Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN: 2957-3920

Volume 2, Issue 2, February 2023

Research on Chili Pepper Recognition and Localization Method Based on Binocular Vision

Wang Yanqing[1], Ye Zhenhuan[1], Zhu Zhenglong[1*], Tang Maoyin[2], Wei Meiling[1], Xu Huan[1]

1. College of Engineering, Zunyi Normal University, Zunyi 56300, China;

2. Department of Automobile, Guizhou Aerospace Vocational and Technical College, Zunyi 56300, China

Correspondence author: Zhu Zhenglong

Abstract: In this paper, the collected images are preprocessed by grayscale, image segmentation, binarization, etc., and then the parameters of the camera and the four types of coordinate transformation are analyzed by calibrating the camera, and then the AlexNet network and transfer learning pair are used The pepper was identified, and the correct rate of 90% was obtained, and the loss function was trained by cross-entropy with a loss value of 0.4, and finally the stereo matching and three-dimensional reconstruction were briefly introduced.

Keywords: Binocular Vision; Pepper Identification; Camera Calibration; Neural Networks

1 INTRODUCTION

From 2016 to 2018, the planting area of Zunyi pepper reached 133,300 hectares, 135,500 hectares and 150,000 hectares respectively, and the output value of dried pepper was 6.02 billion yuan, 7.18 billion yuan and 8 billion yuan, respectively, and its planting scale accounted for 3%, 10% and 40% of the global, national and provincial respectively[1] , of which Chaotian pepper accounts for more than 70% of the total pepper planting area in Zunyi City, and in 2020, the planting area of Chaotian pepper in Zunyi City will reach 108,000 hectares, the output will reach 2,037,700 tons, and the planting output value will be 7.050 billion yuan , ranking first in the main pepper producing areas in the country [2].

As one of the key supporting industries for local economic growth in Zunyi, the chili pepper industry is one of the key supporting industries. It is of great significance to study a binocular vision system with pepper recognition and positioning to realize the identification and positioning of pepper, and solve the problems of untimely picking and economic loss caused by lack of labor due to short picking cycle of pepper. Binocular vision is used for the real spatial position positioning of targets, and has been used in many aspects, such as the research of binocular vision on dynamic materials, the research on vehicle detection and tracking and ranging methods, etc. [3-4] 。In this paper, the AlexNet network model is used for the identification of peppers [5], and the binocular camera is placed in parallel structure, and the position of the target image in real space can be obtained by combining the calculation of known conditions to achieve positioning. First, the acquired images are preprocessed for grayscale, image segmentation, and binarization. The internal and external parameters of the camera were obtained by camera calibration[6], and the camera calibration was completed by Zhang Youzheng's calibration method [7]. After the recognition is completed, the stereo matching is carried out, and the identification object is pepper, and the method suitable for matching the object features such as the target edge and contour is used to realize the three-dimensional matching and three-dimensional reconstruction of pepper [8]。

2 BINOCULAR VISUAL RANGING PRINCIPLE AND IMAGE ACQUISITION AND PREPROCESSING

2.1 *BINOCULAR VISUAL RANGING PRINCIPLE*

As shown in Figure 1, the schematic diagram of binocular parallel vision stereoscopic imaging, the baseline distance *B* is the distance from the projection center of the two cameras, and the distance from the camera to the imaging is the focal length f For the same feature of a space object, the coordinates of the image acquired by the left and right cameras are marked respectively, $P(xc, yc, zc,)$ $P_{left} = (x_l, y_l)$ $P_{rightt} = (x_r, y_r)$ because the two cameras are kept on the same plane. The geometric relationship of the triangle is obtained: $y_l = y_r = Y$

$$x_l = f\frac{x_c}{z_c} \quad x_r = f\frac{(x_c - B)}{z_c} \quad f\frac{y_c}{z_c} = Y \qquad (1)$$

