Tracking and Measuring Explosion Points with High-Resolution Reconstruction under Binocular Occlusion

Shipeng Cheng, Meili Zhou, Zongwen Bai*
School of Physics and Electronic Information, Yan'an University, Yan'an, Shaanxi, China
**Corresponding Author.*

Abstract: This paper leverages data and projects from Group A to enhance the application of bomb impact point tracking and measurement using binocular vision. The research involved gathering bomb impact measurement data across various mountain peaks of differing elevations, Using binocular drones to collect data. Nevertheless, challenges such as bomb impact overlap and occlusion within the video data were identified. To tackle these equipment-related obstacles, including bomb occlusion and camera overlap issues, remote sensing image reconstruction networks were utilized to reconstruct bomb impact images that exhibited partial overlap. The processed imagery data was annotated utilizing the labelimg annotation tool, in collaboration with the OpenCV data processing utility, for precise labeling of bomb impact images. Moreover, a multi-object tracking network was developed and trained for the effective tracking of bombs. The central aim of this research is to regress the world coordinates of initial bomb impact points by employing bomb point localization algorithms and image regression networks dedicated to bomb measurement. Furthermore, this paper delves into the inaccuracies found within target point measurements and undertakes an error analysis predicated on the information pertaining to the target. To enhance the operational capability of the airborne observation platform, the research entailed the relocation of the electro-optical pod to a predetermined position, followed by the remote transmission of the gathered data to ground-based equipment. The ground-based equipment is designed to configure parameters, control the electro-optical pod, receive commands, and process image data for conducting intersection measurement calculations. The electro-optical pod itself facilitates high-speed measurements of target impact positions across visible light, infrared, and laser modes, additionally offering capabilities for local data storage. The pod's attitude self-stabilization was accomplished with gyroscopes. Meanwhile, the ground equipment facilitates remote control, parameter setting, data reception, and the execution of intersection measurement calculations based on image data.

Keywords: Super-resolution Reconstruction; Explosion Point Measurement; Binocular Vision; Linear Regression; Error Analysis

1. Introduction

The system described is utilized for monitoring and measuring the impact point positions of ammunition fired from multiple-barrel artillery in ground suppression conditions. Prior to firing, all drones transition into a preparation state and await takeoff instructions. Upon receiving the takeoff command, three drones take off simultaneously and proceed to their designated positions and altitudes for standby. The ground integrated display and control system continuously receives real-time status parameters from the drones, electro-optical pod, GNSS/IMU inertial navigation system positioning parameters, and partially real-time transmitted image data [1]. This data includes real-time previews from visible light high-speed cameras or infrared videos and can be promptly displayed on the ground integrated control system, enabling users to monitor the operational status of the drones and the electro-optical pod. The system setup is depicted in the accompanying Figure 1.

Throughout the firing process, the electro-optical pod diligently adheres to a prearranged plan to survey the target area, performing tri-mode measurements in visible

light, infrared, and laser to record high-speed target impact positions, with all data stored locally. Subsequent to firing, through an analysis of the downloaded visible light and infrared video data, this research harnesses computer vision techniques for image change detection to autonomously identify the bomb impact targets [2]. The system meticulously locates the positions of the targets by analyzing three visible light video images and an infrared image, enabling precise triangulation measurements of the bomb impact targets. By combining the identified positions with ground target information and utilizing observations from multiple angles, advanced spatial triangulation algorithms accurately calculate the three-dimensional coordinates of the bomb impact targets.

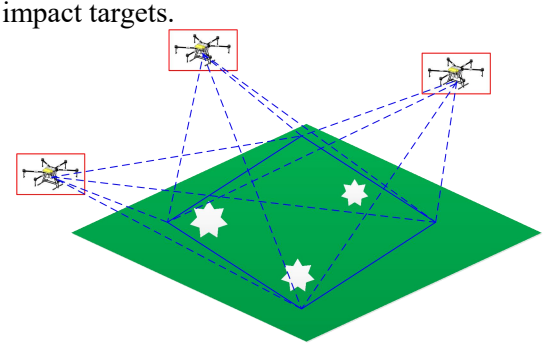

Figure 1. Drone Measurement Scenarios

2. Utilizing Super - Resolution Reconstruction for Precision Bomb Impact Tracking and Measurement in Binocular Visual Occlusion Scenarios

2.1 Hybrid Gaussian Background Modeling

The Gaussian distribution is employed to describe the consistent spatial distribution of all images within a video sequence, aiding in the development of a background Gaussian distribution probability model [3]. By utilizing a predefined probability threshold T, the pixel values undergo evaluation to determine their adherence to the model. If the criteria are met, the pixel is classified as part of the background; otherwise, it is recognized as a moving target. The formula for this classification process is as follows:

$$D(x,y)=\begin{cases}0,\left(I_t(x,y)-\mu_t(x,y)\right)^T\sum_t^{-1}(x,y)\left(I_t(x,y)-\mu_t(x,y)\right)\le T\\ 1,\left(I_t(x,y)-\mu_t(x,y)\right)^T\sum_t^{-1}(x,y)\left(I_t(x,y)-\mu_t(x,y)\right)>T\end{cases} \quad (1)$$

The threshold T can be adjusted either through adaptive thresholding or manual adjustments to reduce computational load. When dealing with a grayscale image input, the aforementioned equation can be streamlined into a one-dimensional formula:

$$D(x,y)=\begin{cases}0,(I_t(x,y)-\mu_t(x,y))/\sigma_t(x,y)\le T\\ 1,(I_t(x,y)-\mu_t(x,y))/\sigma_t(x,y)>T\end{cases} \quad (2)$$

where $\sigma_t(x,y)$ represents the mean square deviation of the Gaussian distribution of (x,y) pixels[4]. The Gaussian background model implements background updates using the following formula:

$$\mu_{t+1}(x,y)=(1-\alpha)\mu_t(x,y)+\beta I_t(x,y) \quad (3)$$

$$\sigma_{t+1}^2=(1+\beta)\sigma_t^2(x,y)+\beta(I_t(x,y)-\mu_t(x,y))^T(I_t(x,y)-\mu_t(x,y)) \quad (4)$$

Where α is the model update rate, and β is the variance update rate.

The background distribution can be accurately represented by a weighted combination of multiple Gaussian distributions, as depicted in Figure 2 below:

Figure 2. Multi-Gaussian Model Weighted Hybrid Representation of the Background

The process for the hybrid Gaussian model is outlined as follows:

A. Initialize the matrix parameters for each Gaussian model.

B. Use the T-frame data image from the video to train the Gaussian mixture model, with the initial Gaussian distribution derived from the first frame pixel.

C. Compare each pixel with the mean of the current Gaussian distribution; if the difference is within 3 times the variance, assign the pixel to the distribution and update its parameters.

D. If the pixel does not match the existing Gaussian distribution, create a new Gaussian distribution based on the pixel. [5].

2.2 Bomb Point Tracking and Target Identification

Upon examining the target and explosion point detection segment, it has been discerned that the system is required to simultaneously recognize multiple target types within the algorithm. Additionally, it should be capable of fulfilling the tracking task for the target along its motion trajectory [6]. Consequently, a

deep learning-based tracking algorithm has been chosen to detect explosive points and targets. The system utilizes a Fully-Convolutional Siamese Network for tracking purposes, adept at pinpointing the location of both the target and the detonation point. It also tracks and recognizes the trajectory of the explosion point from the moment the fuse is activated, through the fireball's emergence, to the eventual dispersion of smoke [7].

The algorithm is anticipated to proficiently track and pinpoint the behavioral trajectory of the target object. The objective is to precisely ascertain the pixel location of the target within the image with maximum accuracy. Moreover, it is designed to align the pixel coordinate point in the image with the actual coordinate position in the real world, utilizing a transformation matrix for precise correspondence [8].

The system employs an enhanced Object-tracking model to detect explosive points and designated targets. The detailed procedure is depicted in Figure 3.

