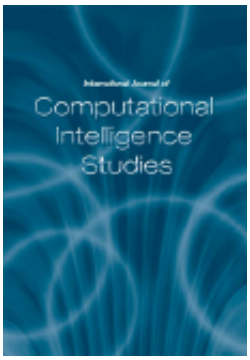
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Computational Intelligence Studies

ISSN online: 1755-4985 - ISSN print: 1755-4977
https://www.inderscience.com/ijcistudies

New media interaction in art design based on deep learning binocular stereo vision

Yongchao Liu, Ziping Zhao

DOI: 10.1504/IJCISTUDIES.2023.10060753

Article History:

Received:	01 September 2022
Last revised:	05 December 2022
Accepted:	22 September 2023
Published online:	05 April 2024

New media interaction in art design based on deep learning binocular stereo vision

Yongchao Liu* and Ziping Zhao

School of Art,
Qingdao Agricultural University,
Qingdao, 266109, China
Email: yongchaoliuqau@163.com
Email: zpzhao@qau.edu.cn
*Corresponding author

Abstract: Advances in computer vision technology and the widespread promotion of artworks have led to a profound connection between the two. Highly imitated fakes are often found in art works, which damage consumers as well as creators' own interests, so the study improves the accuracy of art works authenticity identification through the development of computer vision technology. The research is based on machine binocular stereo vision technology, the convolution neural network structure of single shot multi-box detector (SSD) is fused and trained to establish a high-precision object recognition model, which recognises objects by matching the feature points of binocular images. In the experiment, the SSD model has a stable loss value of 0.9 in the loss function performance test, and the overlap rate of the model is around 0.85, which indicates that the model has a high accuracy of object recognition. In the feature point matching algorithm, the parallax value of the multi-feature point fusion matching algorithm is stable in the range of 67 to 75 after filtering. The model proposed in the study has high accuracy in object recognition, which can play an important role in artwork authenticity identification.

Keywords: single shot multi-box detector; SDD; convolutional neural network; object recognition; art design; multi-feature point fusion matching algorithm.

Reference to this paper should be made as follows: Liu, Y. and Zhao, Z. (2023) 'New media interaction in art design based on deep learning binocular stereo vision', *Int. J. Computational Intelligence Studies*, Vol. 12, Nos. 3/4, pp.238–254.

Biographical notes: Yongchao Liu obtained his BE in Animation from Tiangong University in 2011. He obtained ME and PhD in Visual Communication Design from Kangwon National University in 2015 and 2019, respectively. Presently, he is working as a Lecturer in College of Art, Qingdao Agricultural University. His areas of interest are media interaction design, digital media design and visual communication design.

Ziping Zhao obtained his BE in Art Design from Harbin University of Science and Technology in 2003. He obtained ME in School of Animation and Media from Qingdao Agricultural University in 2014. Presently, he is working as an Associate Professor in College of Art, Qingdao Agricultural University. His area of interest is visual communication design.

1 Introduction

At present, the identification of the authenticity of art works is usually carried out by human eyes. However, with the continuous improvement of the imitation degree, the identification of authentic and fake works is becoming more and more difficult. Therefore, the authenticity of works of art can be identified through machine vision technology (Viguerie et al., 2020). Numerical stereo vision technology is the use of the human vision binocular imaging principle, the object in binocular imaging to form the position of the left and right two images, through the image matching algorithm for parallax to get the three-dimensional coordinates in space, in the acquisition of the object three-dimensional coordinates using the characteristics of the convolutional neural network structure for object recognition (Akundi and Reyna, 2021; Lin et al., 2021). However, it has become a key concern for research in improving the image matching speed as well as accuracy. Image matching techniques mainly include image feature point matching and image stereo matching. The image stereo matching algorithm uses image feature point matching algorithm because of its long running time and large error, which gets faster running speed and accuracy but generates insufficient feature points in recognising objects (Wang et al., 2020b). The study proposes multi-feature point fusion matching (MPFM) algorithm to solve the above problems of insufficient feature points and missing location information. The aim is to improve the accuracy of object recognition in order to promote the normal development of art and design and to protect the interests of consumers and designers.

2 Related works

At present, many researchers are related to this field and have made many researches and improvement methods on machine vision recognition technology and SSD neural network structural model. Zhang et al. (2020) constructed a panoramic photographic image acquisition system through machine vision technology to collect and analyse the images of corn ears, so as to achieve the purpose of breeding excellent maize varieties. The image was processed by lab colour space and adaptive threshold, and the relevant indicators were obtained. The scale invariant feature transform (SIFT) algorithm was used to extract local images. The experimental results show that the recognition accuracy of the proposed model is improved by 7.25% compared with that of ordinary panoramic images. Gorai et al. (2021) studied the intelligent coal representation model of machine vision technology. This model captures samples from different angles, converts images to different colour spaces for feature extraction, and combines stepwise linear regression algorithm to optimise features extracted from images, and applies this model to experiments. The R-square value of fixed coal is divided into 0.89, 0.92, 0.84, and 0.92. This model demonstrates the potential of machine vision technology in object recognition function. Talaat et al. (2020) used image processing technology and wireless power transmission (WPT) technology to accurately identify and locate electric vehicles through machine vision technology combined with Internet of things, realised automatic charging of electric vehicles, and obtained the optimal WPT value of electric vehicles to overcome the battery charging time problem. Souto et al. (2020) found that machine vision technology is difficult to apply in the harsh environment of hot rolling shop, so they

improved the machine vision technology and proposed a system model combining manual digital vision and optical flow measurement to accurately identify the quality and state of processed materials. This model can more safely and accurately identify the state and material quality during hot rolling mill operation in experiments. Zhao et al. (2021) proposed the automatic tool setting technology using machine vision technology on the damaged surface of potassium dihydrogen phosphate crystal (KDP). The distance between the tool tip and the crystal was obtained by using the ranging ability of machine vision, and the tool tip could be set quickly and accurately through the recognition rate of the tool tip. In the experimental results, the automatic tool setting time is 130 s, and the tool setting accuracy is better than 1.359 μm, which indicates that the tool setting technology under machine vision has more reliable security, accuracy and positive efficiency.