Let , $D = x_l - x_r$ D is parallax, which can be obtained by the calibration of the camera, and the three-dimensional coordinates of the feature point P in the camera coordinate system can be calculated:

$$x_c = \frac{B \cdot x_l}{D} \quad y_c = \frac{B \cdot Y}{D} \quad z_c = \frac{B \cdot f}{D} \qquad (2)$$

In addition, let the image depth be d, and the image depth d can be obtained according to the similarity triangle principle, that is, $.d = \frac{B \cdot f}{D}$

2.2 Image Acquisition And Preprocessing

2.2.1 Image Acquisition

The binocular camera is shown in Figure 1, and two cameras can be adjusted

The distance of the image head can be fixed by a tripod in practical applications, and the two cameras are placed horizontally, and the relevant parameters of the cameras are shown in Figure 3.

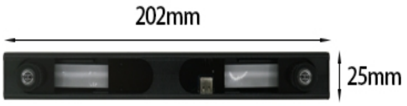


Figure 1 The camera used in the project

The camera is used to collect the pepper in the field, and the image of the monocular and binocular camera can be obtained by adjusting different resolutions. A total of 2 00 photographs were collected and, owing to space limitations, six illustrative images are listed, as shown in figure 2.

Figure 2 Images of some parts were acquired

2.2.2 Image Preprocessing

The collected images contain a lot of useless information such as stems and leaves, natural environment, etc., so image preprocessing needs to be used to remove irrelevant information and obtain purer feature images.

1 Grayscale

The acquired image is a color image, consisting of three channels of red, green, and blue, that is, the RGB diagram, which is represented in the computer as a three-dimensional array The purpose of grayscale is to convert the three-dimensional array into a one-dimensional array, reduce the difficulty of calculation, this paper uses the weighted average method for image grayscale, the formula is as follows, where is the converted grayscale value, , $f(x, y) R(x, y) G(x, y)$, respectively the image in red$B(x, y)$ Value on the green and blue channel:

$$f(x, y) = 0.299R(x, y) + 0.587G(x, y) + \ 0.114B(x, y)$$

The original image is grayscaled as shown in Figure 3.

Volume 2, Issue 2, February 2023

FIGURE 3 IMAGE GRAYSCALE

2 Median filtering

When the collected images are imported into the computer due to electromagnetic interference and environmental influences, they often contain a lot of noise, and the median filtering method is used to denoise so that the surrounding pixel values are close to the true value, thereby eliminating isolated noise points.

3. Image binarization

Image binarization is the conversion of an image into two gray levels of only 0 and 255, which makes it easy to distinguish between target and useless information, as shown in Figure 4 。

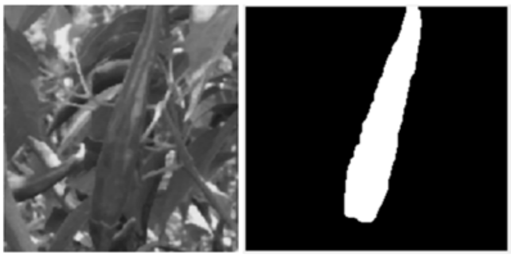


FIGURE 4 IMAGE BINARIZATION

3 BINOCULAR CAMERA CALIBRATION AND PEPPER IDENTIFICATION

3.1 CALIBRATION OF BINOCULAR CAMERAS

The calibration of the camera determines the internal and external parameters of the camera, and the position of each point in the three-dimensional space in the image coordinate system can be obtained, so as to determine the three-dimensional spatial position of the object through the two-dimensional image collected by the camera. The four coordinate systems and the mutual conversion relationship between them can obtain the internal and external parameters of the camera, which are introduced below:

Conversion of four coordinate systems

1 Conversion of world coordinate system to camera coordinate system

The world coordinate system is a three-dimensional coordinate system, (x_w, y_w, z_w) which is proposed to describe the position of the target in the real world, the coordinates reflect the real position of the object, the camera coordinate system is a three-dimensional coordinate system, the optical axis of the camera is the z-axis, the (x_c, y_c, z_c) horizontal right direction is the x-axis, and the vertical downward direction is y axis, which is the bridge between the world coordinate system and the image coordinate system, as shown in Figure 5.

