APPLICATION OF K-MEANS CLUSTERING IN IDENTIFICATION OF MACHINING LOCATION OF MECHANICAL PARTS

Zhuoyi Tian & Linan Zhong †
Department of Mathematics, College of Science
Yanbian University, Yanji, Jilin 13302

*This work is supported by Natural Science Foundation of China under number 11761072.
† Corresponding author: Li-nan Zhong, E-mail:lnzhong@ybu.edu.cn

ABSTRACT

In this paper , we uses K-means clustering algorithm, feature extraction, curve fitting based on sparse matrix, Hough transform and Sobel edge detection method, the mathematical model is established, and using MATLAB 2016 a, SPSS, such as word calculator software, data, image processing, and for clamping of automatic industrial manufacturing line, automatic computer intelligent identification parts in packaging process location problem, and comprehensive identification of parts the coordinates of each position, given mechanical arm grab behavior parameters, check up the accuracy of the target parts location identification.

Keywords: K-means clustering algorithm, sparse matrix, curve fitting.

INTRODUCTION

Under the new international environment, the Chinese government promotes the implementation of the "made in China 2025" plan based on the general trend of international industrial reform. China needs to make continuous innovation in advanced manufacturing fields such as the intellectualization of major equipment with high technology content. Intelligent transformation based on industrial manufacturing production line is the basis of advanced manufacturing intelligence. In the automatic production line of industrial manufacturing, in the process of clamping and packaging, it is necessary to use the computer to automatically identify the position of parts based on image processing, and the parts are automatically transported to specific positions by the mechanical hand. The accuracy and rapidity of position recognition directly lead to the quality of manufacturing. Accordingly, it is particularly important to study how to use computers to quickly identify parts. FIG. 1 shows the outline diagram of a part, and FIG. 2 shows the position diagram of parts before and after handling. According to the part profile data DATA2 and DATE3, the methods to determine the center of a single or multiple parts and identify the positions of different parts are given.

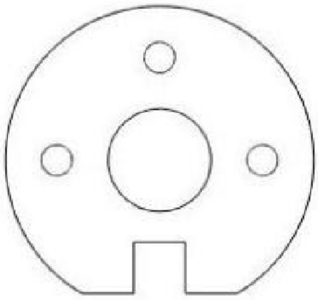

Figure 1 Schematic diagram of part outline

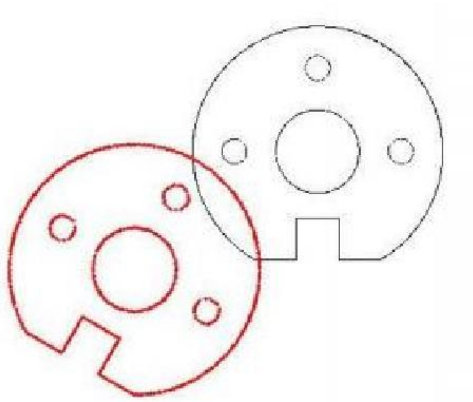

Figure 2 Location drawing of parts
(note: red is the outline of parts placed at any position on the plane; Black is the outline drawing of the standard position of parts)

Establishment and solution of the model

The progress of sparse matrix

The data file DATA1 is the standard position placed after the manipulator grabs the part, and the part contour data can be regarded as the template data of the captured part. After loading the data, it was found that the matrix contained a large number of 0 elements. In order to improve the efficiency of data storage and speed of model operation, sparse matrix was used for processing. If the number of elements with a value of 0 in a matrix is far more than the number of non-zero elements, and the arrangement is irregular and can be found, the matrix is called sparse matrix. Conversely, when the number of non-zero elements is far more than the number of zero elements, the matrix is called a dense matrix. The density of a matrix is defined as the ratio of the number of non-zero elements to the number of elements in the matrix. In the matrix I obtained after the loading of DATA1, most data is 1,0 takes up a small part (0 data is parts). Take the data and conduct sparse processing to obtain the results. It can be seen that matrix I of 140-560 is sparse to obtain 1062 point data, and the data two are greatly reduced.