Wang et al. (2020a) proposed feature enhanced SSD network (FESSD), which can fuse the context information of multi-scale features using ASFP and several atrous ratios. After introducing a new area-weighted loss function, multiple targets can be treated equally, effectively improving the detection accuracy of the model. The validity of the model was verified by AIR-OBJ dataset. Jang et al. (2019) proposed a single-shot face correlation analysis model, in which FACE-SSD is a simple single-person face detection architecture that can detect multiple categories simultaneously without modifying its network structure to achieve real-time performance. In the experimental results, the model achieves 95.67% accuracy of face smile and 90.29% accuracy of attribute detection and recognition. Gamanayake C's research team constructed SSD-MobileNet and SSD-SqueezeNet neural network architectures in the CNN model, which showed better performance in the filtering pruning method. By benchmarking the datasets, the results show that the proposed network architecture has higher accuracy (Gamanayake et al., 2020). Luo et al. (2020) proposed to combine SSD neural network architecture with RGB-D images to resolve shape differences between different objects, generating correlation scores by adjusting position, size, and orientation to facilitate final detection of non-maximal inhibition. Experimental results show that the proposed model achieves high accuracy and fast detection speed in NYUv2 datasets. Du et al. (2020) applied SSD neural network structure to SAR images to suppress clutter and reflect image saliency, which effectively improved the representation ability in complex scenes. In this model, multi-level fusion modules are combined into a unified framework to achieve joint training of the models. The proposed S-SSD model has superior detection performance in miniSAR data.

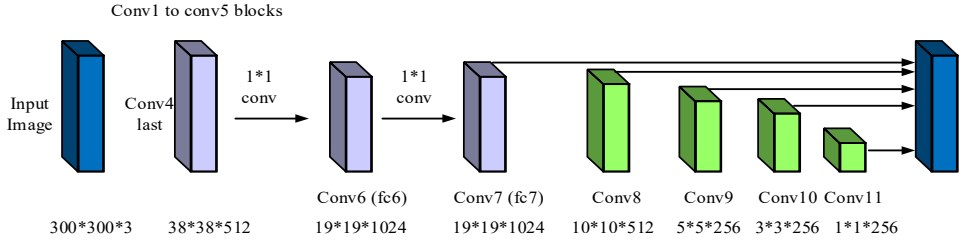
Current machine vision technology has a variety of object recognition models, and image detection capabilities are used in many fields. However, there are fewer applications in artwork recognition. The study applies SSD models to the authentication of artworks with the aim of ensuring the normal creation of artworks and protecting the interests of consumers.

3 Object recognition model construction based on deep learning binocular stereo vision

3.1 Object recognition model construction based on single neural network image detection

The study uses SSD to build an object recognition model, which is a standard convolutional neural network in the field of image classification. By subsequently adding convolutional layers, decreasing the feature scale of the convolutional layers, predicting the object classification and the offset of the default frame, thus achieving multi-scale feature map detection, and finally using the role of non-maximum suppression (NMS) to eliminate the redundant frame to complete the object detection and classification (Matsunobu et al., 2021). The SSD network model guarantees both real-time and accuracy. The network structure is shown in Figure 1.

Figure 1 SSD network structure diagram (see online version for colours)



The region proposal network (RPN) algorithm in the SSD network is mainly used for network generation detection to generate the bias of the default box for objects with different scale features and the confidence of the training objects, where the size of the default box is fixed. The calculation of the different scales of the default box is performed by the maximum and minimum default box scales, and the specific expressions are shown in equation (1).

$$S_k = S_{\min} + \frac{S_{\max} - S_{\min}}{m-1}(k-1) \tag{1}$$

In equation (1), S is used to indicate the scale of the default box size, the number of default boxes with the symbol k, m indicates the number of feature maps. The feature maps are different. The maximum value minus the minimum value gets the scale except the minimum value area. The scale of a single feature map is obtained by evenly dividing the number of feature maps. The scale of a single feature map multiplies the corresponding number of features to get the scale of multiple feature maps. The minimum and maximum scale values of the default box are 0.2 and 0.95 respectively, and the determination of the extreme value is determined by the performance of the computer itself. Therefore, the ratio of the default box can be expressed by the width and height. The formula for calculating the width and height is shown in equation (2).

$$\begin{cases} w_k^a = S_k\sqrt{a_r} \\ h_k^a = S_k / \sqrt{a_r} \end{cases} \tag{2}$$

In equation (2), a_r is the common sense value, its value range is $a_r \in \{1, 2, 1/2, 3, 1/3\}$, w_k^a and h_k^a respectively, the width and height of the default box. When the ratio value of the default box is obtained, the position of the centre coordinates of the default box can be determined, and the centre coordinates are calculated as shown in equation (3).

$$\begin{cases} x = \dfrac{i + 0.5}{|f_k|} \\ y = \dfrac{j + 0.5}{|f_k|} \end{cases} \tag{3}$$

In equation (3), x, y denotes the centre coordinates of the default frame, $|f_k|$ denotes the size of the k feature map, and i, j are constants whose values are between 0 and $|f_k|$. The NMS algorithm of SSD convolutional neural network can remove the default frames with lower scores to reduce the overlapping frames. The default frames in the first row of each column are arranged in order, and the overlap rate is calculated separately for the subsequent default frames whose value size is compared with the threshold size, and finally the NMS is performed for all categories, and the default frames whose confidence does not satisfy the threshold are deleted (Castro-Zunti et al., 2020).

