Research on Depth Image Tracking Based on Aerial Refueling

Shuyao ZHANG*, Xianyun QIAN, Jun WANG, Zhenkai FAN
AVIC The First Aircraft Institute, Xi’an 710089, China
Corresponding author’s e-mail: 752698934@qq.com

Abstract: Aiming at the problem of the current aerial refueling docking which is difficult to accurately identify the floating cone sleeve in a complex environment, the Meanshift algorithm can efficiently track moving targets in the plane, there is not very good solution strategy for the pose measurement of the cone sleeve and the movement in the optical axis direction. The Time-Of-Flight (TOF) in the depth camera can directly collect depth images and grayscale images, providing the distance information displayed by a flat camera. By using the Meanshift algorithm to iterate in the three-dimensional space, the Kalman filter is added to achieve robustness to the tracking of moving targets, effectively solve the problem of occlusion of the target, and provide the coordinate position of the target. In the experiment, 600 frames of images are extracted for comparison. The effect of the TOF camera has higher accuracy and stability than that of the flat camera. It still has a stable tracking effect after using the props to cover the cone model, which is better than the traditional plane Meanshift algorithm.

1. Introduction
In the past 30 years, artificial intelligence (AI) technology has gradually entered the frontline of scientific research and production. With high precision and high computing power, this technology has a pivotal position in both engineering and academic directions. For example, all kinds of bionic robots, traffic road monitoring, stock market fluctuation forecast and system simulation, etc. Among them, machine vision technology can drive itself to make decisions by actively acquiring external information, and has the ability to fully perceive the world. Image recognition and tracking technology occupy a unique position. Therefore, the introduction of machine vision technology into the aviation field can quickly promote the trend of equipment intelligence.

In the process of soft air refueling, the pilot of the receiver needs to operate the aircraft to actively dock the "cone sleeve" dropped by the refueling machine. During this period, it will encounter the complicated airflow environment and the influence of strong and weak light. The docking is difficult and the pilot operation is difficult. high. The introduction of machine vision technology in this link will help pilots identify, locate, and track the swing of the refueling cone faster, improve the success rate of docking, and reduce the difficulty of aerial refueling. However, the complexity of the air environment brings full challenges to traditional machine vision technology. For example, factors such as light intensity, background color close to the target or obscured, will cause serious changes in the efficiency, accuracy and robustness of target tracking. Difference. This year, scholars have gradually combined the two, combining technologies such as deep learning and template matching and have made remarkable achievements [3-4]. However, because planar images are susceptible to light interference and accurate depth information cannot be obtained, this paper uses a depth TOF camera for research, combining the planar MeanShift algorithm and Kalman filter to explore from another perspective.
2. MeanShift algorithm
As early as 1975, Fukunaga K, Hostetler LD [1-2] and others first proposed the mean shift algorithm (MeanShift, MS) in related papers. In the following 50 years, the mean shift algorithm has been in the research of researchers. Propelled by continuous development, it has become a more mature member of the current machine recognition algorithms.

2.1. Principle concept
The target tracking algorithm based on MeanShift is a non-parameter density estimation method. The algorithm has obvious characteristics, small calculation amount, simple principle, and high convergence efficiency. Therefore, the recognition effect has a certain degree of robustness when a small part of it is occluded with good adaptability.

MeanShift algorithm is a tracking algorithm based on image template segmentation. The algorithm first decomposes the space $R_d$ into d dimensions, and extracts n sampling points $x_i$ from them, where $i=1,2,...,n$, from which the MeanShift vector at the reference point $x$ can be given, and its basic form is:

$$M_h(x) = \frac{1}{k} \sum_{x_i \in s_h} (x_i - x)$$  \hspace{1cm} (1)

$$s_h(x) = \left\{ y : (y-x)^T (y-x) \leq h^2 \right\}$$  \hspace{1cm} (2)

In formula (1), $s_h$ represents the high-dimensional sphere area with radius h, and the value of y that meets the condition is shown in formula (2). $k$ indicates that among the n sample points $x_i$, $k$ fall into the area $s_h$, $(x_i - x)$ is the offset of the sample point relative to the determined reference point $x$.