Detection of outliers

Firstly, the local anomaly factor LOF algorithm based on the distance-based anomaly detection algorithm is considered. This method needs to calculate index data such as the KTH distance and its domain, the accessible distance and its density, and the local discrete factor, and draw a direct view graph of the corresponding data. Therefore, the operation of this method is too complex for the processing of large data sets. However, the distance-based outlier detection method does not need to know the distribution model of data sets, and it is applicable to any dataset that can calculate the distance between objects, and the calculation method is simple. Therefore, this paper adopts this algorithm to detect outliers. If the data Set is the Set of object o more than p part of the object and the distance it all gets d, then the object o is called the data Set on the Set of DB (p, d). Outliers based on the distance of the definition of outliers is applicable to arbitrary dimension data Set, the parameter p indicates that with the objects of distance gets d outliers by the smallest proportion of data sets. The distance-based outlier detection method can easily customize the distance function between objects, and the Euclidean distance calculation function is one of them. The definition of Euclidean distance is as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

Above all, m is the number of dimensions of the data object. The mean steps of distance - based outlier detection algorithm as follows:

(1) Select a data object at random.

(2) Calculates the Euclidean distance between other data and the selected data object. If the proportion of the data object whose distance is greater than d is greater than p, the data object is judged to be an outlier.

(3) Select the next non-repeating data object.

(4) Repeat (2) until all data objects are detected.

After detecting and removing outliers, the position of standard parts is shown in FIG.3.

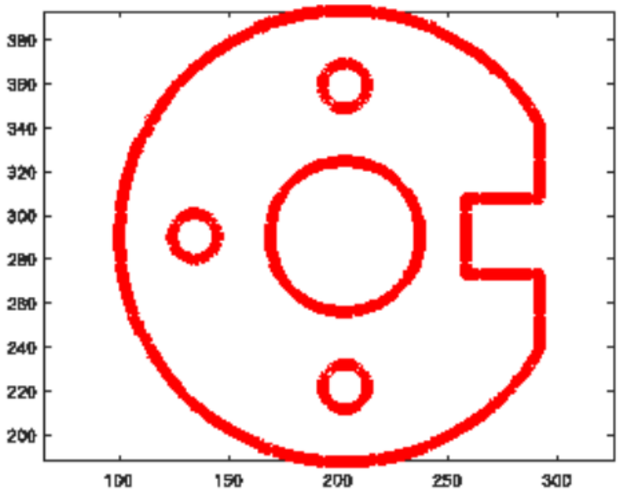

Figure 3 Location of standard parts

Location recognition model

As can be seen from FIG 4, after removing outliers, the shape of standard parts can roughly include three small circles, a large circle, a large arc and an open rectangle. By clustering the parts points, the distribution of parts can be determined.

The core idea of k-means clustering algorithm: randomly select K points from the data set, and each point represents the clustering center of each class. Then calculate the distance from the remaining samples to the clustering center and divide it into the nearest cluster. Then, the average value of each cluster was calculated, and the clustering center was divided again until there was no significant change in the two adjacent clustering centers, indicating that the clustering center converges.

The algorithm's termination can be any one of these:

(1) No data objects are reassigned to different clusters;
(2) The center of the cluster is convergence;
(3) The sum of squared errors are locally minimal;

The steps of K-means algorithm [1] are as follows :

(1) So I'm going to randomly choose a k sample from the data set of D as the initial k of a mass vector : $\{\mu_1, \mu_2 \cdots \mu_k\}$;
(2) When $\mathrm{n} = 1,2, \cdots \mathrm{N}$
   a) Cluster division C is initialized as $C_t = \emptyset$, $t = 1,2, \cdots k$
   b) For $\mathrm{i} = 1,2, \cdots \mathrm{m}$, compute the distance between sample x_i and the vector μ_j $(j = 1,2, \cdots k)$ is $d_{ij} = \|x_i - \mu_j\|_2^2$. x_i is the smallest d_{ij} corresponding category lambda I, or $C_{\lambda_i} = C_{\lambda_i} \cup \{x_i\}$.
   c) Calculate the new center of mass μ_j for all sample points in Cj (j=123).
   d) If all k center of mass vectors haven't changed, then go to step (c).
(3) Output cluster partitioning

The calculations in this article are as follows:[5]

(1) Input N data, and select $\mathrm{K}(k < N)$ data as the initial clustering center.
(2) For the remaining elements, the distance to the center of the clustering point of each cluster was calculated, and the point was divided into the nearest class.
(3) Recalculate the clustering center of each cluster
(4) Compared with the previous center of the cluster, if the center of the cluster changes, go to (2); otherwise, end the iteration and output the result.