The focus of the SSD convolutional neural network model is to map the real frame to the real frame category to the default frame (De Gregorio et al., 2020). The model calculates the confidence of the default box with the target confidence using the softmax loss regression function and its position loss function using the softmax L1 loss regression function. The SSD convolutional neural network loss function is shown in equation (4).

$$L(x, c, l, g) = \frac{1}{N}\left(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)\right) \tag{4}$$

In equation (4), N indicates the default number of matched boxes, and if the number is changed to 0, the loss function value is 0. The confidence loss function and the position loss function are represented by L_{conf} and L_{loc}, respectively, and l indicates the predicted box and g indicates the real box. If the loss function of confidence belongs to multi-category confidence, the loss function is represented as shown in (5), where $\hat{c}_i^p$ denotes the category of confidence loss function.

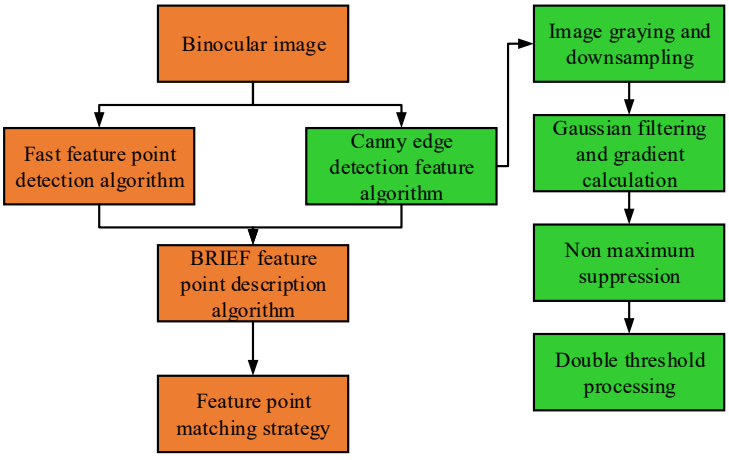
$$\begin{cases} L_{conf}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^p \log\left(\hat{c}_i^p\right) - \sum_{i \in Neg} \log\left(\hat{c}_i^0\right) \\ \hat{c}_i^p = \dfrac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)} \end{cases} \tag{5}$$

3.2 Object recognition and localisation model construction based on visual image feature point matching

The combination of Canny edge detection algorithm and speeded up robust features (SURF) algorithm yields C-SURF algorithm, which is an edge detection feature point matching algorithm. Based on this algorithm, the features from accelerated segment test

(FAST) matching algorithm is improved to obtain a matching algorithm for multi-feature point fusion of binocular images, which does not have the fixity of the expected rotation of the scale and therefore has excellent results in detection speed (Shao et al., 2020). The specific flowchart of this algorithm is shown in Figure 2.

Figure 2 MPFM algorithm flow (see online version for colours)



The Canny edge point detection algorithm uses a three-phase criterion as the evaluation criterion for the superiority of edge detection by Gaussian smoothing followed by derivation. One of them is the low error rate detection criterion, which minimises the error detection errors and achieves the most jealous edge detection effect, whose signal-to-noise ratio needs to reach the maximum value. The mathematical addition formula of the signal-to-noise ratio is shown specifically in equation (6).

$$SNR = \frac{\left|\int_{-w}^{+w} G(x) f(x) dx\right|}{\sigma\sqrt{\int_{-w}^{+w} f^2(x) dx}} \tag{6}$$

In equation (6), $f(x)$ denotes a Gaussian filter with a boundary range of $[-w, +w]$, then the image formed after Gaussian filtering is represented by $G(x)$, and the mean squared difference of the Gaussian filter template is represented by σ. The second is the optimal localisation criterion, which is based on the principle that the smaller the difference between the detected edge points and the edge points of the real situation, the more accurate it is. Its specific calculation formula is shown in equation (7).

$$Location = \frac{\left|\int_{-w}^{+w} G'(x) f'(x) dx\right|}{\sigma\sqrt{\int_{-w}^{+w} f'^2(x) dx}} \tag{7}$$

The third is the single-edge response criterion, which satisfies the second-order derivative of $f(x)$, whose mathematical specific expression is shown in equation (8).

$$d(f') = \pi \frac{\left| \int_{-\infty}^{+\infty} f'^2(x)dx \right|}{\left| \int_{-\infty}^{+\infty} f''(x)dx \right|} \tag{8}$$

According to Canny's principle of edge detection, the 3 * 3 Gaussian filter template is used as an example. Because the number of feature points detected in the original image is large and the information is redundant, which will affect the matching speed of the algorithm, the downsampling approach is adopted. The mathematical expression of Gaussian filtering is shown in equation (9).

$$\begin{cases} G(x, y) = \dfrac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \\ f_s(x, y) = f(x, y) * G(x, y) \end{cases} \tag{9}$$

In equation (9), the $f(x, y)$ is represented the input data, then $f_s(x, y)$ for the data processed by the Gaussian filter, and the two-dimensional Gaussian function is represented by $G(x, y)$. Let the squared variance value be 0.64, and the coordinate value and variance are substituted into the template to obtain the standard Gaussian filter template by normalisation, as shown in Figure 3.