The basic vector of the MeanShift algorithm $M_h(x)$ is that in the d-dimensional space, first select a reference point $x$, then select k sample points relative to the reference point $x$ to obtain the drift vector, and finally average. It can be seen that the sample points in all regions are sampled from the probability density function $f(x)$. According to probability statistics, it can be known that if the probability density increases in the direction where the probability density increases the most, the sampled samples in the region almost all fall along the direction of the probability density gradient. Obviously, the corresponding Meanshift vector $M_h(x)$ points to the direction with the largest probability density gradient. As shown in Figure 1, the great circle is $s_h$, and the result of the basic vector $M_h(x)$ is to shift the reference point to the place where the probability density of the sampling point is the largest.
However, in the classic MeanShift algorithm, all sampling points that fall into the space $s_h$, no matter how far away from the reference point $x$, $M_h(x)$ have the same contribution to the drift vector. However, in a complex air environment, when the tracking target is occluded or the background color is similar, the pixel value of the outer layer is easily disturbed, resulting in reduced reliability. Therefore, the central pixel will be given greater weight. The weight of the pixel value is slightly smaller, so the concepts of kernel function and weight coefficient are introduced. Thus, the form of MeanShift can be extended to:

$$M(x) = \frac{\sum_{i=1}^{n} G_H(x_i - x)\omega_i(x_i)(x_i - x)}{\sum_{i=1}^{n} G_H(x_i - x)\omega_i(x_i)}$$  \hfill (3)$$

Among them, $\{x_i\}_{i=1,2,\ldots,n}$ are $n$ pixels in the area where the target is located. $\omega(x_i) \geq 0$ is a weight assigned to the sample point, $G(x)$ is the unit kernel function, and $\omega(x_i) \geq 0$ is the weight, that is, the weight of the sample point, which represents the importance of the sample point $x_i$.

$$\omega_i = \sum_{u=1}^{m} \frac{q_u}{P_u(y_o)} \delta[b(x_i) - u]$$  \hfill (4)$$

$$G_H(x_i - x) = |H|^{-\frac{3}{2}} G(H^\frac{1}{2}(x_i - x))$$  \hfill (5)$$

Refer to different distances to assign different weights to pixels. The commonly used MeanShift expansion form is:

$$M_h(x) = \frac{\sum_{i=1}^{n} G(\frac{x_i - x}{h})\omega(x_i)(x_i - x)}{\sum_{i=1}^{n} G(\frac{x_i - x}{h})\omega(x_i)}$$  \hfill (6)$$
2.2. Improved tracking algorithm for color interference problem

At the same time, because the MeanShift algorithm is extremely sensitive to colors, the ordinary RGB color space is likely to cause recognition errors, which is not conducive to achieving the tracking effect. Therefore, this article converts the RGB color space to HSV (Hue, Saturation, Value) color space. The distribution is more even. Taking the histogram of the H component, the appearance probability of the color can be obtained, and instead of the original value of the original pixel, the probability distribution diagram of the color can be obtained [5].

3. TOF camera

Although the 2D camera has developed more mature in the tracking and recognition direction, it can also have a higher tracking efficiency for multiple targets [6-7]. However, a 2D camera can only obtain the planar features of the tracking target. In a complex aerial environment, it is susceptible to the influence of illumination and it is difficult to capture the position information of the target. Therefore, this article uses a depth camera for image recognition to provide a target cone. Multiple poses and depth information.

3.1. Introduction to TOF camera principle

Depth cameras are divided into binocular cameras, structured light cameras and depth cameras. Although binocular cameras and structured light cameras have high positioning accuracy, they are highly dependent on external light sources. During aerial refueling missions, strong and weak light, forward light and backlight will be encountered, so the recognition and tracking of the camera will be affected, resulting in loss of frames and thus loss of tracking targets. The TOF camera will shield the external light source, project an infrared light source, and adjust the visual distance range according to the modulated light source power.