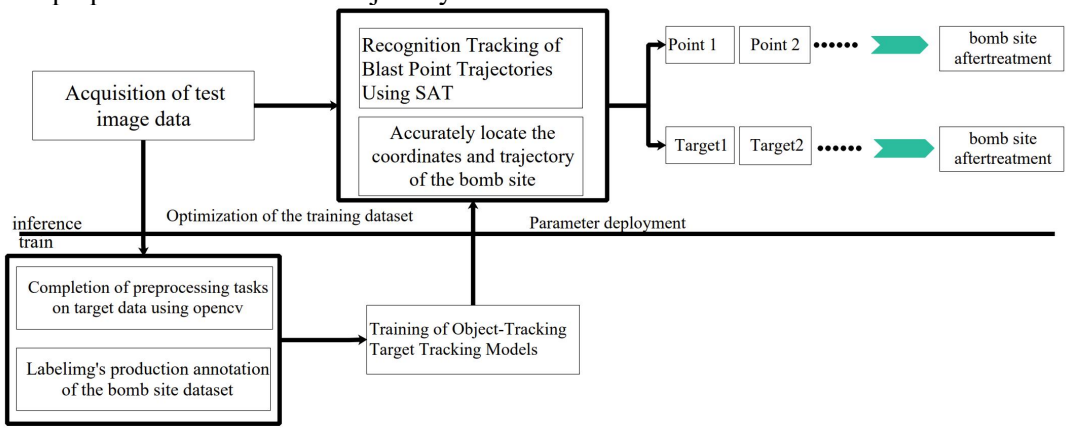

Figure 3. Schematic Diagram of the Workflow of the Bomb Point Target Tracking Model

Prior to delving into the tracking model, it's essential to outline the algorithm for tracking a solitary target. The efficacy of tracking a single target hinges largely on the comparison of features and logical deduction. This feature comparison forms the core of numerous studies. Given the initial frame of the target's image, the most direct method to ascertain the target's position in subsequent frames is by contrasting the upcoming frame's image with the target's image features, typically within a sliding window or object proposal framework. The feature bearing the closest resemblance is deemed the target object[9]. The architecture of the Object-Tracking network is illustrated in Figure 4.

To refine the object tracking algorithm, it is trained using the maximal image search technique. The training regimen for the algorithm is grounded in the discriminative approach, leveraging both positive and negative exemplars. Logical functions are employed as loss functions to systematically train the network.

$$l(y, v) = log(1 + exp(-yv)) \quad (5)$$

Where v is the actual value score of a single sample and candidate pair, and $y \in \{-1,1\}$ is the label of its groundtruth. By using image pairs that contain example images and larger searches, this paper take advantage of the fully convolutional nature of the network during training[10]. This will result in a plot $D \to R$ generated by the score v, with each pair efficiently yielding many samples. This article define the loss of the score plot as the average of the individual losses.

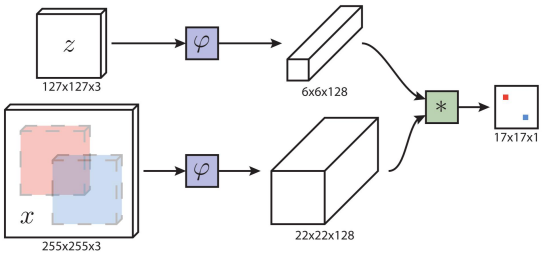

Figure4. Diagram of a Single-target Target Tracking Network

$$L(y, v) = \frac{1}{|D|}\sum_{u \in D} l(y[u], v[u]) \quad (6)$$

Each position in the score graph is required to have the real label $y[u] \in \{-1,1\}$, and the parameter θ of the convolutional network is obtained by applying stochastic gradient descent to the problem.

The image extraction process involves two consecutive frames from the video, ensuring

the presence of objects in both. These frames are spaced no more than T frames apart, and during training, classes that do not pertain to the objects are disregarded. The object proportions within each frame are normalized, maintaining the original aspect ratio of the images. Additionally, if the score plot elements fall within the radius R from the center, this is taken into account alongside the network's stride k.

$$y[u] = \begin{cases} +1 & if\ k||u-c|| \le R \\ -1 & others \end{cases} \quad (7)$$

For enhancement: Elements within the score plot are deemed positive.

In the single-target tracking process, the initial condition is set to the first frame, where the target is assigned a mask. This mask is then predicted for the target in each subsequent frame, effectively transforming the task into one of video target segmentation.

Drawing from the single-target tracking algorithm's insights, the SAT algorithm model meticulously classifies target objects within the tracklet on a pixel-by-pixel basis. It employs a tracking algorithm to continuously observe the segmented target, segments it within the tracked target frame, and utilizes the SAT target segmentation tracking task as a foundation for recognizing and tracing the explosion point's motion trajectory. This methodology aids in partially reconstructing the explosion's complete motion trajectory, allowing for the accurate localization and dynamic monitoring of the event. The extensive network architecture for the SAT segmentation classification task is illustrated in Figure 5.

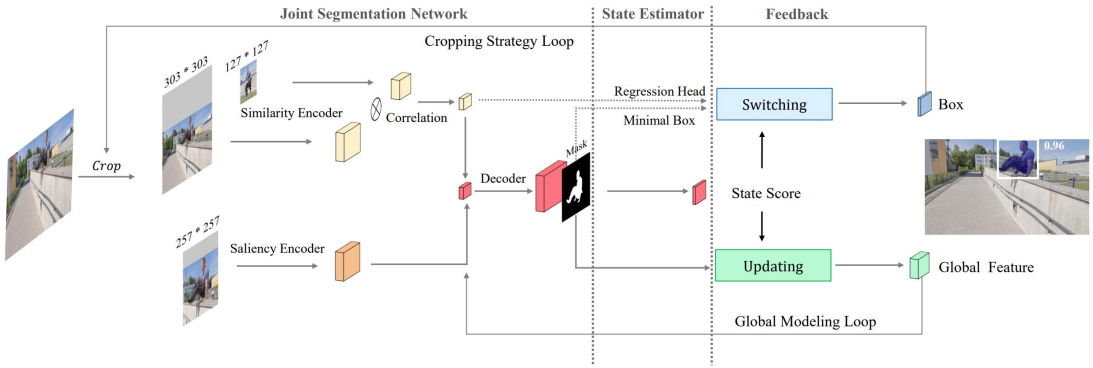

Figure 5. Diagram of the Overall Network Structure of SAT Multi-objective Tracking Task

To enhance the sentence: The system's test image data is initially captured at a resolution of 1080×1920×1. Therefore, it is necessary to modify the network's input parameters to retain the integrity of the initial subtle target information as it is processed by the network.

In the process of dataset construction, the categorization of output identifiers is restructured, amalgamating all classes into two primary groups: explosive points and target points. The annotation process is executed with precision using labelimg, taking into account pre-existing conditions to ensure accuracy[10]. The system extracts video features pertinent to target tasks via 3D convolution, with a particular focus on temporal data along the T-axis for tracking purposes. This approach incorporates functionalities for recognizing trajectories and pinpointing pixel coordinates. In the process of tracking and identifying targets, the four coordinate points defining the entire target in pixel space are determined. The complete motion trajectory of the explosion point, from ignition to smoke formation, is calibrated. Recognition criteria are limited to 6 targets and 7 types of targets at the explosion point.

2.3 Precision Bomb Impact Point Regression Module

To refine the sentence: Considering the ammunition's trajectory and the explosion's timing may not coincide with exact single-pixel precision, it is crucial to determine the exact pixel coordinates of the explosion point. This is achieved by analyzing the locations of the fireball and the smoke observed after the explosion. The algorithmic rationale for this procedure is detailed in Figure 6.

For a refined expression: Training an accurate pixel coordinate regression model for the explosion point necessitates the use of manually annotated optical sequences depicting the explosion's fireball and smoke. This is complemented by data on wind

direction, speed, the ammunition's entry angle, and velocity, along with a precise correlation of explosion points. Such a detailed method enables the extraction of regression model parameters for the explosion point's coordinates. Additionally, a reverse model is constructed to link the wind conditions, ammunition dynamics, and the progression of the fireball and smoke back to the explosion point, which assists in the precise annotation of bomb coordinates. When trained on this rich dataset, the regression model demonstrates enhanced robustness and stability, benefiting from the integration of expert insights.[11].