Figure 3 Weight matrix of Gaussian filter (see online version for colours)

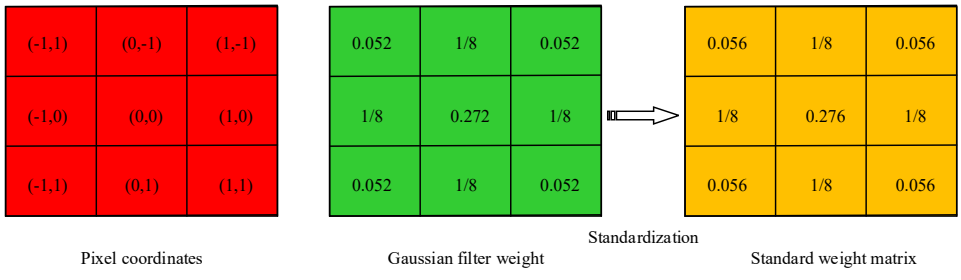


Figure 3 represents the weight matrix, where the gradient and direction of the pixel points in the image need to be calculated. Where the gradient calculation can be transformed to find the greyscale variation of the pixel, the partial derivative of the horizontal and vertical coordinate directions, and finally the convolution template is used as the partial differential approximation of the horizontal and vertical coordinates. The gradient direction, azimuth angle, and vertical gradient edge direction of the pixel points are obtained. The angles are divided into four directions, and non-maximal suppression is applied to each direction, and the magnitude is compared and the maximum value is retained. A double-threshold algorithm is used to detect and edge connect the gradient image by selecting two thresholds, and the pixel point is taken as an edge feature point if the gradient value of the image pixel is not less than the threshold value, and vice versa. If the pixel gradient value is between the two thresholds, it is judged whether there is a point larger than the threshold value in the eight pixel points around the pixel point, and if it exists, the pixel point can be used as an edge feature point. Finally, the extracted feature points are mapped to the original image. After the Canny edge detection algorithm, the positions of the feature points are extracted, and now the feature

descriptors are constructed for the detected feature points by improving the SURF algorithm. The SURF algorithm uses the integral image to improve the efficiency of the feature point descriptor calculation, and the integral image is calculated as in equation (10).

$$I(x, y) = \sum_{i<x, j<y} I(i, j) \tag{10}$$

In equation (10), the $I(x, y)$ represents the integral image. Since the scale does not change in binocular images, the scale of the feature points S is determined as 1. The detected feature points are centred on a square space, which can be divided into 4 * 4 subspaces, and each subspace has 5 * 5 pixel regions, and the corresponding vectors of Haar wavelets in all subspaces are calculated in the horizontal and vertical directions, and their values are recorded as *dx* and *dy*, respectively. The response vectors of the pixel regions are Gaussian weighted to make them robust to noise interference, and the normalisation of the descriptors removes the situation caused by the brightness variation between the left and right of the image. Finally, the feature vectors of all subspaces are counted.

An optimal feature detection method is obtained in the FAST algorithm, using a feature point detection method based on machine learning. Corner point detection is used for multiple images to be detected, and consecutive pixel points as well as threshold values are used as set values, denoted by n, t respectively. The extracted feature points are used as training samples to divide the feature points, and the specific division rules are shown in equation (11).

$$S_{p\to x}\begin{cases} d, & I_{p\to x} \le I_p - t \\ s, & I_p - t \le I_{p\to x} \le I_p + t \\ b, & I_{p\to x} \ge I_p + t \end{cases} \tag{11}$$

The study uses the ID3 algorithm to build a decision tree for the feature points after conducting training, and if the decision division is performed is determined by the first X pixel point, the entropy value obtained from the set P is K, which is calculated as shown in equation (12).

$$H(P) = (c + \overline{c}) \log_2 (c + \overline{c}) - \left(c \log_2 (c) + \overline{c} \log_2 (\overline{c})\right) \tag{12}$$

In equation (12), the number of corner points is expressed by c and the number of non-corner points is expressed by $\overline{c}$, through which the information gain can be obtained, and the decision tree is formed to segment the maximum position of the information gain, and the decision tree can be used for detection when meeting similar scenes for feature point detection. The NMS method is used to solve the problem of easy aggregation of feature points. Each detected corner point is assigned a quantisation value V, and the quantisation values of neighbouring corner points are compared to eliminate the smaller quantisation values. The set of pixel points in a particular domain with greyscale values greater than $I_p + t$ is denoted by S_b and the set of pixel points with greyscale values less than $I_p - t$ is denoted by S_d. The quantisation value V is defined as shown in equation (13).

$$V = \max\left(\sum_{x\in S_b} \left|I_{p\to x} - I_p\right| - t, \sum_{x\in S_d} \left|I_p - I_{p\to x}\right| - t\right) \tag{13}$$

In the multi-feature point fusion detection algorithm, feature point detection is greatly enriched by the ability to detect not only feature points of edges but also corner point features (Wang et al., 2021). BRIEF descriptors are binary feature descriptors, so their feature point descriptors are binary codes that appear when neighbouring pixel points are compared in greyscale values. Compared to floating point data, the binary robust independent elementary features (BRIEF) descriptor algorithm has significantly less computation of feature vectors and the rate of matching is improved. The noise immunity of the descriptor is improved by filtering the image with a Gaussian template of size 9 * 9 and a variance σ of 2. The specific way of its filtering is shown in equation (14).

$$G(x, y) = \frac{(\pi\sigma^2)^{-1}}{2} \cdot e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{14}$$

After obtaining the feature point p, the comparison points are selected in the block of pixels W * W around p, and the intensity values of the selected pixel points are compared by the τ method, and the comparison rules of the τ method are shown in equation (15).

$$\tau(p, u, v) = \begin{cases} 1, & p(v) \geq p(u) \\ 0 & \end{cases} \tag{15}$$

$p(\cdot)$ in equation (15) indicates the greyscale value at the selected point.

