Research on RGB-D image recognition technology based on feature fusion and machine learning

Qingbo Sun
Shandong University of Political Science and Law, 250014
20385489@dlvct.edu.cn

Abstract: The three-dimensional RGB-D image contains not only the color and texture information of the two-dimensional image, but also contains the surface geometry information of the target. This article analyzes the RGB-D image recognition methods, including stereo vision technology, structured light technology, etc. By studying the application points of RGB-D image recognition technology under the background of feature fusion and machine learning, the purpose is to improve the richness of image recognition content and provide reference for the smooth development of the follow-up work.

1. Introduction
The currently applied RGB-D image not only contains the color and texture information of the two-dimensional image, but also contains many types of geometric information. Especially with the integration of Kinect, Xtion and other equipment, the level of refinement and clarity of RGB-D image content has also been continuously improved. The optimization of the RGB-D image recognition technology system based on feature fusion and machine learning can not only improve the accuracy of the image recognition results, but also has a positive meaning for improving the value of image utilization.

2. RGB-D image recognition method analysis

2.1 Stereo Vision Technology

Figure 1 Schematic diagram of binocular vision
The stereo vision method is a traditional way of acquiring depth images. Its principle is to calculate the parallax through the photos of the same scene from multiple perspectives, and obtain the three-dimensional information of the objects in the scene according to the parallax. A general stereo vision system is composed of several ordinary industrial CCD cameras, which require manual camera calibration before acquiring the depth map, which is generally calibrated by the positive checkerboard method. As shown in Figure 1, O_L and O_R are the centers of the two cameras respectively, and the three-dimensional coordinates of point A can be obtained from the similar relationship of triangles. The stereo vision method does not require high-precision sensors, and only a few ordinary cameras are needed to obtain the depth map, but its acquisition of the depth map requires a series of complex algorithms to operate on the original image.

2.2 Structured light technology
Structured light technology is currently a more advanced depth map acquisition method. Because structured light measurement has obvious advantages such as fast speed, high precision, non-contact, easy implementation and automation, it has been widely used in industrial inspection, machine vision, film and television special effects, etc. field. Projectors and cameras are the basic components of structured light systems. Once the projector projects the coded pattern onto the scene to be measured, the camera can capture and shoot the projected pattern, and the acquisition of the three-dimensional information of the scene is the result of matching the code with the projected pattern. Therefore, coding and matching are two important factors for constructing a structured light system. Take Kinect as an example, its structure is as follows.

Figure 2 Schematic diagram of Kinect structure

The principle of Kinect's acquisition of color images is the same as that of ordinary cameras. This article focuses on the principle of Kinect sensor's acquisition of depth images. Kinect essentially uses structured light technology to obtain the depth distance. It emits a light source through an infrared emitter. This light source is called laser speckle, which is a diffraction spot formed randomly by laser irradiating the rough surface of an object or penetrating ground glass. These spots will change with the distance of the object from the Kinect sensor. Microsoft recorded the patterns of these laser speckles in the entire space, and made a one-to-one correspondence with the distance.
3. Application points of RGB-D image recognition technology under the background of feature fusion

3.1 RGB-D image extraction

3.1.1 Gist feature extraction
From the perspective of previous RGB-D image extraction, there are problems such as poor content extraction, high dimensionality of feature data, and low accuracy of algorithm calculation results. In view of this kind of situation, in the process of Gist feature extraction, the extraction algorithm needs to be reasonably selected. Currently, more algorithms are used including HOG algorithm, SIFT algorithm, SURF algorithm, etc. At the same time, in the process of algorithm application, it is also necessary from the perspective of gradient analysis. To start the analysis to improve the accuracy of the calculation results. Moreover, in the calculation process, the overall calculation amount is relatively large, which is 150% higher than the total amount of calculation data in the past, which also requires more reliable algorithms to complete information sorting. In the process of Gist feature extraction, feature data extraction will be completed from naturalness, openness, roughness, expansion and steepness, and then Gabor filters are used to process the content to obtain richer data information. An example of graph extraction is shown in Figure 3.