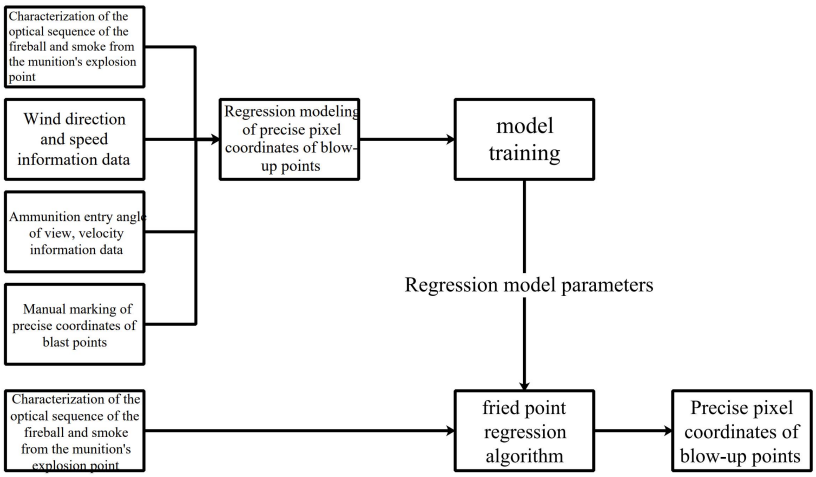

Figure 6. Exact Frying Point Regression Algorithm logic

To refine the sentence: The exact pixel coordinate regression model for the explosion point can be formulated either autonomously, employing diverse regression models and pre-existing data, or through amalgamation with the explosion point detection model. In the integrated approach, the detailed pixel data of the explosion point is incorporated into the detection model's input, with the precise pixel coordinates serving as the model's output. The step-by-step algorithmic development of the complete linear regression model is outlined below.

To refine the sentence: The linear regression algorithm model, designed to ascertain the exact location of the explosion point, posits that there exist n pixel coordinate positions for the explosion point. Each feature is associated with a corresponding weight value, indicative of the parameter's influence on the explosion point's position. The model is conceptualized as the aggregate of the product of features and their respective weights, augmented by an offset value. The mathematical representation is as follows:

$$y = w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n + b \quad (8)$$

As the system aims to regressively predict the coordinates of the initial explosion point, the offset term b is fixed at 0. In this context, $[x_0\ x_1\ x_2 ... x_n]$represents the regression position information and parameter matrix or vector associated with the explosion point at different stages, including the ignition, light spot creation, smoke formation, and smoke dissipation processes.

$$y = w_0 * x_0 + w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n + b(9)$$

The weight w can also be written in the form of a matrix:

$$w = [w_0\ w_1\ w_2 ... w_n] \quad (10)$$

The weights in the regression monitoring task of bomb point represent the impact of various parameters, such as wind direction and the position of the monitored coordinate point (observed by the UAV), on the final prediction result[10]. These weights are represented by vectors.

$$Y = XW^T \quad (11)$$

The loss function is a crucial algorithm that assesses the model's quality to a certain extent. In regression problems, this article commonly employs the mean square error (MSE) as a measure of loss.Within the regression algorithm, MSE serves as the ultimate metric for evaluating the loss and performance of the regression model. Specifically, in determining the exact location of the explosion point, the regression model provides a criterion for measuring the discrepancy between the expected and actual coordinate positions of the explosion point. The loss function is defined as follows:

$$Loss\ function = \frac{1}{m}\sum_{i=1}^{m}(\widehat{y_i} - y_i)^2 (12)$$

Here, $\widehat{y_i}$ Symbolizes the sample's forecasted

value, corresponding to the bomb point's coordinate location as projected by the regression model. Meanwhile, y_i represents the actual value of the sample, indicating the real position in world coordinates after converting the bomb point's pixel coordinates through the matrix. A lower value of the loss function indicates a closer alignment between the predicted and true values.

2.4 Location of the Explosion Point

The determination of the bombing point's direction is achieved by integrating the pixel coordinates captured by the target recognition algorithm with the pose data of the single camera, which is acquired during the calibration phase. The autonomous system ascertains the bombing point's direction by applying the principle of perspective transformation, which projectively converts the pixel coordinates gathered by the UAV into the bombing point's vector. Figure 7 depicts the structural schematic of the autonomous system's approach to establishing the bombing point's orientation.

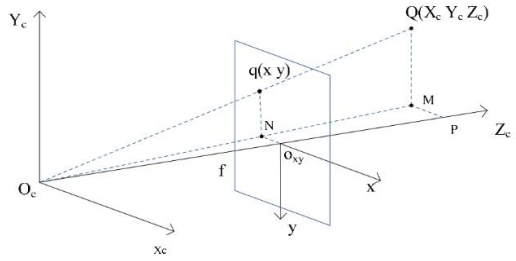

Figure 7. Single-machine Bomb Point Direction Monitoring Map

Suppose that the point Q is a measured point in the spatial field of view (world coordinate system), and the point is projected onto the single-camera imaging plane, $q(x, y)$ represents the imaging point of the camera's coordinate system, and its pixel point is (u, v). Let M be the projection matrix, and since the camera system has successfully completed the calibration, this part is known, then there are:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = M \begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{bmatrix} \begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} \quad (13)$$

In the formula above, Z_c Denotes the z-axis coordinates of point Q within the camera's reference frame. where the camera is positioned[11]. Meanwhile, $(u, v, 1)$ indicates the homogeneous coordinates that match the pixel coordinates of the dual projection points on the imaging surface. and $(X_m, Y_m, Z_m, 1)$ denotes the homogeneous coordinates of the measured point Q in the world coordinate system. With the knowledge of coordinates $(u, v, 1)$ and the projection matrix M, a system of three equations can be constructed. However, when only two equations are available, the direction of the measured point Q between the cameras can still be obtained.

In the multi-machine joint explosion point solution, the pixel coordinates of the explosion point(q1, q2, q3) taken by the three cameras of Lianli are used to construct a non-homogeneous linear equation system, and the world coordinates of the explosion point (Q1, Q2, Q3) are obtained by using the least squares method. The system processes a video feed from three cameras, capturing the target and the explosion point's pixel coordinates; it then computes the global coordinates for each explosion site.

(1) The binocular vision convergence fixed-point model refers to a positioning model in which two cameras capture images of the object being measured from different positions within the same scene, thereby obtaining the distance information of the target point being measured[12]. From the previous single-camera monitoring, a single calibrated camera can determine the direction of the measured target, while two cameras can determine the position information of the measured target through the intersection of directional rays, assuming that the measured point is captured simultaneously, as depicted in Figure 8.

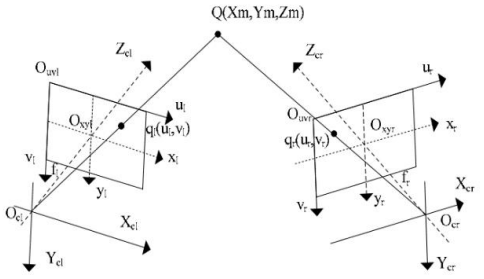

Figure 8. Schematic Diagram of Binocular Vision Convergence Fixed-point Model

Assuming that the fixed point Q is a measured point in the spatial field of view, and the point is projected onto the imaging plane of the left and right cameras, q_l, q_r the pixel coordinates of the two points have been detected to be (u_l, v_l), (u_r, v_r) , and then assuming that the

two cameras in the model have completed the calibration work, and the M_l, M_r are their respective projection matrices, then there are:

$$Z_{cl}\begin{bmatrix}u_l\\v_l\\1\end{bmatrix}=M_l\begin{bmatrix}X_m\\Y_m\\Z_m\\1\end{bmatrix}=\begin{bmatrix}m_{11}^l & m_{12}^l & m_{13}^l & m_{14}^l\\m_{21}^l & m_{22}^l & m_{23}^l & m_{24}^l\\m_{31}^l & m_{32}^l & m_{33}^l & m_{34}^l\end{bmatrix}\begin{bmatrix}X_m\\Y_m\\Z_m\\1\end{bmatrix} \quad (14)$$

$$Z_{cr}\begin{bmatrix}u_r\\v_r\\1\end{bmatrix}=M_r\begin{bmatrix}X_m\\Y_m\\Z_m\\1\end{bmatrix}=\begin{bmatrix}m_{11}^r & m_{12}^r & m_{13}^r & m_{14}^r\\m_{21}^r & m_{22}^r & m_{23}^r & m_{24}^r\\m_{31}^r & m_{32}^r & m_{33}^r & m_{34}^r\end{bmatrix}\begin{bmatrix}X_m\\Y_m\\Z_m\\1\end{bmatrix} \quad (15)$$