4 Analysis of the results of object recognition based on deep learning binocular stereo vision

4.1 Performance analysis of object recognition model with SSD convolutional neural network

In this study, objects commonly seen in daily life were selected as training samples through PASCAL VOC2011-2 dataset. The PASCAL VOC2011-2 dataset contains 11,530 picture data. The dataset is roughly divided into 20 categories, including people, animals, vehicles and indoor objects, which are commonly used for detection tasks. Since the SSD network is a framework formed on the caffe deep learning approach, the network weights in the caffe deep learning framework can be used as the initial values for the training of its own network structure by means of migration learning. This migration learning training method can improve the convergence rate of the network structure, reduce the maximum number of iterations, and thus improve the training efficiency of the model. The learning rate of the model training is 0.001, the number of iterations of the model and the number of samples of the reverse recursion are 16,000 and 100 respectively, and the value of the loss function is output every ten times of training, and the trend of the loss function is shown in Figure 4.

The training error curve in Figure 4 results in a gradual decrease in the loss function value of this model as the number of model iterations increases, and at 6,000 iterations, the loss function curve starts to converge and the curve trend stabilises, with the loss function value fluctuating around 0.9. Therefore, it is proved that the performance of the model is improved with the training of migration learning. To verify the superiority of the SSD model performance, the YOLO network model is added as a comparison. The YOLO network model cannot be trained by the migration learning method, so the

training iterations of the model are set to 4 * 104 times, the learning rate and the reverse recursive sample parameters are kept constant, and the output time of the loss function is increased to 20 times with ten intervals. The trend of the loss function change in the YOLO model is shown in Figure 5.

Figure 4 Curve of loss function value in SSD network model (see online version for colours)

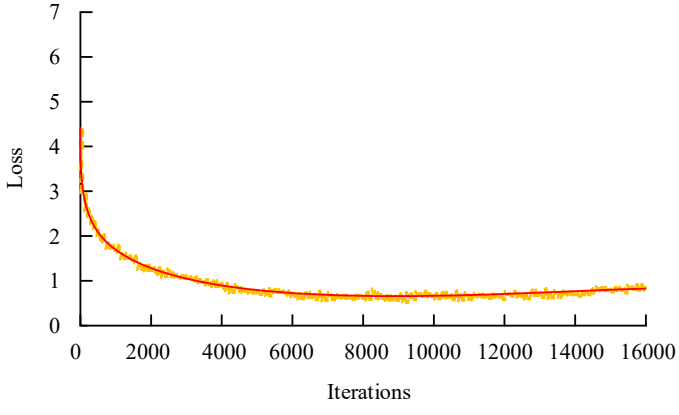
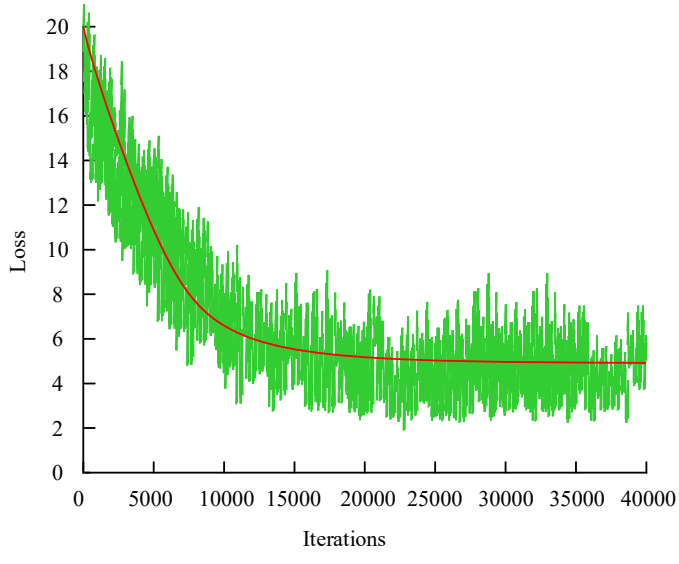


Figure 5 Loss function value curve in YOLO network model (see online version for colours)



The trend of the loss function of the YOLO model in Figure 5 is the same as that of the SSD model, but the final degree of convergence and the range of change after convergence are different between them in different ways of definition. The value of the loss function of the YOLO model is 157.53 at the beginning of the iteration, which decreases to 13.91 after 20 iterations, and the convergence of the loss function occurs at 25,000 iterations with Its value fluctuates around 5.00. Compared with the SSD model, the YOLO model is slightly inferior to the SSD model in terms of overall performance.

Figure 6 shows the average overlap between the labelled frames of the training samples of the SSD model and the detection frames of the network structure.

Figure 6 Changes in the average overlap rate between the label box of SSD model training samples and the detection box of network structure (see online version for colours)