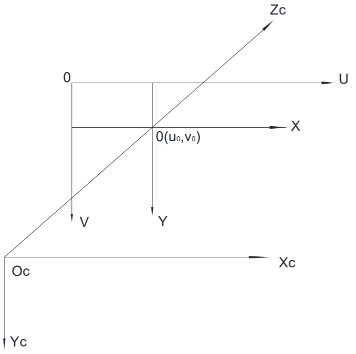


FIGURE 5 SEVERAL COORDINATE SYSTEMS

The conversion of the world coordinate system to the camera coordinate system is a rigid body transformation, which is realized by the rotation matrix R and the translation matrix T, and the conversion relationship between the two is expressed in the form of a matrix as follows:

$$\begin{pmatrix} X_c \\ y_c \\ Z_c \end{pmatrix} = R * \begin{pmatrix} X_w \\ y_w \\ Z_w \end{pmatrix} + T = \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_w \\ y_w \\ Z_w \\ 1 \end{pmatrix} \tag{3}$$

$$= LW * \begin{pmatrix} X_w \\ y_w \\ Z_w \\ 1 \end{pmatrix}$$

where R is the 3×3 matrix, which is the result of multiplying the three rotation matrices, representing the three directions of the rotation of the coordinate system. The translation matrix T is a 3×1 matrix, and LW is the rotation matrix R and the translation matrix T is the external parameter matrix of the camera.

2 Conversion of camera coordinate system to image physical coordinate system

The image physical coordinate system (X, Y) is introduced according to the projection relationship, which is convenient for further obtaining pixel coordinates, and the coordinate origin is the intersection point of the image pixel coordinate system (u, v) and the optical axis of the camera, as shown in Figure 5. The conversion of the camera coordinate system to the physical coordinate system of the image can be obtained by the geometric projection relationship:

$$\frac{x}{f} = \frac{x_c}{z_c} \qquad \frac{y}{f} = \frac{y_c}{z_c} \tag{4}$$

Sorted out:

$$z_c \cdot x = x_c \cdot f$$
$$z_c \cdot y = y_c \cdot f \qquad (5)$$

Represent them in matrix form:

$$Z_c \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{pmatrix} \qquad (6)$$

3 Conversion of image physical coordinate system to image pixel coordinate system

The image physical coordinate system (u, v) is a two-dimensional coordinate system, for pixels, the image collected by the camera is saved in the form of numerical values in the computer, the coordinate origin is in the upper left corner of the image, the U-axis direction is horizontal to the right, and the v-axis direction is horizontal down, as shown in Figure 9 shown.

The image physical coordinate system and the image pixel coordinate system are different representations in the two-dimensional plane, the difference is that the image physical coordinate system is a continuous concept, and the image pixel coordinate system is a discrete concept, and the two can be obtained by digital discretization:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \qquad (7)$$

where is the unit length of x direction including the number of pixels, and the unit length of y direction includes the number of pixels, due to the different coordinate origin of the two coordinate systems, the origin of the physical coordinate system of the image is:$\alpha\beta$ (u_0, v_0)

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha & 0 & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \qquad (8)$$

Merge the transformation matrix of the camera coordinate system to the physical coordinate system and the transformation matrix of the image physical coordinate system to the pixel coordinate system:

$$Z_c \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha & 0 & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{pmatrix} g\, Z_c \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} \alpha & 0 & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} f_c & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = KP_c \qquad (9)$$

Among them, K is the internal parameter matrix of the camera, which is obtained by the camera calibration

The internal and external parameters of the camera, which can then determine the position of points in three-dimensional space in two-dimensional images, are calibrated by Matlab and Zhang Zhengyou chessboard method calibrates the camera, and the calibrated checkerboard is shown in Figure 6.