Z_{cl}, Z_{cr} respectively represent the coordinates of the depth direction of Refers to the location of point Q within the coordinate system of the dual-camera setup.and m_ij represent the element values in the two projection matrices of M_l, M_r $(u_l, v_l, 1), (u_r, v_r, 1)$ respectively represent the homogeneous coordinates corresponding to the pixel coordinates of the two projection points on their respective imaging planes, and $(X_m, Y_m, Z_m, 1)$ represent the homogeneous coordinates corresponding to the measured point Q in the world coordinate system. The simultaneous formulas can eliminate the Z_{cl}, Z_{cr} in the above equation, and then obtain two sets of linear equations for (X_m, Y_m, Z_m) as follows:

$$\begin{cases}(u_lm_{31}^l-m_{11}^l)X_m+(u_lm_{32}^l-m_{12}^l)Y_m+(u_lm_{33}^l-m_{13}^l)Z_m=m_{14}^l-u_lm_{34}^l\\(v_lm_{31}^l-m_{21}^l)X_m+(v_lm_{32}^l-m_{22}^l)Y_m+(v_lm_{33}^l-m_{23}^l)Z_m=m_{24}^l-u_lm_{34}^l\\(u_rm_{31}^r-m_{11}^r)X_m+(u_rm_{32}^r-m_{12}^r)Y_m+(u_rm_{33}^r-m_{13}^r)Z_m=m_{14}^r-u_rm_{34}^r\\(v_rm_{31}^r-m_{21}^r)X_m+(v_rm_{32}^r-m_{22}^r)Y_m+(v_rm_{33}^r-m_{23}^r)Z_m=m_{24}^r-u_rm_{34}^r\end{cases} \quad (16)$$

The two sets of linear equations in the above equation are essentially two projection line equations composed of the measured point Q and the two cameras in the model. According to the principle of geometry, these two straight lines intersect at the same point, that is, the O_1q_1 of the straight line and the O_2q_2 of the straight line should intersect at the point $Q(X_m, Y_m, Z_m)$ [13] . In theory, the three-dimensional spatial coordinates of the measured point are attainable through the joint resolution of the aforementioned pairs of linear equations. However, the practical application is often marred by noise interference, which skews the projection lines. To counteract this, an estimation derived from the least squares method is utilized in lieu of the true 3D spatial coordinates of the measured point. This method enables the determination of the coordinates for point Q, thereby enhancing the accuracy of distance measurement.

$$\begin{bmatrix}X_m\\Y_m\\Z_m\end{bmatrix}=(A^TA)^{-1}A^TB \quad (17)$$

In the above formula:

$$A=\begin{bmatrix}u_lm_{31}^l-m_{11}^l & u_lm_{32}^l-m_{12}^l & u_lm_{33}^l-m_{13}^l\\v_lm_{31}^l-m_{21}^l & v_lm_{32}^l-m_{22}^l & v_lm_{33}^l-m_{23}^l\\u_rm_{31}^r-m_{11}^r & u_rm_{32}^r-m_{12}^r & u_rm_{33}^r-m_{13}^r\\v_rm_{31}^r-m_{21}^r & v_rm_{32}^r-m_{22}^r & v_rm_{33}^r-m_{23}^r\end{bmatrix} \quad (18)$$

$$B=\begin{bmatrix}m_{14}^l-u_lm_{34}^l\\m_{24}^l-u_lm_{34}^l\\m_{14}^r-u_rm_{34}^r\\m_{24}^r-u_rm_{34}^r\end{bmatrix} \quad (19)$$

2.5 Post-explosion Point Treatment

2.5.1 Cross-validation of the location of the explosion point

The positioning cross-verification is realized by using the trinocular vision fixed-point model, which is built on the basis of the binocular vision convergence ranging model through the reasonable placement of three cameras, and at the same time adjusts the position relationship between the cameras to make their respective optical axes at a certain angle to each other, that is, the model is composed of three sets of binocular ranging models. Assuming that the projection points of any point Q in space on the imaging surfaces of a, b and c are q_1, q_2 and q_3 respectively, the schematic diagram of the trinocular ranging model is shown in Fig.9. Ideally, the measured values of the binocular ranging model composed of camera A and camera B should be Q_1 , the measured value of the binocular ranging model composed of camera B and camera C should be Q_2 , and Q_3 the measured value of the binocular ranging system composed of camera A and camera C should overlap with the actual measured point Q at one point, that is, the three straight lines of $O_{ca}q_1, O_{cb}q_1, O_{cc}q_1$ intersect to the same point Q. However, in the actual ranging environment, based on the existence of errors in the binocular ranging system, the three points of Q_1, Q_2, Q_3 do not overlap with the actual measured point Q[14]. Instead, the three straight lines of $O_{ca}q_1, O_{cb}q_1, O_{cc}q_1$ intersect in pairs to form different three points in space Q_1, Q_2, Q_3 , and the three-dimensional spatial coordinates of these three points can be obtained by binocular ranging algorithm.

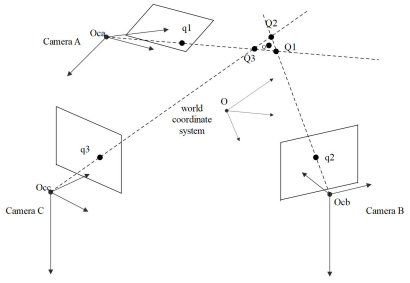

Figure 9. Schematic Diagram of a Three-eye Vision Fixed-point Model

Through the above theoretical analysis, it can

be seen that the binocular ranging system will have a certain error in the positioning of the measured point, especially in the depth direction, so the system proposes a triocular vision joint ranging algorithm to optimize the measured values of the Q_1, Q_2, Q_3 points, and then obtain more accurate distance information of the measured target. That is, let the spatial coordinates of the measured point (X_m, Y_m, Z_m), then the objective function is used to estimate the actual coordinates of the Q point optimally, and the detailed deduction is as follows:

$$F = min\left(\|Q - Q_1\| + \|Q - Q_2\| + \|Q - Q_3\|\right) \quad (19)$$

Assuming that the pixel coordinates of the three points of q_1, q_2 and q_3 projected by point Q on the imaging planes of a, b and c cameras have been detected to be $(u_a, v_a), (u_b, v_b), (u_c, v_c)$, and then assuming that the three cameras in the model have completed the calibration work, and the M_a, M_b, M_c are their respective projection matrices, then there are:

$$Z_{ca}\begin{bmatrix} u_a \\ v_a \\ 1 \end{bmatrix} = M_a\begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11}^a & m_{12}^a & m_{13}^a & m_{14}^a \\ m_{21}^a & m_{22}^a & m_{23}^a & m_{24}^a \\ m_{31}^a & m_{32}^a & m_{33}^a & m_{34}^a \end{bmatrix}\begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} \quad (20)$$

$$Z_{cb}\begin{bmatrix} u_b \\ v_b \\ 1 \end{bmatrix} = M_b\begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11}^b & m_{12}^b & m_{13}^b & m_{14}^b \\ m_{21}^b & m_{22}^b & m_{23}^b & m_{24}^b \\ m_{31}^b & m_{32}^b & m_{33}^b & m_{34}^b \end{bmatrix}\begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} \quad (21)$$

$$Z_{cc}\begin{bmatrix} u_c \\ v_c \\ 1 \end{bmatrix} = M_c\begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11}^c & m_{12}^c & m_{13}^c & m_{14}^c \\ m_{21}^c & m_{22}^c & m_{23}^c & m_{24}^c \\ m_{31}^c & m_{32}^c & m_{33}^c & m_{34}^c \end{bmatrix}\begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} \quad (22)$$

In the formula, M_i represents the projection matrix of each of the three cameras and $M_i = A_i[R_i T_i]$ (where $i = a, b, c$), $(u_a, v_a, 1), (u_b, v_b, 1), (u_c, v_c, 1)$ represent the homogeneous coordinates corresponding to the pixel coordinates of the three projection points on their respective imaging planes, and $(X_m, Y_m, Z_m, 1)$ represents the homogeneous coordinates corresponding to the measured points Q in the world coordinate system. On the basis of the previous section bit algorithm, the three-dimensional spatial coordinates of the three points of Q_1, Q_2, Q_3 can be obtained by using the least squares method, that is, $Q_i(X_{mi}, Y_{mi}, Z_{mi})$ (where $i = 1,2,3$), and it can obtain the projection line equation of twice the number of cameras on the basis of traditional binocular ranging, which can further enhance the anti-noise ability, as shown in the following formula[15].