TOF cameras usually use a sine wave modulation method. The phase shift of the sine wave at the transmitting end and the receiving end is proportional to the distance between the object and the camera. The distance can be measured through the phase shift, that is, for the optical signal of the modulation frequency f, the wavelength is $\frac{c}{f}$, Therefore, the received light phase delay 0-2π corresponds to the flight distance $0 \sim \frac{c}{f}$, so the distance d is

$$d = \frac{1}{2} \cdot \frac{c}{f} \cdot \frac{\varphi}{2\pi} \tag{7}$$

The above $\frac{1}{2}$ is to consider the back and forth distance of light. Relative to the phase of the transmitted signal, the reflected light generates an integral gated signal with a phase shift of 0°/90°/180°/270°, and the received signal is integrated to obtain:

$$\varphi = \arctan \left( \frac{Q_2 - Q_1}{Q_3 - Q_1} \right) \tag{8}$$

$$A = \frac{\sqrt{(Q_1 - Q_2)^2 + (Q_1 - Q_3)^2}}{2} \tag{9}$$

$$B = \frac{Q_1 + Q_2 + Q_3 + Q_4}{4} \tag{10}$$

Equation (8) represents the phase of the received signal, equation (9) represents the received signal strength, and equation (10) represents the intensity of ambient light interference.

During the working process of the TOF camera, the system collects the reflected light for a complete time regardless of the distance of the target object.
3.2. Image preprocessing of TOF camera

Based on the TOF camera’s ranging and framing principle, the collected image is a depth map, and there will be more noise points in the initial picture. This noise point is caused by active errors and passive errors. Therefore, preprocessing of the collected images is essential.

In this paper, the depth image directly collected by TOF is grayed out and a histogram is drawn [8].

Next, perform threshold segmentation on the target: set a binary mask image mask, and only retain the area corresponding to the mask 1 in the depth map to complete the segmentation.

$$\text{mask}(x, y) = \begin{cases} 
1, & f(x, y) \leq \tau \\
0, & f(x, y) > \tau 
\end{cases}$$  \hspace{1cm} (11)

In specific operations, there is no guarantee that the object to be recognized is closest to the TOF camera, so the depth map cannot strictly divide the foreground and the background. If the image is still segmented using formula (11), the result will contain the depth value outside the object to be recognized. Therefore, this article sets two depth values to ensure good results for segmented images.

$$\text{mask}(x, y) = \begin{cases} 
1, & \tau_1 \leq f(x, y) \leq \tau_2 \\
0, & \text{others}
\end{cases}$$  \hspace{1cm} (12)

After the depth threshold is cut, the result of Figure 3 is obtained. It can be found that due to the noise characteristics of the TOF camera, there is obvious noise at the edge of the image. In order to eliminate the interference of noise, the method of contour detection is used to search and retain the largest connected domain and eliminate the small area of noise area, and the result can be as good as the picture. In this way, the final mask image of the object can be obtained to segment the target object.
4. 3D target tracking

4.1. MeanShift algorithm in depth image

The MeanShift algorithm is widely recognized as a color-sensitive target tracking algorithm, but the core of the algorithm is to track the arbitrary distribution represented in the probability distribution map, that is, find a matching method with higher confidence for the distribution map of every two frames [9].

Different from RGB-D images, TOF cameras can only obtain depth images. This information only contains depth information. There is a peak near the gray value of 200, which is the cone of the tracked target.
After smoothing and filtering the probability, the average depth value of the cone sleeve can be obtained. Because there are other background environments interference behind the cone sleeve, the cone sleeve part corresponding to the identification frame should be given a larger weight in the weight distribution, and the inner circle of the cone sleeve is assigned a larger weight than the umbrella sleeve part.

From the above, this article introduces the kernel function $K(n)$:

$$K(n) = A \cdot e^{-\frac{(x-b)^2}{\sigma^2}}$$  \hspace{1cm} (13)

Then the depth probability distribution can be expressed as:

$$H(n) = \frac{1}{M \cdot N} \sum_{(x,y) \in I} K(n)[\delta(I(x,y) - n)]$$ \hspace{1cm} (14)

$$K(n) = e^{-\frac{(x-H_{sleeve})^2}{\sigma^2}}$$ \hspace{1cm} (15)

Where $n=1, 2,...$, $K$, $H_{sleeve}$ is the average depth value of the cone sleeve, $M$ and $N$ represent the length and width of the search box respectively, $K$ is the number of overall depth values; $\delta$ is the unit pulse function; $\sigma$ is the umbrella sleeve Part of the weight of the kernel function parameter; the target template of the MeanShift algorithm of formula (17), the back projection is calculated in the next frame, and the target position can be obtained by iteration.