The monocular and binocular cameras were used to acquire chessboard images for monocular and bicular targeting, the acquired images were shown in Figure 7 and Figure 8, and the calibration effect and calibration results were shown in Figure 9, 10 and Figure 11, respectively The internal and external parameters of the calibrated camera are shown in Table 1.

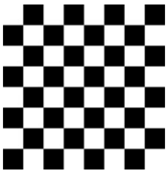


FIGURE 6 THE CHESSBOARD USED IN ZHANG ZHENGYOU'S CALIBRATION METHOD

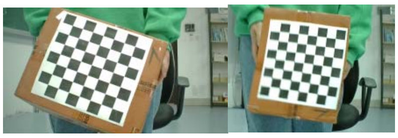


FIGURE 7 MONOCULAR LEFT AND RIGHT CHECKERBOARD CHARTS

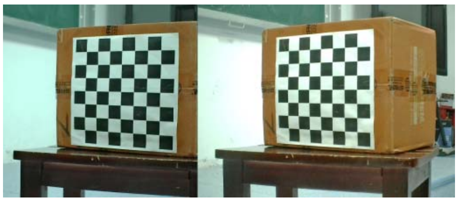


FIGURE 8 BINOCULAR CHECKERBOARD IMAGE

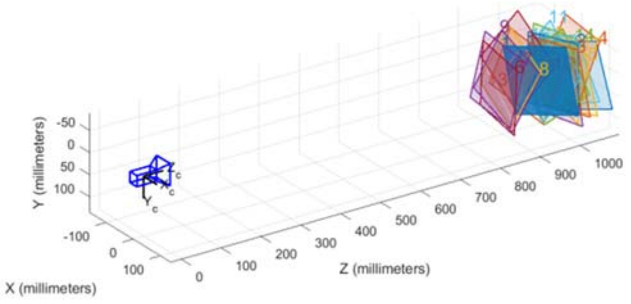


FIGURE 9 SINGLE-TARGET RENDERING

TABLE 1 CAMERA CALIBRATION RESULTS

<table>
<tr><td colspan="2">formeter</td><td>Left camera</td><td>Right camera</td></tr>
<tr><td colspan="2">intrinsic parameter</td><td>$\left(\begin{matrix}3075.1 & 33.8992 & 509. \\ 0 & 2897.8 & 28. \\ 0 & 0 & \end{matrix}\right.$</td><td>$\left(\begin{matrix}1945.6 & 6.7406 & 682 \\ 0 & 1883.0 & 404 \\ 0 & 0 & \end{matrix}\right.$</td></tr>
<tr><td colspan="2">aberration</td><td>$k_1 = -1.9864$、$k_2 = 8.5$
$P_1 = 0.2331$、$P_2 = 0.02$</td><td>$k_1 = -1.9864$、$k_2 = 8.$
$P_1 = 0.2331$、$P_2 = 0.0$</td></tr>
<tr><td rowspan="2">external parameter</td><td>rotation matrix</td><td colspan="2">$\begin{pmatrix}0.9973 & -0.363 & 0.0645 \\ 0.0369 & 0.9993 & -0.0078 \\ -0.0642 & 0.0102 & 0.9979\end{pmatrix}$</td></tr>
<tr><td>Translation matrix</td><td colspan="2">$(110.0411 \quad 3.0849 \quad -7.6937)$</td></tr>
</table>