$$Q_i: \begin{bmatrix} X_{mi} \\ Y_{mi} \\ Z_{mi} \end{bmatrix} = \left(A_i{}^T A_i\right)^{-1} A_i{}^T B_i, i = 1,2,3 \quad (23)$$

where:

$$A_1 = \begin{bmatrix} u_a m_{31}^a - m_{11}^a & u_a m_{32}^a - m_{12}^a & u_a m_{33}^a - m_{13}^a \\ v_a m_{31}^a - m_{21}^a & v_a m_{32}^a - m_{22}^a & v_a m_{33}^a - m_{23}^a \\ u_b m_{31}^b - m_{11}^b & u_b m_{32}^b - m_{12}^b & u_b m_{33}^b - m_{13}^b \\ v_b m_{31}^b - m_{21}^b & v_b m_{32}^b - m_{22}^b & (v_b m_{33}^b - m_{23}^b \end{bmatrix} \quad (24)$$

$$B_1 = \begin{bmatrix} m_{14}^a - u_a m_{34}^a \\ m_{24}^a - u_a m_{34}^a \\ m_{14}^b - u_b m_{34}^b \\ m_{24}^b - u_b m_{34}^b \end{bmatrix} \quad (25)$$

$$A_2 = \begin{bmatrix} u_b m_{31}^b - m_{11}^b & u_b m_{32}^b - m_{12}^b & u_b m_{33}^b - m_{13}^b \\ v_b m_{31}^b - m_{21}^b & v_b m_{32}^b - m_{22}^b & v_b m_{33}^b - m_{23}^b \\ u_c m_{31}^c - m_{11}^c & u_c m_{32}^c - m_{12}^c & u_c m_{33}^c - m_{13}^c \\ v_c m_{31}^c - m_{21}^c & v_c m_{32}^c - m_{22}^c & (v_c m_{33}^c - m_{23}^c \end{bmatrix} \quad (26)$$

$$B_2 = \begin{bmatrix} m_{14}^b - u_b m_{34}^b \\ m_{24}^b - u_b m_{34}^b \\ m_{14}^c - u_c m_{34}^c \\ m_{24}^c - u_c m_{34}^c \end{bmatrix} \quad (27)$$

$$A_3 = \begin{bmatrix} u_a m_{31}^a - m_{11}^a & u_a m_{32}^a - m_{12}^a & u_a m_{33}^a - m_{13}^a \\ v_a m_{31}^a - m_{21}^a & v_a m_{32}^a - m_{22}^a & v_a m_{33}^a - m_{23}^a \\ u_c m_{31}^c - m_{11}^c & u_c m_{32}^c - m_{12}^c & u_c m_{33}^c - m_{13}^c \\ v_c m_{31}^c - m_{21}^c & v_c m_{32}^c - m_{22}^c & (v_c m_{33}^c - m_{23}^c \end{bmatrix} \quad (28)$$

$$B_3 = \begin{bmatrix} m_{14}^a - u_a m_{34}^a \\ m_{24}^a - u_a m_{34}^a \\ m_{14}^c - u_c m_{34}^c \\ m_{24}^c - u_c m_{34}^c \end{bmatrix} \quad (29)$$

It can be simplified to:

$$\begin{aligned} F = min\left(\|Q - Q_1\| + \|Q - Q_2\| + \|Q - Q_3\|\right) \\ = (X_m - X_{m1})^2 + (Y_m - Y_{m1})^2 + (Z_m - Z_{m1})^2 \\ + (X_m - X_{m2})^2 + (Y_m - Y_{m2})^2 + (Z_m - Z_{m2})^2 \\ + (X_m - X_{m3})^2 + (Y_m - Y_{m3})^2 + (Z_m - Z_{m3})^2 \end{aligned} \quad (30)$$

By the fact that the sum of squares of the dispersions of the variables and their arithmetic mean is the smallest, the optimal estimate of the measured point Q can be obtained.

2.5.2 Explosion point anomaly detection

During the capture of the explosion point, images from a particular camera may become unfocused, obscured by fog, or obstructed by smoke on the field. In such instances, the explosion point exception handling module will invoke redundant image data to substitute the compromised data and reinitiate the detection and positioning process for the explosion point.[16]. In scenarios with multiple explosion points, the angle of observation can lead to an adhesion effect among the points captured by the camera. This can significantly compromise the accuracy of the explosion point detection module. To rectify this, it is essential to employ redundant image data from different angles. As explosion points may exhibit varying degrees of adhesion when viewed from distinct perspectives, it is necessary to analyze three separate sets of image data. After identifying the explosion points, their potential adhesion must be tracked across these three data sets to ensure precise recognition.

2.5.3 Cross-validation of multi-mode

information

During the detonation of ammunition, environmental elements like smoke and dust from the explosion can impair the optical lens's ability to capture the explosion point, thereby influencing the outcome. Consequently, thermal imaging data captured by the infrared lens is utilized for the detection and localization of the explosion point. This data also serves to cross-validate the findings from the optical lens, ensuring the precision of the explosion point's coordinate results.

2.5.4 Analysis of sources of error

The error source analysis conducted in this study is divided into two main categories. The first involves assessing the algorithm model's accuracy by evaluating the detected explosion points and the algorithm's training precision. The second category is dedicated to examining the errors related to the alignment between the world coordinates and pixel coordinates, which are crucial for pinpointing the exact location of the explosion point as determined by the algorithm.

In evaluating the error of the tracking model, the mean error across individual losses within the tracking task serves as a metric to assess the algorithm's precision. This section delves into a quantitative analysis of the target's relative error, employing a binocular measurement system. It involves a detailed examination of the variance between the world coordinates of the anticipated bomb point and the coordinates obtained from the pixel data of the specific point via the regression model. The actual coordinates of the explosion site are ascertained through Real-Time Kinematic (RTK) positioning, pinpointing the crater created by the blast to acquire the true global coordinates of the detonation point.

To ascertain the global coordinates of the bomb points predicted by regression, the pixel coordinates generated by the model are converted through matrix transformations. Error analysis is performed by juxtaposing the precise coordinates of the predicted bomb point against the actual detonation point's global coordinates, taking into account a range of external and internal factors, and applying a linear regression model to accurately determine the bomb point's location.

$$y = w_0 * x_0 + w_1 * x_1 + w_2 * x_2 + \ldots + w_n * x_n + b \quad (31)$$

Among the linear regression models, $[x_0\ x_1\ x_2 \ldots x_n]$ respectively represents the evolutionary attitude of the explosion point, $w = [w_0\ w_1\ w_2 \ldots w_n]$ is expressed as the influence of various internal parameters on pixel coordinates, and the specific influence parameters can be expressed as wind direction, altitude, drone position, etc. In the above analysis, the error of the regression model is also briefly analyzed, in which the computer memory temperature, the temperature of the working environment, the humidity of the working environment, the degree of anti-interference of external noise, the maximum wind resistance, and the location of the UAV deployment can be used as factors affecting the internal parameter changes. In addition, the system uses the binocular measurement system by default to perform regression analysis and prediction on the exact world coordinate position of the explosion point[17].

2.5.5 An error analysis model is established to focus on the analysis of the error of parameter calibration

Calibrating the camera parameters and computing the three-dimensional coordinates of the spatial target feature points are essential processes that hinge on the accurate calculation of the feature points' coordinates on the imaging plane. This necessitates a thorough analysis of the error-inducing factors affecting the coordinates of the imaging points on the imaging surface. The primary contributors to calibration error in camera parameters include optical system lens distortion, the camera's internal structural setup, and the quantization error and noise from electronic components. The optical system's lens distortion leads to aberrations in the principal ray, as depicted in figure 10. This distortion shifts the principal point on the image plane and the imaging point away from their ideal locations, impacting the camera's parameter calibration and the subsequent computation of the target feature point's three-dimensional coordinates.