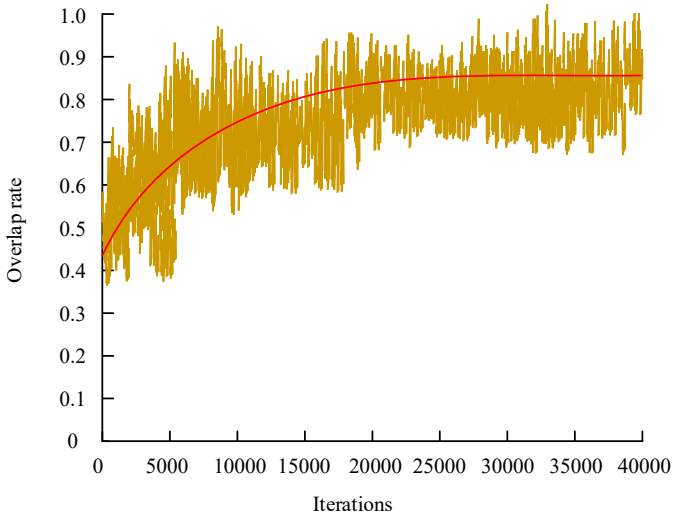
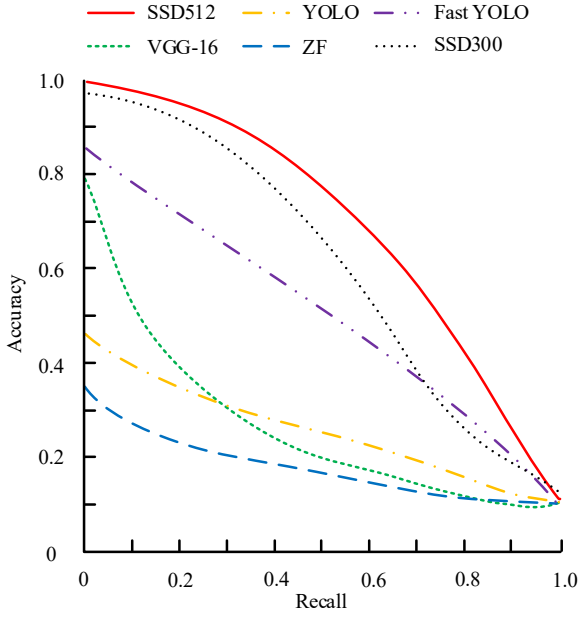


Figure 7 Performance analysis of six object recognition algorithms in PASCAL VOC2012 test set (see online version for colours)



The change of the overlap rate in Figure 6 can reflect that the value of the overlap rate increases with the number of iterations. The higher the value of the overlap rate, the better the training effect of the model. When the number of iterations reaches the

maximum, the overlap rate of the model reaches about 0.85. The accuracy and recall of the model are calculated by the overlap rate. The recall rate is taken as the abscissa and the accuracy rate as the ordinate. The area enclosed by the graph is a P-R curve, which can be used as the evaluation index of object recognition. Figure 7 P-R curve of each target detection model.

In Figure 7, the test set of PASCAL VOC2012 is used to analyse the performance of six object recognition algorithms. From the curve trend, it can be seen that SSD512 object recognition algorithm is superior to other algorithms when it is close to real-time detection.

Table 1 Accuracy results of items trained in the model

Recognition algorithm	*Fast R-CNN*	*Faster R-CNN*	*YOLO*	*SSD300*	*SSD512*
Bicycle	79.50	80.90	68.30	81.30	85.80
Bottle	39.40	53.20	23.30	48.70	54.30
Automobile	77.00	83.10	57.00	77.20	82.60
Bus	82.70	84.20	69.40	84.10	87.30
Cat	87.80	87.50	82.50	87.20	87.10
Chair	43.90	53.10	37.30	55.80	58.90
Table	56.10	60.10	49.60	64.60	65.50
Dog	85.80	85.90	78.30	85.60	86.00

Table 1 shows the results of the accuracy analysis of the eight object categories trained, and according to the data in the table, it is found that the accuracy of different object recognition in the same model varies, which is related to the number of samples trained, the quality of the images and the nature of the features of the object categories. The accuracy rate of SSD512 is higher than that of other algorithms in all kinds of object recognition. The reason is that SSD512 model structure adopts multi-scale feature map convolution structure, which makes the feature extraction and recognition of items easier and more accurate, thus obtaining the highest recognition degree. In cat category, each algorithm has no significant difference in its object features, so its recognition accuracy is similar without significant difference.

4.2 Performance test results analysis of binocular image feature point matching algorithm

The study uses binocular camera approach to acquire binocular images, and the performance of four algorithms, SIFT, SURF, FAST, and ORB, is analysed in comparison with C-SURF and MPFM, respectively. Table 2 shows the experimental data results of binocular image feature point matching.

From the analysis of the experimental data in Table 2, the SIFT and SURF algorithms have a large gap with the proposed C-SURF in terms of shortening the running time and increasing the number of matching points, and the number of matching points is less than half of the proposed algorithm, and the SIFT and SURF algorithms detect fewer target feature points, which will lead to instability in the calculation of parallax and affect the accuracy of subsequent object localisation. Although the operation time of the C-SURF algorithm is slightly longer than that of the FAST and oriented fast and rotated brief

(ORB) algorithms, the target matching points are much higher than these two algorithms, and it is difficult to support the subsequent binocular localisation with fewer target matching points. Therefore, the performance of C-SURF algorithm is still stable in target recognition and localisation with less texture. The performance comparison between C-SURF algorithm and MPFM algorithm is shown in Table 3.

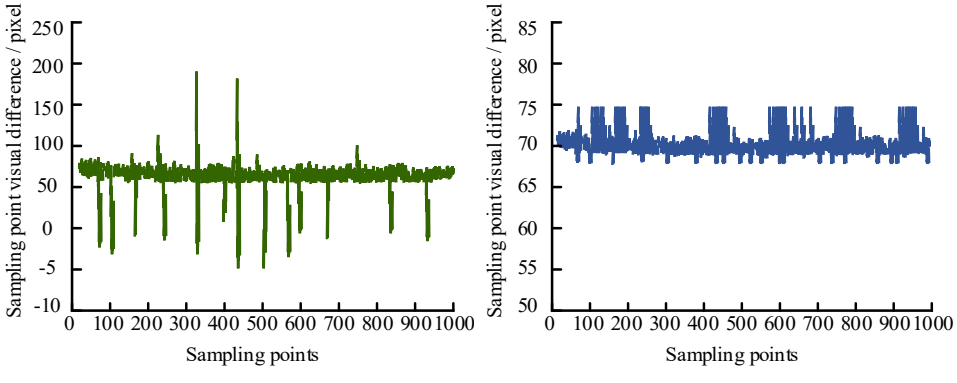
Table 2 Experimental data results of binocular image feature point matching