4.2. Application of Kalman in depth image

Kalman filter is an efficient autoregressive filter. It is believed that the value of the next state can be predicted from the previous state. After the prediction, it is compared and corrected with the previous state. It has a perfect match for target tracking.

For 2D image recognition, the system can be defined as:

$$X(k|k-1) = AX(k-1|k-1) + BU(k) + W(k)$$ \hspace{1cm} (16)

$$P(k|k-1) = AP(k-1|k-1)A^T + Q$$ \hspace{1cm} (17)

$$K(k) = P(k|k-1)H^T \cdot [HP(k|k-1)H^T + R]^{-1}$$ \hspace{1cm} (18)

$$X(k|k) = X(k|k-1) + K_s(k)(Z(k) - HX(k|k-1))$$ \hspace{1cm} (19)

$$P(k|k-1) = [I - K(k)H]P(k|k-1)$$ \hspace{1cm} (20)
Among them, equations (16) and (17) are prediction equations, equations (18)-(20) are filter equations, \( x(k-1|k-1) \) and \( x(k|k) \) respectively represent the prediction of the state at time \( k-1 \) at time \( k \), and the state at time \( k \) after filtering; \( P(k|k-1) \) and \( P(k|k) \) are the covariance at time \( k \) obtained at \( k-1 \), and the covariance at time \( k \) after filtering; \( K(k) \) is the Kalman gain at time \( k \), \( Z(k) \) is the depth measurement position at time \( k \), \( A \), \( B \) and \( H \) are state transition matrix, control input matrix and transition matrix respectively.

In the case of different dimensions, each dimension is independent of each other for the Kalman filter, so for the depth map collected by the TOF camera, only the average depth value in the optical axis direction needs to be considered. Assuming that the center mass of the target cone sleeve is \( (x_k, y_k) \) in the plane coordinates, the average speed is \( (v_{x_k}, v_{y_k}) \), and the closest point of the cone sleeve in the optical axis direction is used as the reference point \( F \), then the coordinates of point \( F \) is \( z_k \), and the speed is \( v_{z_k} \).

Because the cone sleeve is in a relatively stable environment when refueling in the air, and its moving speed suddenly increases, so its pose will not change too quickly in every two consecutive frames, so the default is to move at a constant speed. Then assume the sampling interval of every two frames is \( \Delta t \).

The motion state vector of the target is:

\[
X_k = \begin{bmatrix} x_k & y_k & z_k & v_{x_k} & v_{y_k} & v_{z_k} \end{bmatrix}^T
\]

The observed state vector \( Z(k) \) is:

\[
Z(k) = \begin{bmatrix} x_k & y_k & z_k \end{bmatrix}^T
\]

The state transition matrix \( A \) is:

\[
A = \begin{bmatrix}
1 & 0 & 0 & \Delta t & 0 & 0 \\
0 & 1 & 0 & 0 & \Delta t & 0 \\
0 & 0 & 1 & 0 & 0 & \Delta t \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

The observation matrix \( H_k \) is:

\[
H_k = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0
\end{bmatrix}
\]

Let the noise matrix be an independent Gaussian white noise matrix. The 3D MeanShift algorithm combined with the Kalman filter can be used to identify and track the target position.

5. Experimental results
In this paper, the algorithm effect was verified in the laboratory, and the simulation model of the refueling cone sleeve was purchased. The camera adopts SmartTof, and its depth image resolution is 320*240, which can reach 8 meters distance measurement, and the accuracy is about 0.5%-2%. The frame rate is 30fps, and depth images and grayscale images can be collected directly. This article uses depth images to start the experiment. Hardware environment: CPU: Inter CORE 8th Gen, memory: 8GB, Windows10; software environment: Python3.8.
Air refueling generally begins to enter the docking state when the distance between the oil receiving pipe and the cone sleeve is about 3 meters, so in the experimental simulation, the camera is at the same distance from the target. In actual operation, the camera is fixed at a position of about 15° obliquely below, and 600 frames of video are taken. The chase model is held in front, and the cone sleeve is shaken by imitating the movement of the cone. With the adjustment of the positioning posture, the experimental effect is shown in the figure.