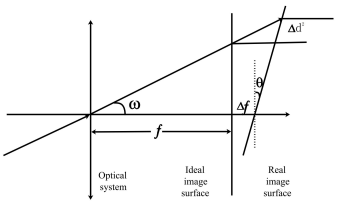

Figure 10. The Influence of CMOS Installation Error on the Coordinates of the Imaging Point

θ is the angle of the actual imaging surface from the ideal imaging surface, and Δf is the distance of the actual imaging surface from the ideal imaging surface, then the influence of the imaging surface installation error on the coordinates of the imaging point can be expressed by the following formula:

$$\Delta d_2 = \frac{cos\omega_1 + cos\theta(\Delta f + f \cdot tan\omega_1 \cdot tan\theta)}{cos(\omega_1 + \theta)} \quad (32)$$

In addition, the measurement error of the position of the imaging point of the calibration object plane is: $D = |f \cdot \tan(\omega \cdot \delta_n) - f \cdot tan\omega|$ The selected calibration method and equipment can ensure that the maximum deviation of the calibration affecting the position of the imaging point can be controlled at 0.073 pixels[18].

Due to machining and assembly inaccuracies, there is a misalignment between the camera's coordinate system and that of the cubic mirror. To quantify the angular discrepancy between the two systems, a collimator, theodolite, and a high-precision bidirectional turntable are employed. The collimator is set up on a horizontal plane, ensuring that the turntable's horizontal axis is both level and orthogonal to the collimator's optical axis. Similarly, the vertical axis is adjusted to be perpendicular, creating a coordinate system with the turntable and collimator that is congruent with the camera's coordinate system.

2.5.6 Analysis of the error between the real-world coordinates and the world coordinates of the exact bomb point location

In the detection and measurement of the binocular camera, $Q = \begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix}$ is used to represent the coordinates of the world spatial information, and the linear regression algorithm is used to deduce the position coordinates of the initial explosion point, and the error analysis method is used to measure the explosion point and the corresponding real coordinates.

$$Z_{cl}\begin{bmatrix} u_l \\ v_l \\ 1 \end{bmatrix} = M_l \begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11}^l & m_{12}^l & m_{13}^l & m_{14}^l \\ m_{21}^l & m_{22}^l & m_{23}^l & m_{24}^l \\ m_{31}^l & m_{32}^l & m_{33}^l & m_{34}^l \end{bmatrix} \begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} \quad (33)$$

$$Z_{cr}\begin{bmatrix} u_r \\ v_r \\ 1 \end{bmatrix} = M_r \begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11}^r & m_{12}^r & m_{13}^r & m_{14}^r \\ m_{21}^r & m_{22}^r & m_{23}^r & m_{24}^r \\ m_{31}^r & m_{32}^r & m_{33}^r & m_{34}^r \end{bmatrix} \begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} \quad (34)$$

The influence of various parameters in the pixel coordinate on the error is considered into the parameter coordinate transformation matrix A, the specific method uses the following formula to derive the whole parameter transformation matrix A, the parameter conversion matrix A plays an important role in the conversion between coordinates, firstly, the two pixel coordinate information composed of the binocular measurement system are consociated, and the formula of converting it into a linear coordinate equation is as follows:

$$\begin{cases} Z_{cl} \cdot u_l = M_l \cdot X_m = m_{11}^l \cdot X_m + m_{12}^l \cdot Y_m + m_{13}^l \cdot Z_m + m_{14}^l \\ Z_{cl} \cdot v_l = M_l \cdot Y_m = m_{21}^l \cdot X_m + m_{22}^l \cdot Y_m + m_{23}^l \cdot Z_m + m_{24}^l \\ Z_{cl} \cdot 1 = M_l \cdot X_m = m_{31}^l \cdot X_m + m_{32}^l \cdot Y_m + m_{33}^l \cdot Z_m + m_{34}^l \end{cases} \quad (35)$$

$$\begin{cases} Z_{cr} \cdot u_r = M_r \cdot X_m = m_{11}^r \cdot X_m + m_{12}^r \cdot Y_m + m_{13}^r \cdot Z_m + m_{14}^r \\ Z_{cr} \cdot v_r = M_r \cdot Y_m = m_{21}^r \cdot X_m + m_{22}^r \cdot Y_m + m_{23}^r \cdot Z_m + m_{24}^r \\ Z_{cr} \cdot 1 = M_r \cdot X_m = m_{31}^r \cdot X_m + m_{32}^r \cdot Y_m + m_{33}^r \cdot Z_m + m_{34}^r \end{cases} \quad (36)$$

Simultaneous derivation of the above equations yields the following system of linear equations:

$$\begin{cases} (u_l m_{31}^l - m_{11}^l)X_m + (u_l m_{32}^l - m_{12}^l)Y_m + (u_l m_{33}^l - m_{13}^l)Z_m = m_{14}^l - u_l m_{34}^l \\ (v_l m_{31}^l - m_{21}^l)X_m + (v_l m_{32}^l - m_{22}^l)Y_m + (v_l m_{33}^l - m_{23}^l)Z_m = m_{24}^l - u_l m_{34}^l \\ (u_r m_{31}^r - m_{11}^r)X_m + (u_r m_{32}^r - m_{12}^r)Y_m + (u_r m_{33}^r - m_{13}^r)Z_m = m_{14}^r - u_r m_{34}^r \\ (v_r m_{31}^r - m_{21}^r)X_m + (v_r m_{32}^r - m_{22}^r)Y_m + (v_r m_{33}^r - m_{23}^r)Z_m = m_{24}^r - u_r m_{34}^r \end{cases} \quad (37)$$

The parameter transformation matrix A is obtained by using the linear equation system, the parameters in matrix A are modeled, and error analysis is conducted with a particular emphasis on analyzing the error of the target. The transformation matrix is represented in the following form for quantitative analysis of errors.

$$M_l = \begin{bmatrix} m_{11}^l & m_{12}^l & m_{13}^l & m_{14}^l \\ m_{21}^l & m_{22}^l & m_{23}^l & m_{24}^l \\ m_{31}^l & m_{32}^l & m_{33}^l & m_{34}^l \end{bmatrix} M_r = \begin{bmatrix} m_{11}^r & m_{12}^r & m_{13}^r & m_{14}^r \\ m_{21}^r & m_{22}^r & m_{23}^r & m_{24}^r \\ m_{31}^r & m_{32}^r & m_{33}^r & m_{34}^r \end{bmatrix} \quad (38)$$

The error is therefore taken into account and the resulting matrix equation is

$$M_l + \sigma_{sum} = \begin{bmatrix} m_{11}^l & m_{12}^l & m_{13}^l & m_{14}^l \\ m_{21}^l & m_{22}^l & m_{23}^l & m_{24}^l \\ m_{31}^l & m_{32}^l & m_{33}^l & m_{34}^l \end{bmatrix} + \begin{bmatrix} \sigma_{11}^l & \sigma_{12}^l & \sigma_{13}^l & \sigma_{14}^l \\ \sigma_{21}^l & \sigma_{22}^l & \sigma_{23}^l & \sigma_{24}^l \\ \sigma_{31}^l & \sigma_{32}^l & \sigma_{33}^l & \sigma_{34}^l \end{bmatrix} \quad (39)$$

$$M_r + \sigma_{sum} = \begin{bmatrix} m_{11}^r & m_{12}^r & m_{13}^r & m_{14}^r \\ m_{21}^r & m_{22}^r & m_{23}^r & m_{24}^r \\ m_{31}^r & m_{32}^r & m_{33}^r & m_{34}^r \end{bmatrix} + \begin{bmatrix} \sigma_{11}^r & \sigma_{12}^r & \sigma_{13}^r & \sigma_{14}^r \\ \sigma_{21}^r & \sigma_{22}^r & \sigma_{23}^r & \sigma_{24}^r \\ \sigma_{31}^r & \sigma_{32}^r & \sigma_{33}^r & \sigma_{34}^r \end{bmatrix} \quad (40)$$

Combined with the above derivation process, the parameter position conversion matrix is formed.

