, 2012.

[8] M. Babaee, D. T. Dinh, and G. Rigoll, “A deep convolutional neural network for background subtraction,” arXiv preprint arXiv: 1702.01731, 2017.

Xiaodong Hu received the B.E. degree in Department of Electronic Science and Technology, from University of Science and Technology of China in 2016. He is currently studying in Beijing Institute of Tracking and Telecommunications Technology. His research interests include space robot, deep learning, deep reinforcement learning.