![Figure 3 Example diagram of Gist feature extraction](image)

3.1.2 Color-depth map extraction
In the process of data information extraction, it is also necessary to do a good job in the extraction of color-depth maps. In the specific application process, based on the extracted information obtained by the existing Gist algorithm, sub-networks will be used to further improve the image content. Usually, the image is densified to refine the content of the required extraction part, and then the feature points are extracted using the Gist algorithm, and the weighting coefficient is calculated at the same time, so as to obtain the local Gist feature that meets the application requirements. Take the schematic diagram in Figure 3 as an example. In the specific processing process, first, cover the entire picture with a 16×16 grid, and then extract the Gist features in each grid; second, for the existing The divided grid is divided again and organized into 4×4 grids, that is, each small grid is divided into 16 small grids. Calculate the weights according to the correlation between each other to obtain the required analysis results.
3.2 Feature dimensionality reduction processing

After finishing the above part of the content processing, all the parameter information can be learned based on the Gist feature. At the same time, the basic information in the RGB image will be extracted, including color information, texture information, and geometric information, as shown in Figure 4, after segmentation. The picture is converted into $16 \times 16 \times (4 \times 4)$ small grid graphics, and the feature dimension that needs to be extracted has been greatly improved. In order to speed up the calculation, the principal component analysis method will also be used to assist the calculation process. This method belongs to the data dimensionality reduction processing method often used in the process of graph feature extraction. Its principle is shown in Figure 5. The projection point is selected in the coordinate system, and the known high-latitude data is mapped to the low-latitude data by means of the vector projection method. In the latitude space, the projection components are then used for orthogonal processing to obtain the required linear vector to meet specific application requirements. According to the projection results, the amount of projection in different dimensions can also be obtained, and the main direction of the analyzed data information and the criticality of the characteristic content can be understood, so as to complete the dimensionality reduction process. For example, $16 \times 16 \times (4 \times 4)$ sub-grid patterns were obtained by the previous division. After dimensionality reduction processing, the total dimensional value can reach 131072, which has rich data information characteristics.
3.3 Characteristic content expression
When performing feature content expression processing, the application process includes the following points: First, perform feature content extraction. According to the bag-of-words model in the initial state, the graphics will be subdivided into several sub-modules in the application, and then the content of these pixels will be subdivided into several sub-modules. Organize so that it can form a representative vector according to a certain rule. In RGB-D images, the Gist algorithm will also be used to optimize the application content in the initial state to obtain more analytically valuable application data. Second, use the K-means clustering algorithm to complete the sample point division processing. In specific applications, it is also necessary to sort out the correlation between the initialization position and the clustering effect and running time, and select 30%-50% of the applications Data, thereby enhancing the use value of the analysis results.

3.4 Simulation experiment analysis

3.4.1 Preparation for simulation experiment
In the construction of this simulation experiment, the RGB-D database is used as a basic reference, and all the pictures in the database are used as simulation experiment data. In terms of image acquisition, the Kinect sensor is used to capture them. The total number of captured RGB images is 5863 sheets, the resolution of each picture is adjusted to 1280×640, in order to facilitate subsequent data analysis work in an orderly manner.

3.4.2 Simulation experiment process
According to the basic content provided by the simulation experiment environment, in the whole experiment process, first, adjust the size of the pictures in the data set, the specific adjustment size is 512×512, and then according to the content in 2.1-2.3, for each picture The information undergoes Gist feature processing to obtain corresponding reference data. Second, use the K-means clustering algorithm to perform in-depth processing, and in the application of the algorithm, set the K value in the algorithm to 500. Third, determine the number of training data in the experiment. In this simulation experiment, 4876 pictures are selected for training, and the data set used for testing is (5863-4876)=987 pictures. The ten-fold cross-validation method is used for content verification, thereby obtaining image recognition accuracy data.