$$A = \begin{bmatrix} u_l m_{31}^l - m_{11}^l & u_l m_{32}^l - m_{12}^l & u_l m_{33}^l - m_{13}^l \\ v_l m_{31}^l - m_{21}^l & v_l m_{32}^l - m_{22}^l & v_l m_{33}^l - m_{23}^l \\ u_r m_{31}^r - m_{11}^r & u_r m_{32}^r - m_{12}^r & u_r m_{33}^r - m_{13}^r \\ v_r m_{31}^r - m_{21}^r & v_r m_{32}^r - m_{22}^r & v_r m_{33}^r - m_{23}^r \end{bmatrix} \quad (41)$$

$$B = \begin{bmatrix} m_{14}^l - u_l m_{34}^l \\ m_{24}^l - u_l m_{34}^l \\ m_{14}^r - u_r m_{34}^r \\ m_{24}^r - u_r m_{34}^r \end{bmatrix} \quad (42)$$

The error is therefore taken into account and the resulting matrix equation is

$$A + \sigma_{sum} = \begin{bmatrix} u_l m_{31}^l - m_{11}^l & u_l m_{32}^l - m_{12}^l & u_l m_{33}^l - m_{13}^l \\ v_l m_{31}^l - m_{21}^l & v_l m_{32}^l - m_{22}^l & v_l m_{33}^l - m_{23}^l \\ u_r m_{31}^r - m_{11}^r & u_r m_{32}^r - m_{12}^r & u_r m_{33}^r - m_{13}^r \\ v_r m_{31}^r - m_{21}^r & v_r m_{32}^r - m_{22}^r & v_r m_{33}^r - m_{23}^r \end{bmatrix} + \begin{bmatrix} \sigma_{11}^{sum} & \sigma_{12}^{sum} & \sigma_{13}^{sum} \\ \sigma_{21}^{sum} & \sigma_{22}^{sum} & \sigma_{23}^{sum} \\ \sigma_{31}^{sum} & \sigma_{32}^{sum} & \sigma_{33}^{sum} \end{bmatrix} \quad (43)$$

The whole conversion relationship is represented in the form of a matrix, and the algebraic remainder is represented by $A_{ij} = (-1)^{i+j} \cdot M_{ij}$. The coordinates of the whole world are deduced in combination with various

error factors:

$$\begin{bmatrix} X_m \\ Y_m \\ Z_m \end{bmatrix} = A^{-1} \cdot B = \frac{A^*}{|A|} \cdot B = \frac{1}{|A|} \cdot \begin{bmatrix} A_{11} & \cdots & A_{13} \\ \vdots & \ddots & \vdots \\ A_{41} & \cdots & A_{43} \end{bmatrix} \cdot B = \frac{1}{\begin{vmatrix} u_l m_{31}^l - m_{11}^l & u_l m_{32}^l - m_{12}^l & u_l m_{33}^l - m_{13}^l \\ v_l m_{31}^l - m_{21}^l & v_l m_{32}^l - m_{22}^l & v_l m_{33}^l - m_{23}^l \\ u_r m_{31}^r - m_{11}^r & u_r m_{32}^r - m_{12}^r & u_r m_{33}^r - m_{13}^r \\ v_r m_{31}^r - m_{21}^r & v_r m_{32}^r - m_{22}^r & v_r m_{33}^r - m_{23}^r \end{vmatrix}} \begin{bmatrix} A_{11} & \cdots & A_{13} \\ \vdots & \ddots & \vdots \\ A_{41} & \cdots & A_{43} \end{bmatrix} \cdot \begin{bmatrix} m_{14}^l - u_l m_{34}^l \\ m_{24}^l - u_l m_{34}^l \\ m_{14}^r - u_r m_{34}^r \\ m_{24}^r - u_r m_{34}^r \end{bmatrix} \quad (44)$$

Where $Q = \begin{bmatrix} X_m \\ Y_m \\ Z_m \end{bmatrix}$ is expressed as the three-dimensional coordinates of the world's position.

Taking into account the factors affecting the error, the resolution of the visible light camera is not less than 1920*1080, the size range of 16.3cm*9.2cm, and the coverage of the entire target is 700m×700m. The pixel size corresponding to the killing range of 1000 square meters is 96*54, 96*54*0.084mm*0.084mm=36.57mm^2, which indicates the proportion of the explosion point in the actual picture.

The representation of the parameter transformation matrix, which takes into account the error factor, is shown below

$$Q + \sigma_{sum} = \begin{bmatrix} X_m \\ Y_m \\ Z_m \end{bmatrix} = (A + \sigma_{sum})^{-1} \cdot B = \frac{1}{\begin{vmatrix} u_l m_{31}^l - m_{11}^l + \sigma_{11}^{sum} & u_l m_{32}^l - m_{12}^l + \sigma_{12}^{sum} & u_l m_{33}^l - m_{13}^l + \sigma_{13}^{sum} \\ v_l m_{31}^l - m_{21}^l + \sigma_{21}^{sum} & v_l m_{32}^l - m_{22}^l + \sigma_{22}^{sum} & v_l m_{33}^l - m_{23}^l + \sigma_{23}^{sum} \\ u_r m_{31}^r - m_{11}^r + \sigma_{31}^{sum} & u_r m_{32}^r - m_{12}^r + \sigma_{32}^{sum} & u_r m_{33}^r - m_{13}^r + \sigma_{33}^{sum} \\ v_r m_{31}^r - m_{21}^r + \sigma_{41}^{sum} & v_r m_{32}^r - m_{22}^r + \sigma_{42}^{sum} & v_r m_{33}^r - m_{23}^r + \sigma_{43}^{sum} \end{vmatrix}} \cdot \begin{bmatrix} A_{11} & \cdots & A_{13} \\ \vdots & \ddots & \vdots \\ A_{41} & \cdots & A_{43} \end{bmatrix} \cdot \begin{bmatrix} m_{14}^l - u_l m_{34}^l \\ m_{24}^l - u_l m_{34}^l \\ m_{14}^r - u_r m_{34}^r \\ m_{24}^r - u_r m_{34}^r \end{bmatrix} \quad (45)$$

After the error is calculated by MSE, the direct error between the coordinates of the exact bombing point position and the world coordinates is calculated as $MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \leq 1m \approx 7 * 4$ pixels. The coverage converted to pixel coordinates can be expressed as a size of 7*4 pixels. It shows that the error between the predicted regression position coordinates of the bomb point and the real-world point is very controllable, which can meet the requirements of the required coordinates and error range[19].

Then the matrix form of the Taylor expansion of $f(x_1, x_2, ..., x_n)$ at $x^{(0)}$ is expressed as $f(x) = f(x^{(0)}) + \nabla f(x^{(0)})^T \Delta X + \frac{1}{2}\Delta X^T G(X^{(0)})\Delta x + \ldots$, $\nabla f(x^{(0)}) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}\right]^T_{x^{(0)}}$ and this formula represents $f(x)$ in $x^{(0)}$ he partial derivative at ((0)), where $x^{(0)}$ represents only the position of the target coordinate point. The world coordinates are deflected by the world coordinates of the target coordinates.

$$G(x^{(0)}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots \ \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}_{x^{(0)}} \quad (46)$$

is expressed as a Hesse matrix at $f(x)$ at $x^{(0)}$. The Hesse matrix is a symmetric matrix of n*n order composed of the second-order partial derivative of the objective function f at the point x.

2.5.7 Optimize the error analysis function

Before using the camera, the internal parameter K of the camera can be obtained after calibrating the camera, and the pixel coordinates of the feature points can be obtained through feature point matching, and then the normalized spatial coordinates corresponding to the pixels can be obtained according to the above model combined with the transformation matrix, namely:

$$\frac{X}{Z} = X = (u - cx)/f_x \quad (47)$$

$$\frac{Y}{Z} = Y = \frac{u - cy}{f_y} \quad (48)$$

$$Z = 1 \quad (49)$$

When the drone is shooting, the model of the camera is:

$$ZP_{uv} = Z\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K(RP_w + t) = KTP_w \quad (50)$$

Expressing $T = (R|t)$ in the form of Lie algebra, then

$$s_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = Kexp(\xi) \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} \quad (51)$$

Then the optimization equation for the whole can be expressed as

$$\xi^* = \underset{\xi}{argmin} \frac{1}{2}\sum_{i=1}^{n} \left\| u_i - \frac{1}{s_i} Kexp(\hat{\xi}) P_i \right\|_2^2 \quad (52)$$

G-N and L-M are solved by optimizing the equations.

3. Experiment

The exact pixel coordinates are [100.001 199.99911 1], [149.99893 299.99997 1], and the corresponding exact world coordinate location points are [7700, 73407, 0.0], [-66029, -605035, 0.0].