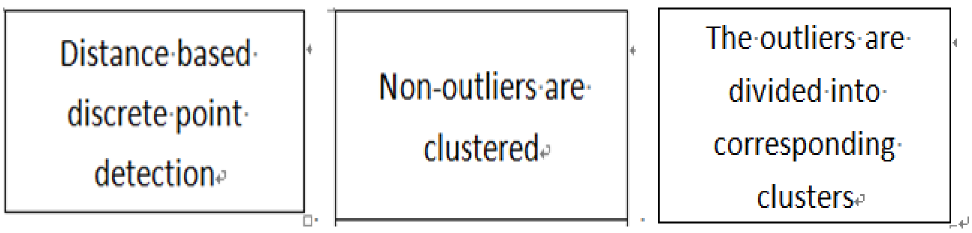

Figure 4 ; K-means algorithm based on outlier detection

Set the total number of clustering categories K=4, and the clustering result is shown in FIG.5. The clustering center C contains four points, the coordinates of which are (120291), (202207), (276293) and (201375) respectively. The iterative process of clustering is shown in FIG. 6. It can be seen that the clustering center is determined after 5 iterations.

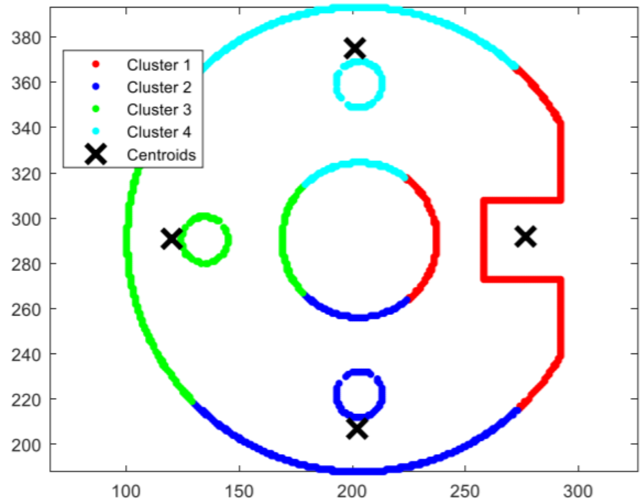

Figure 5 K-means clustering results

```
Replicate 1, 7 iterations, total sum of distances = 49523.
Replicate 2, 8 iterations, total sum of distances = 49523.
Replicate 3, 6 iterations, total sum of distances = 49523.
Replicate 4, 12 iterations, total sum of distances = 49523.
Replicate 5, 8 iterations, total sum of distances = 49521.
Best total sum of distances = 49521
```

Figure 6 the process clustering iterative

Part Angle and position identification

Define the placement Angle of parts as shown in Figure 7. The arrow pointing to the right is 0, which is the standard position as shown in Figure 7.

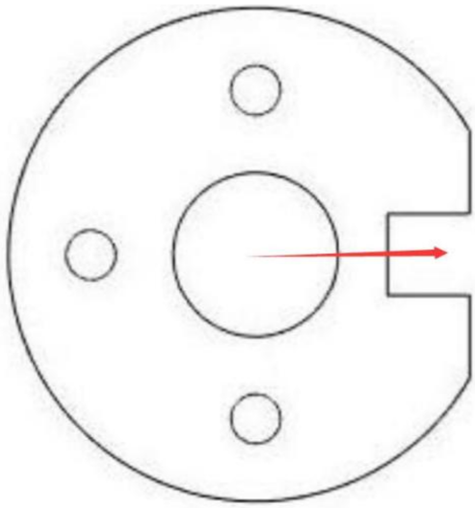

Figure 7 the standard location

As for point A (x_A, y_A), the closest point to T (x_T, y_T) in the clustering center from the center of mass, the Angle between ray TA and positive X-axis is defined as being the Angle of a part. The formula [2] is as follows:

$$\theta = a\tan\left(\frac{y_A - y_T}{x_A - x_T}\right)$$

Where, the coordinate approximation of parts can be expressed by the clustering center point set C (including four pointsC_1 $C_2 C_3$ and C_4), and the calculation formula is as follows:

$$x = \frac{xC_1 + xC_2 + xC_3 + xC_4}{N}$$

$$y = \frac{yC_1 + yC_2 + yC_3 + yC_4}{N}$$

Finally, the center coordinate of the part is (x, y).

By means of K-means clustering recognition algorithm, "DATA2.mat" data is substituted into the data. The coordinate of the center point of the part is (192 , 287) and the part Angle is 115.0836.