Matching method	*Number of feature points (left/right)*	*Number of matches*	*Matching time (seconds)*	*Target matching point*
SIFT	969/867	93	1.134	7
SURF	1,234/1,107	89	0.967	12
FAST	1,867/1,943	13	0.254	6
ORB	663/643	20	0.126	5
C-SURF	795/851	216	0.353	46

Table 3 Performance comparison results of c-surf algorithm and MPFM algorithm

Matching method	*Number of feature points (left/right)*	*Number of matches*	*Matching time (seconds)*
C-SURF	1,172/997	182	0.250
MPFM	2,803/2,547	487	0.330

Figure 8 Parallax fluctuation before and after filtering, (a) fluctuation of visual difference before filtering (b) parallax fluctuation after filtering (see online version for colours)



Compared with MPFM, C-SURF has fewer feature detection points, which are mostly distributed at the target edges, and fewer feature points are detected in local areas. MPFM solves the problem of fewer local feature detection points, but too many feature detection points increase the running time of the algorithm. With the current computer performance, the effect of real-time processing can be fully achieved, so the research of the improved algorithm can solve the problems existing in the current binocular image feature point matching. The C-SURF algorithm is used to detect the fixed target and collect parallax data. The parallax value within the target range can be calculated by

matching points of the left and right images of the target, and the fluctuation of parallax value before and after filtering is shown in Figure 8.

Figure 8(a) shows that the parallax value at this time has a large fluctuation range, which may adversely affect the subsequent positioning accuracy. Figure 8(b) shows the parallax value after median filtering, which eliminates the parallax value of anomalies and finally stabilises in the range of (67, 75), with high positioning accuracy. The C-SURF algorithm is used to analyse the object localisation error for different distances, and the results in Figure 9 are obtained.

Figure 9 The results of C-SURF algorithm (see online version for colours)

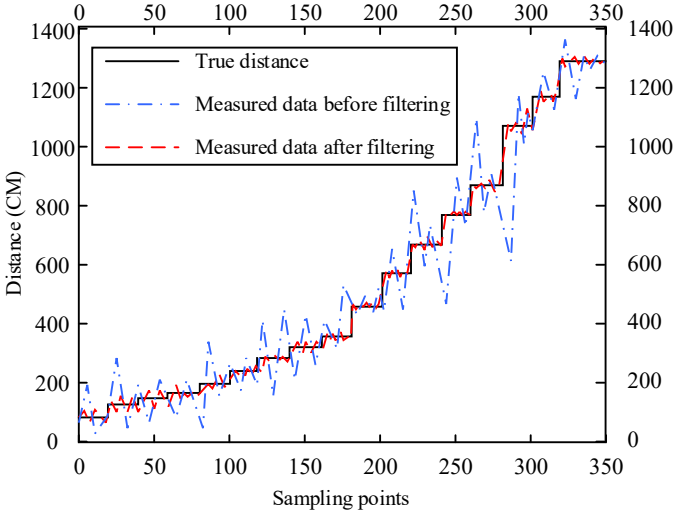


Table 4 Statistical table of positioning error under different distance conditions

Measured distance (cm)	*SIFT*	*SURF*	*C-SURF*	*MPFM*
91.1	97.2	98.4	82.6	91.8
191.1	197.3	196.7	193.0	192.4
291.1	298.6	295.4	292.7	293.9
391.1	398.8	396.4	393.8	395.4
491.1	499.7	499.2	493.6	494.3
591.1	600.4	602.4	594.8	597.8
691.1	701.2	704.9	694.5	700.3
791.1	806.9	807.4	796.2	800.6
891.1	908.7	910.6	900.4	901.4
991.1	1,009.5	1,011.4	1,001.2	1,002.3
1,091.1	1,116.9	1,120.7	1,111.0	1,121.4
1,191.1	1,227.3	1,224.7	1,212.9	1,207.0

The results of the real range and the actual range in Figure 9 reflect that the range error after filtering is significantly less than the range error before filtering, which is closer to the real range. The statistics of the localisation errors were conducted in different distance

cases, which included four algorithms, SURF, SIFT, C-SURF and MPFM, and each group of tests contained 100 test points, and the statistics are shown in Table 4.

From the error statistics in Table 4, we know that the longer the detection distance, the larger the error value under the same algorithm. At the same distance, the errors of SURF algorithm and SIFT algorithm are higher than those of C-SURF algorithm and MPFM algorithm. The error statistics between C-SURF and MPFM are similar, which shows that the feature point matching algorithm proposed in the study has excellent object recognition and location functions. Based on the analysis of the above results, the multi-feature extraction model proposed in the study is more reliable than the single feature extraction model of other algorithms. Recognition in single features has certain limitations. In true and false objects, there are an inevitably similar feature, which leads to recognition and classification errors. Multi feature extraction can verify the authenticity of an item positively or negatively through different features. Therefore, the method proposed by the study has more advantages.