The world coordinates of the target calculated by vector space to coordinate function $P_c = R(P_w - C)$ are [7700.001290525472, 73407.76817339347, 0.0], [-66029.50444612742, -605035.9517151915, 0.0]. The corresponding

internal reference matrix is

$$R = \begin{bmatrix} -0.91536173 & 0.40180837 & 0.02574754 \\ 0.05154812 & 0.18037357 & -0.98224649 \\ -0.39931903 & -0.89778361 & -0.18581953 \end{bmatrix}$$

The internal reference matrix, the world coordinate position points detected by the UAV and the internal reference matrix of the camera are combined with the pixel matrix to generate the error vector function $V = \frac{1}{|P_w - C|} \cdot [|P_c| + |P_w - 1| \cdot |P_c| \cdot \frac{1}{|P_w - C|}] \cdot cos\theta$ to obtain the offset of the final world coordinate position $\begin{vmatrix} \Delta x \\ \Delta y \\ \Delta z \end{vmatrix} = \begin{vmatrix} 0.001290525472 & 0.50444612742 \\ 0.76817339347 & 0.9517151915 \\ 0 & 0 \end{vmatrix}$。

3.1 Real-time Solution Time for the Target

In the real-time calculation process of single-target video, under the premise of the current downlink bandwidth of 80 Mbs, the 25 fps real-time image information displayed on the ground integrated display console requires 5 Mbs bandwidth. When transmitting high-speed real-time information for 5 seconds, the theoretical calculation formula is described as follows: $\frac{(1920\times1080\times8)/2\times1000fps\times5s}{(75\times1024\times1024)} = 527s = 9min$, and the additional amount of data to be transmitted is 20%-30% for subcontracted transmission, and the estimated return time is about 12 minutes.

When the data is transmitted over a wired optical fiber channel, the transmission bandwidth rate is 500Mbs, and the transmission uses three USB3.0 data transmission lines. The 25fps real-time screen information displayed on the ground integrated display console requires 5 Mbs bandwidth. Assuming that the length of the video to be measured is 5 minutes, when transmitting 300 seconds of high-speed real-time information, the theoretical calculation formula is described as follows $\frac{(1920\times1080\times8)/2\times1000fps\times300s}{(1485\times1024\times1024)} = 1598s = 26\text{min}$ and the subpackage transmission needs to increase the transmission data by 20%-30%, and the transmission process includes three camera data volumes, so the estimated backhaul time is about 30 minutes.

By analyzing two distinct channel transmission modes, along with the solution time for detecting and identifying multiple explosion points, we can enhance the real-time calculation of the bomb point [20]. During this process, three UAVs simultaneously segment and track the target bomb point. The real-time solution time encompasses several components: the data transmission duration, the tracking and segmentation phase of the bomb point target, the time taken by the regression algorithm to revert to the initial bomb point coordinates, the pixel positioning of the bomb point target, and the time required for converting to world coordinates.

Taking the transmission process of short-distance wireless channel as an example, the transmission time of the three UAVs is 12 minutes when transmitting and solving the video for 5s, the tracking time of the target cutting and tracking algorithm is 7 minutes, the time for the regression algorithm to reverse calculate the position of the initial explosion point is 5 minutes, and the time for world coordinate conversion and target positioning is 33 seconds. Real-time solution time = transmission + detection + pushback + positioning = 9min + 0.55min + 0.55min + 0.017min = 15.617min

Taking the transmission of wired USB3.0 channel as an example, the SAT transmission frame rate is 150 frames per second, and 900,000 frames of pictures need to be transmitted, so the tracking algorithm takes 100 minutes. The linear regression algorithm also takes 100 minutes, and the real-time solution time = transmission + detection + pushback + positioning = 30 min + 99.8 min + 99.8 min + 1.53 min = 231.13 min.

4. Conclusions

This paper introduces the application of explosion point tracking and measurement under binocular vision, which is supported by the project and data provided by Group A. This paper collected the bomb point measurement data at different heights on the top of the mountain by collecting the data of the pushback, and noticed that there were problems such as ghosting and occlusion of the explosion point in the data video. In order to solve the problem of explosion occlusion and camera ghosting caused by equipment problems, this article used the remote sensing image reconstruction network to reconstruct some explosion point images with ghosting. The processed data images are annotated by the labelimg annotation tool combined with the data processing tool OpenCV, and the multi-target tracking network is used to complete the training of the multi-target

tracking model, so as to realize the tracking processing of the explosion points. In this paper, this article focus on the task of regression of the world coordinate point of the initial bomb point through the bomb point location algorithm and image regression network to realize the bomb point measurement.

Acknowledgments

This work was produced by the Computer Vision and Artificial Intelligence Team of Yan'an University, which was supported by:

A. Major Special Project of the Ministry of Science and Technology - Theory and Method of Security and Reliability of Green Internet of Things Empowered by Artificial Intelligence, Project No.: 2022YFE0138600, Sub-project Leader. 2023.01-2026.12, 8.5 million yuan, under research.

B. Presided over the National Natural Science Foundation of China's "Research on the Theory and Application of Cross-media Collaborative Deep Security Situational Awareness in Artificial Intelligence", project number: 62266045, January 2023-December 2026, 330,000 yuan, under research.

References

[1] Ash, Jordan T. and Ryan P. Adams. “On Warm-Starting Neural Network Training.” arXiv: Learning (2019): n. pag.

[2] Irwan Bello, William Fedus, Xianzhi Du, Ekin D Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. arXiv preprint arXiv:2103.07579, 2021.

[3] Cubuk, Ekin Dogus et al. “Randaugment: Practical automated data augmentation with a reduced search space.” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019): 3008-3017.

[4] Dai T, Cai J, Zhang Y ,et al. Second-order Attention Network for Single Image Super-Resolution // 2019 IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR).IEEE, 2019. DOI:10.1109/CVPR.2019.01132.

[5] Dai, Tao, Hua Zha, Yong Jiang and Shutao Xia. “Image Super-Resolution via Residual Block Attention Networks.” *2019 IEEE / CVF International Conference on Computer Vision Workshop (ICCVW)* (2019): 3879-3886.

[6] Zamir S W, Arora A, Khan S,et al. Restormer: Efficient Transformer for High Resolution Image Restoration//2021. DOI:10.48550/arXiv.2111.09881.

[7] Zhou D, Yu Z, Xie E,et al. Understanding The Robustness in Vision Transformers. 2022. DOI:10.48550/arXiv.2204.12451.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv, abs/2010.11929*.

[9] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In ICLR, 2022.

[10] Tu Z, Talebi H, Zhang H, et al. MaxViT: Multi-Axis Vision Transformer. arXiv e-prints, 2022. DOI: 10.48550/arXiv.2204.01697.

[11] Tu Z, Talebi H,Zhang H, et al. MAXIM: Multi-Axis MLP for Image Processing. 2022. DOI: 10.48550/arXiv.2201.02973.

[12] He, Tong, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie and Mu Li. “Bag of Tricks for Image Classification with Convolutional Neural Networks.” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018): 558-567.

[13] He X, Mo Z, Wang P, et al. ODE-Inspired Network Design for Single Image Super-Resolution // 2019 IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.DOI:10.1109/CVPR.2019.00183.

[14] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv: 1606.08415, 2016.

[15] Hu J, Shen L, Sun G, et al. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP (99). DOI: 10.1109/TPAMI.2019.2913372.

[16] Huang G, Sun Y, Liu Z, et al. Deep Networks with Stochastic Depth. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-46493-0_39.

[17] Kim J, Lee J K, Lee K M .Accurate

Image Super-Resolution Using Very Deep Convolutional Networks. IEEE, 2016. DOI: 10.1109/CVPR.2016.182.

[18] Li Z, Yang J, Liu Z, et al. Feedback Network for Image Super-Resolution // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 0 [2024-03-07]. DOI: 10.1109/CVPR.2019.00399.

[19] Liang, Jingyun et al. "SwinIR: Image Restoration Using Swin Transformer." *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2021): 1833-1844.

[20] Lim, Bee, Sanghyun Son, Heewon Kim, Seungjun Nah and Kyoung Mu Lee. "Enhanced Deep Residual Networks for Single Image Super-Resolution." *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017): 1132-1140.