3.4.3 Arrangement of experimental results

Table 1 Analysis of recognition accuracy of different algorithms

| Use algorithm                          | Recognition accuracy |
|----------------------------------------|----------------------|
| HOG+SVM algorithm                      | 79.3%                |
| SIFT+SPM algorithm                     | 86.3%                |
| KD+SVM algorithm                       | 90.3%                |
| Gist image recognition algorithm       | 93.2%                |

From the experimental results in Table 1, it can be understood that the Gist image recognition algorithm used in this article has achieved 93.2% accuracy in image recognition, while the traditional HOG+SVM algorithm has the lowest accuracy, only 79.3%, and the accuracy of the other two types of algorithms The image recognition accuracy of the KD+SVM algorithm reaches 90.3%, and it also has a high image recognition accuracy. However, the algorithm proposed in this article still has a lot of room for improvement. Analyzing the causes of image recognition errors, it is understood that the background color in the confused image is more than 70% similar, which leads to algorithm recognition errors. Therefore, in the subsequent development, it is necessary Focus on the background color distinction to meet the corresponding management needs [1].
4. Application points of RGB-D image recognition technology in the context of machine learning

4.1 Neural network design
When the neural network is designed based on the actual situation, the specific design steps are as follows: First, preprocess the collected images, filter out the required picture information from the RGB-D image, and perform simple preprocessing on its related content. Second, rely on the neural network to establish the corresponding neural network structure, in which the elements in the RGB-D image will be identified, so as to obtain reliable data analysis results. Third, for the extraction of image network features, the specific extraction process can refer to the related algorithms in Chapter 2. Fourth, the neural network recognition system is trained according to the existing database, so that it can continue to learn deeply and meet the recognition needs. Fifth, perform feature fusion processing, and then output the corresponding image recognition results, evaluate the accuracy of the recognition results, and continue to train them to obtain higher accuracy information [2].

4.2 Feature content integration
After the RGB image is obtained, the feature content fusion processing needs to be completed according to the correlation. In the specific fusion processing process, the feature dimensions that need to be extracted need to be improved on the original basis to improve the reliability of the processing results. In order to speed up the calculation, the optimal weight algorithm is also used to assist the calculation process. This method is a data feature fusion processing method often used in the process of graphic feature fusion. Its principle is to select information that meets the optimal calculation requirements from the known data set based on the corresponding basic conditions to meet specific application requirements. According to the calculation results, different types of weights can also be obtained, and the main direction of the analyzed data information and the criticality of the characteristic content can be understood, so as to complete the optimal weight calculation [3].

4.3 Pseudo-colorization of the depth map
After completing the above graphics processing, you need to do a good job of pseudo-color processing. Its main function is to obtain the required color image after processing the existing grayscale image, but the color image is not the true color of the original image, so this process is called pseudo-color processing. In the specific processing process, the corresponding temperature data will be directly drawn on the interface during the infrared imaging design process of the system, and then the acquired temperature data will be processed to obtain the corresponding gray-scale image, and finally obtained by the processing method Pseudo-color image [4].

4.4 Simulation experiment analysis

4.4.1 Preparation for simulation experiment
In the construction of this simulation experiment, the RGB-D database is used as the basic reference, and all the pictures in the database are used as the simulation experiment data. In terms of image acquisition, the Kinect sensor is used to capture them from different angles, and the angle is set to 30 °, 60 ° and 90 °, the total number of captured RGB pictures is 35,263, the resolution of each picture is adjusted to 640 × 640, in order to facilitate the subsequent data analysis work in an orderly manner [5].