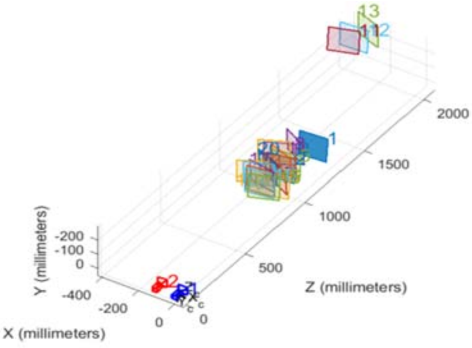
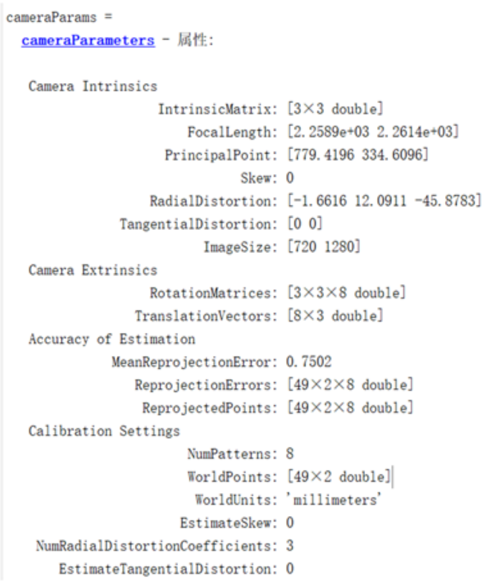


FIGURE 10 DOUBLE-TARGET RENDERING



```
cameraParams =
  cameraParameters - 属性:

   Camera Intrinsics
                    IntrinsicMatrix: [3×3 double]
                        FocalLength: [2.2589e+03 2.2614e+03]
                     PrincipalPoint: [779.4196 334.6096]
                               Skew: 0
                   RadialDistortion: [-1.6616 12.0911 -45.8783]
               TangentialDistortion: [0 0]
                          ImageSize: [720 1280]
   Camera Extrinsics
                   RotationMatrices: [3×3×8 double]
                 TranslationVectors: [8×3 double]
   Accuracy of Estimation
              MeanReprojectionError: 0.7502
                 ReprojectionErrors: [49×2×8 double]
                  ReprojectedPoints: [49×2×8 double]
   Calibration Settings
                        NumPatterns: 8
                        WorldPoints: [49×2 double]
                         WorldUnits: 'millimeters'
                       EstimateSkew: 0
    NumRadialDistortionCoefficients: 3
       EstimateTangentialDistortion: 0
```

FIGURE 11 CALIBRATION RESULTS

3.2 *CAYENNE PEPPER IDENTIFICATION*

The acquired images are preprocessed for pepper identification. In this paper, an improved convolutional neural network is used for image recognition.

Convolutional neural network is a type of neural network, it has deep learning, local connection, weight sharing and other characteristics, can accurately identify the characteristics of pepper, mainly divided into convolutional layer, pooling layer, fully connected output layer. This paper uses the AlexNet model to identify peppers. In the AlexNet model, the input is 2 24×224×3 pixels, including 5 The layers of convolution and the three layers are fully connected, and the convolutional layers are 1 1×11, 5×5 and three 3×3, respectively of convolution kernels. The pooling method adopts the Mega pooling method; The activation function is the ReLu function; The model is output as a neuron of class name length through the softmax function, and the activation function adopts the softmax probability value; The model is optimized using the SGD optimizer, the loss function is the cross-entropy loss function, and the network model is shown in Figure 12.