Figure 8 Target cone recognition

If the MeanShift algorithm of the traditional 2D plane is used, when the received light or the target background color is similar, the target will be lost when there is target interference. The TOF camera can directly collect the depth information of the image, that is, the distance information of the target can be calculated more accurately. It has inherent advantages over flat cameras.

Figure 9 Planar tracking disturbed by background

| Tracking algorithm | Number of test frames | Track the frame correct | Target lost frame | Accuracy |
|--------------------|-----------------------|-------------------------|-------------------|----------|
| Plane tracking     | 600                   | 566                     | 34                | 94.3%    |
| Noted Algorithm    | 600                   | 584                     | 16                | 97.3%    |
| KCF                | 600                   | 575                     | 25                | 95.9%    |

It can be found from the following pictures that after adding the Kalman filter, the cone can still be tracked stably after being occluded by a small part, without losing the target.

Figure 10 Recognition effect after occlusion

After many rounds of experiments, this algorithm has a better effect on the cone sleeve recognition and tracking in a laboratory environment, especially for a small number of occluded targets, which has been significantly improved compared with the classic algorithm.
6. Summary
This article is based on the MeanShift algorithm of the 2D plane, using the depth image as the basis for target tracking, and introducing the Kalman filter to achieve the function of short-term prediction. Through testing in a laboratory environment, the algorithm in this paper has a significant improvement over planar tracking. It can track steadily even after the target cone is occluded. The recognition accuracy is improved without major changes in processing speed compared to the original algorithm. The autonomous docking of aerial refueling provides a new feasible way. The environment in the air is more complicated than that of the laboratory, and the simulation constraints in the experiment will be gradually increased in the future to further improve the authenticity of the experiment.

References
[1] GU Xingfang, MAO Yaobin, LI Qiujie. Survey on visual tracking algorithms based on mean shift[J]. Computer Science, 2012, 39(12): 16-24
[2] YANG L C. Practical problems and solutions in age trend-line analyses for energetic components[C]/The 43th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, 2007:8-17
[3] FUKUNAGAK,Hoatetlerl. The estimation of the gradient of a density function, with application in pattern recognitions in pattern recognition[J]. IEEE Transactions on Information Theory,1975,21(1),32-40
[4] Quanquan, Wei Zibo, Gao Jun. Overview of modeling and control in the docking stage of flexible flexible aerial refueling[J]. Acta Aeronautica Sinica.
[5] Thomas P R , Bhandari U , Bullock S , et al. Advances in air to air refuelling[J]. Progress in Aerospace Sciences, 2014, 71:14-35.
[6] Lu Yanjun, Wang Shiyu, Zhang Taining, et al. Research on improved tracking algorithm for complex environmental problems, Microelectronics and Computer, 2020, 37(9): 78-82
[7] XU Junyan, CUI Zongyong, LUO Yuanqing, et al. A weighted particle filter for pedestrian tracking in complex scenarios[J]. Journal of Signal Processing, 2017, 33(7): 934-942.(in Chinese)
[8] DING Xiaofeng, SHANG Zhenhong, LIU Hui, et al. Multiple template moving object tracking based on Mean Shift[J]. Computer Engineering and Applications, 2017, 53(6): 141-144.(in Chinese)NING Jifeng, ZHANG Lei, ZHANG D, et al. Robust object tracking using joint color-texture histogram[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(7): 1245–1263.
[9] Bradski GR, Kaehler A. Learning OpenCV[M]. Yu Shiqi, Liu Ruizhen, Trans. Beijing: Tsinghua University Press, 2008.
[10] LI Sujuan, WANG Ji, YAN Baoqing, et al. Application of Kalman filter in the object tracking [J]. Modern Electronics Technique, 2007 (13): 110-112.(in Chinese)