4.4.2 Simulation experiment process
According to the basic content provided by the simulation experiment environment, in the whole experiment process, first, adjust the size of the pictures in the data set, the specific adjustment size is 256×256, and the feature fusion processing is performed on each picture information, relying on Caffe Framework to obtain the corresponding reference data. Second, use the C++/CUDA architecture to perform in-depth processing, and in the application of the framework structure, the matrix parameters
will also be sorted to obtain a more concise data set. Third, determine the application model of the entire experimental process. In this simulation experiment, the Alex Net model will be used as the main test carrier. The average time of this model on image processing is 1.15ms, with the main frequency 2.90GHZ 16G running content Complete the entire simulation experiment and obtain the accuracy data of image recognition [6].

4.4.3 Arrangement of experimental results
From the experimental results in Table 1, it can be understood that the neural network image recognition algorithm used in this article has achieved 92.3% accuracy in image recognition, while the traditional CNN-RNN algorithm has the lowest accuracy, only 86.3%. The other two types of algorithms are accurate The rate is in the upper middle, and the accuracy of the SN-CNN-SVM algorithm for image recognition has reached 91.1%, and it also has a high image recognition accuracy. However, the algorithm proposed in this article still has a lot of room for improvement. Analyzing the causes of image recognition errors, it is understood that the depth recognition results in the confused images are similar, and the similarity exceeds 60%, which leads to algorithm recognition errors. Therefore, in the follow-up In the development, it is necessary to focus on the research of in-depth identification content to meet the corresponding management needs [7].

| Use algorithm | CNN-RNN algorithm | SP+HMP algorithm | SN-CNN-SVM algorithm | Neural network recognition algorithm |
|---------------|-------------------|------------------|----------------------|-------------------------------------|
| Recognition accuracy | 86.3% | 88.6% | 91.31% | 92.3% |

5. Experimental comparative analysis
According to the above experiment, in the application of Gist image recognition method, the recognition accuracy of the feature image is higher, and in the use process, its overall convenience is relatively high, and it can meet the requirements of image recognition in many situations. The neural network recognition algorithm needs to go through iterative data learning in the application, and the initial state is relatively cumbersome. After a certain number of iterations, its recognition accuracy is relatively stable. In the follow-up research, further discussion of the algorithm application is also needed. Improve the reliability of the analysis results [8].

6. Conclusions
To sum up, in the research process of RGB-D image recognition technology, both feature fusion and machine learning models have good application value. Taking a reasonable way to merge their content can not only speed up image recognition, but also improve The accuracy of image recognition results.

References
[1] Yan Jianwei, Zhang Wenyong, Xie Benliang. Random forest tea fresh leaf classification based on feature fusion [J/OL]. Journal of South China Agricultural University: 1-11 [2021-05-01].
[2] Yuan Mengjiao, Dong Yuning. Network video stream classification based on feature fusion and machine learning [J/OL]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2021(02): 100-108[2021-05-01].
[3] Ma Xiaotian, Wang Chenping, Qiu Yanjun. Prediction of pavement friction performance based on the fusion of macro and micro texture features[J/OL]. Journal of Zhejiang University (Engineering Science Edition): 1-11[2021-05-01].
[4] Feng Kaiyuan, Luo Qingbin, Zheng Minghui, Li Chao. Research on malicious webpage detection method based on multi-feature fusion[J]. Journal of Hubei University for Nationalities (Natural Science Edition), 2021, 39(01): 80-85.
[5] Meng Zhen, Wang Hao, Yu Wei, Deng Sanhong, Zhang Baolong. Research on vocal music classification based on feature fusion [J/OL]. Data analysis and knowledge discovery: 1-22 [2021-05-01].

[6] Ma Jing, Cai Wenjie, Yang Li. Research on Heart Sound Recognition and Classification Based on Machine Learning [J]. Chinese Journal of Medical Physics, 2021, 38(01): 75-79.

[7] Jiang Qianyu, Wang Fengying, Jia Lipeng. Malicious code detection method based on perceptual hash algorithm and feature fusion [J]. Computer Applications, 2021, 41(03): 780-785.

[8] Lu Chujie, Li Sihui. Research on building occupancy rate perception model based on multi-sensor fusion [J]. Computer Applications and Software, 2021, 38(01): 58-62.