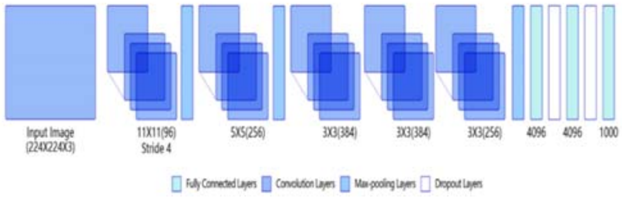
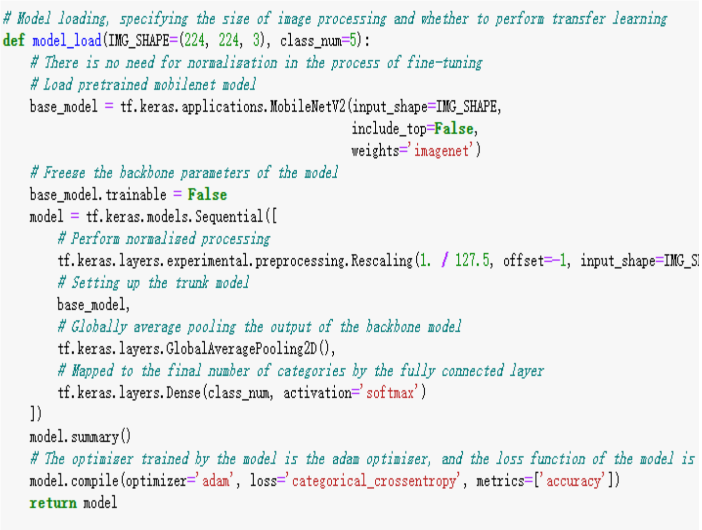


FIGURE 12 ALEXNET MODEL NETWORK STRUCTURE

Transfer learning is introduced into the CNN model, normalized the model, and then pooled by the Average pooling method, and

Volume 2, Issue 2, February 2023

the model can be built by outputting the number of classifications through the full connection layer, and the transfer learning model is shown in Figure 13.



```
# Model loading, specifying the size of image processing and whether to perform transfer learning
def model_load(IMG_SHAPE=(224, 224, 3), class_num=5):
    # There is no need for normalization in the process of fine-tuning
    # Load pretrained mobilenet model
    base_model = tf.keras.applications.MobileNetV2(input_shape=IMG_SHAPE,
                                                   include_top=False,
                                                   weights='imagenet')
    # Freeze the backbone parameters of the model
    base_model.trainable = False
    model = tf.keras.models.Sequential([
        # Perform normalized processing
        tf.keras.layers.experimental.preprocessing.Rescaling(1. / 127.5, offset=-1, input_shape=IMG_S
        # Setting up the trunk model
        base_model,
        # Globally average pooling the output of the backbone model
        tf.keras.layers.GlobalAveragePooling2D(),
        # Mapped to the final number of categories by the fully connected layer
        tf.keras.layers.Dense(class_num, activation='softmax')
    ])
    model.summary()
    # The optimizer trained by the model is the adam optimizer, and the loss function of the model is
    model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
    return model
```

FIGURE 13 TRANSFER LEARNING MODEL

According to the convolutional neural network model training dataset, the dataset is divided into 5 classes, each class has 110 pictures, of which 1 00 are used for training and 10 are used for testing, and the final model training results are shown in Figure 14.