5 Conclusions

The development of computer vision technology has enabled binocular vision with the function of acquiring depth information of the target, making the technology to be applied in many fields. In art design, in order to increase the strength of artwork authenticity identification, the study uses machine deep learning binocular stereo vision technology to identify and discriminate works. The study establishes the SSD convolutional neural object recognition model to find the feature points of the works by matching the feature points of binocular images. The value of loss function in the SSD convolutional neural network model is stable around 0.9, which has obvious advantages over the YOLO model, and the overlap rate of the proposed model fluctuates around 0.85, which indicates that the algorithm proposed in the study has better performance. In the mAP index detection, again the algorithm has the best accuracy performance. The binocular image feature point matching algorithm established in the study showed the best performance when compared with the rest of the different algorithms, having more detected feature points while shortening the running time of the algorithm. The performance of the parallax value has strong stability after median filtering, and the parallax value fluctuates in a small range of (67, 75), indicating that the algorithm has high localisation accuracy. For the performance analysis of the detection error at different distances, the error of the algorithm proposed in the study is smaller than the error of the rest of the algorithm. For the method proposed in the study, there are still shortcomings, the method needs to rely on better hardware equipment to achieve real-time processing capability, and the overlap rate of the frame and the matching accuracy still need to be further improved, so the subsequent study continues to optimise the proposed algorithm to get higher matching accuracy.

Acknowledgements

The research is supported by: Special Program and Key Subject of Art and Science for Youth in Shandong Province in 2020 (QN202008196); Doctoral Foundation of Qingdao Agricultural University (1120708).

References

Akundi, A. and Reyna, M. (2021) 'A machine vision based automated quality control system for product dimensional analysis', *Procedia Computer Science*, Vol. 185, pp.127–134.

Castro-Zunti, R.D., Yépez, J. and Ko, S.B. (2020) 'License plate segmentation and recognition system using deep learning and OpenVINO', *IET Intelligent Transport Systems*, Vol. 14, No. 2, pp.119–126.

De Gregorio, D., Tonioni, A., Palli, G. and Di Stefano, L. (2020) 'Semiautomatic labeling for deep learning in robotics', *IEEE Transactions on Automation Science and Engineering*, Vol. 17, No. 2, pp.611–620.

Du, L., Li, L., Wei, D. and Mao, J. (2020) 'Saliency-guided single shot multibox detector for target detection in SAR images', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 58, No. 5, pp.3366–3376.

Gamanayake, C., Jayasinghe, L., Ng, B.K.K. and Yuen, C. (2020) 'Cluster pruning: an efficient filter pruning method for edge AI vision applications', *IEEE Journal of Selected Topics in Signal Processing*, Vol. 14, No. 4, pp.802–816.

Gorai, A.K., Raval, S., Patel, A.K., Chatteriee, S. and Gautam, T. (2021) 'Design and development of a machine vision system using artificial neural network-based algorithm for automated coal characterization', *International Journal of Coal Science & Technology*, Vol. 8, No. 4, pp.737–755.

Jang, Y., Gunes, H. and Patras, I. (2019) 'Registration-free face-SSD: single shot analysis of smiles, facial attributes, and affect in the wild', *Computer Vision & Image Understanding*, Vol. 182, No. 4, pp.17–29.

Lin, J., Du, Z., Yu, C., Ge, W., Lu, W. Deng, H. and Xu, J. (2021) 'Machine-vision-based acquisition, pointing, and tracking system for underwater wireless optical communications', *Chinese Optical Letters*, Vol. 19, No. 5, pp.25–30.

Luo, Q., Ma, H., Tang, L., Wang, Y. and Xiong, R. (2020) '3D-SSD: learning hierarchical features from RGB-D images for amodal 3D object detection – ScienceDirect', *Neurocomputing*, Vol. 378, No. 5, pp.364–374.

Matsunobu, L.M., Pedro, H. and Coimbra, C. (2021) 'Cloud detection using convolutional neural networks on remote sensing images', *Solar Energy*, Vol. 230, No. 1, pp.1020–1032.

Shao, W., Cao, L., Guo, W., Xie, J. and Gu, T. (2020) 'Visual navigation algorithm based on line geomorphic feature matching for Mars landing', *Acta Astronautica*, Vol. 173, No. 12, pp.383–391.

Souto, M.L., Fernández, A. and Guerra, L. (2020) 'Real machine vision use-cases applied to hot rolling mill plants', *Procedia Manufacturing*, Vol. 51, pp.280–287.

Talaat, M., Arafa, I. and Metwally, H. (2020) 'Advanced automation system for charging electric vehicles based on machine vision and finite element method', *IET Electric Power Applications*, Vol. 14, No. 13, pp.2616–2623.

Viguerie, L.D., Pladevall, N.O., Lotz, H., Freni, V., Fauquet, N., Mestre, M. and Verdaguer, M. (2020) 'Mapping pigments and binders in 15th century gothic works of art using a combination of visible and near infrared hyperspectral imaging', *Microchemical Journal*, Vol. 155, No. 12, p.104674.

Wang, P., Fu, H., Li, X., Guo, J., Lv, Z. and Di, R. (2021) 'Multi-feature fusion tracking algorithm based on generative compression network', *Future Generation Computer Systems*, Vol. 124, No. 32, pp.206–214.

Wang, P., Sun, X., Diao, W. and Fu, K. (2020a) 'FMSSD: feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 58, No. 5, pp.3377–3390.

Wang, Y., Jia, X., Li, X., Yang, S., Zhao, H. and Lee, J. (2020b) 'A machine vision based monitoring system for the LCD panel cutting wheel degradation – ScienceDirect', *Procedia Manufacturing*, Vol. 48, pp.49–53.

Zhang, X., Liu, J. and Song H. (2020) 'Corn ear test using SIFT-based panoramic photography and machine vision technology – ScienceDirect', *Artificial Intelligence in Agriculture*, Vol. 4, No. 2, pp.162–171.

Zhao, L., Cheng, J., Yin, Z., Yang, H., Chen, M. and Yuan, X. (2021) 'Research on precision automatic tool setting technology for KDP crystal surface damage mitigation based on machine vision', *Journal of Manufacturing Processes*, Vol. 64, No. 6, pp.750–757.