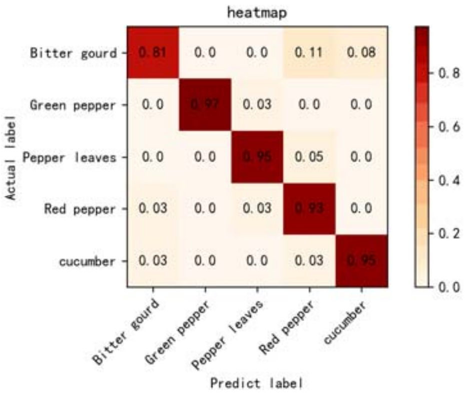


FIGURE 14: PEPPER IDENTIFICATION ACCURACY

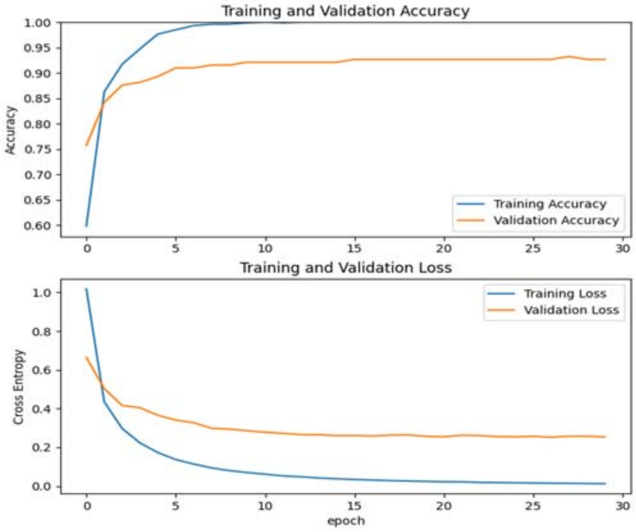


FIGURE 15: CROSS-ENTROPY TRAINING LOSS FUNCTION

Cross-entropy is to determine the closeness of the actual output to the desired output, that is, the smaller the value of the cross-entropy, the closer the two probability distributions, and the cross-entropy loss function trained by the model is shown in Figure 15. As can be seen from the figure, the correct rate of recognition on the validation set is 90%, cross-entropy data is 0.4.

4 STEREO MATCHING AND 3D RECONSTRUCTION

Stereo matching is determined by a specific point in an image in another image corresponding to the position of the point, three-dimensional reconstruction refers to the restoration of the real vision of space on the basis of matching, the spatial target point in the left and right images of the image, combined with the analysis of the corresponding coordinates and camera calibration, can carry out three-dimensional reconstruction of spatial points. The world coordinate system is established with this point, and the world coordinates of any point in space are obtained by the least squares method to complete the three-dimensional reconstruction.

5 CONCLUSION

In this paper, the acquired images are preprocessed by grayscale, image segmentation, binarization, etc., and then the parameters of the camera and the four types of coordinate conversion are analyzed by calibrating the camera, and secondly, it is analyzed AlexNet network and transfer learning were used to identify peppers, and 90% accuracy was obtained, and the loss function was trained by cross-entropy, and the loss value was 0.4, and finally briefly introduces stereo matching and three-dimensional reconstruction.

FUNDS

College Students Innovation Training Project (202110664030), Zunshi Kehe HZ [2020] No. 10, Zunshi Ke Talent [2021] No. 2, Zunhong Kehe Shizi (2022) No. 06, Zunshi Kehe HZ Zi [2022]. No. 41.

REFERENCES

[1]PENG Siyun, LUO Yi, XIE Ting, et al. Analysis and suggestions of "diamond model" of pepper industry competitiveness in Zunyi City[J]. Chili Magazine, 2019, 17(1):31-36

[2]MAO Dong, JIANG Hua, et al. Development and current situation of Chaotian pepper industry in Zunyi[J]. China Vegetables, 2021, 2021(2):7-9

[3]Ma Mengjie. Research on dynamic material identification method based on binocular stereo vision[D]. Chengdu: Southwest Jiaotong University, 2018

[4]ZHU Lei. Research and implementation of vehicle detection, tracking and ranging method based on binocular vision[D]. Changsha: Hunan University, 2020

[5]DING Wenkuan Chili Detection and Recognition Based on Convolutional Neural Network and Machine Vision [D]. Tianjin: Tianjin University of Technology, 2017

[6]Leonard Brosgole. Dynamic size perception as a function of target location in egocentric space [J]. Bulletin of the Psychonomic Society， 1993， 31 (4):

[7]Zhang Z. A Flexible New Technique for Camera Calibration [J]. Tpami， 2000， 22 (11): 1330-1334.

[8]WU Nianxiang, ZOU Huadong. Research on robot pose calibration technology based on stereo vision[J]. Journal of Jinggangshan University (Natural Science Edition), 2015,03:71-75.

[9]DAI Yanfang,SONG Zhigang An Improved Image Segmentation Algorithm Implemented by MatLab[J]. Microprocessor, 2014, 05: 31-33