Center position identification of multiple parts

Part segmentation based on curve fitting

By identifying the dividing line between parts, four points are randomly selected, the coordinates of which are (262.2809304.5311), (244.4018282.1823), (227.4167246.4242) and (187.1889182.0596), and then the segmentation curve is obtained, and a boundary line is obtained by curve fitting. [7] in general, in practical work, there is not necessarily a linear relationship between variables, such as the relationship between blood concentration and time after medication; The relationship between curative effect of diseases and length of treatment; The relationship between toxicant dose and fatality rate is usually a curve. Curvefitting is to select the appropriate curve type to fit the observation data and use the fitting curve equation to analyze the relationship between the two variables. MATLAB curve fitting function Polyfit is used to obtain the boundary line fitting results, as shown in FIG. 8.

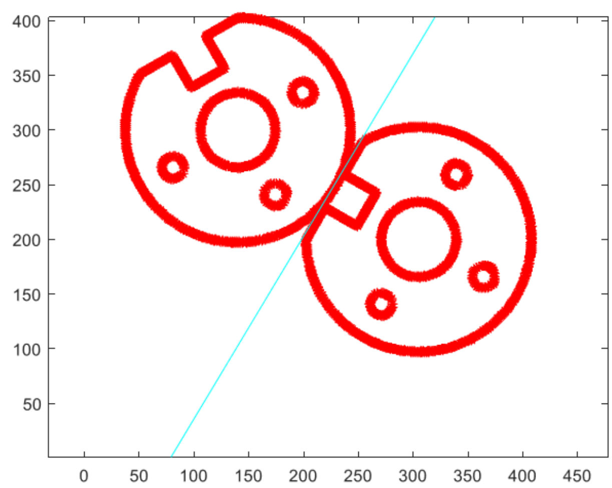

Figure 8 Boundary fitting results

Through modeling and analysis of position recognition of each part, it can be known that the location of parts can be identified efficiently and quickly by adopting K-MEANS clustering recognition algorithm. Clustering feature points of the two parts can be obtained by using K-MEANS clustering recognition algorithm respectively, as shown in FIG. 9. Where, red represents the point set of part 1, green represents the point set of part 2, blue "X" represents the clustering feature point of part 2, and cyan "X" represents the clustering feature point of part 1.

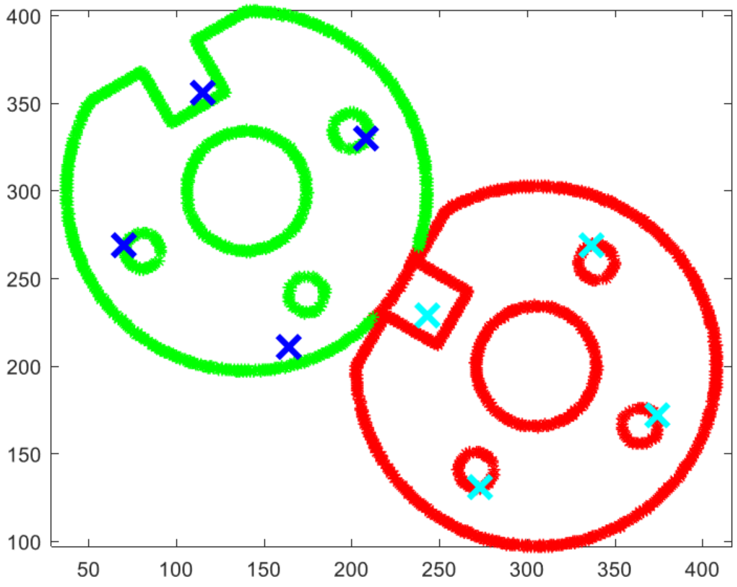

Figure 9 the result of DATA3 clustering recognition

The location information of the two parts obtained by calculation is shown in follow table.

Parts	1	2
Center position	307.75001998750	140.25002936250
Angle	155.1576	112.0386

Model optimization and improvement

Although the model may have some errors, its generalization space is also relatively large. Its edge features and internal features of the extraction and processing, universality is strong, can be applied to many parts of the machine tool processing and packaging. Combined with the application of software, the workload is greatly reduced. It can be seen that this model can be widely used in the identification of related parts in the future, which will provide convenience for similar work. The algorithm is improved to prevent the optimized solution from falling into the local optimal solution

REFERENCES

[1] Liu jianghua. Text retrieval method and validation based on K-means clustering algorithm and LDA subject model [J]. Information science,2017,35(02):16-21+26.
[2] geng qingtian. Research on key technologies of intelligent transportation system

based on image recognition theory [D]. Jilin,2016.
[3]Jiang shuqiang, min weiqing, wang shuhui. Overview and prospect of image recognition technology for intelligent interaction [J]. Computer research and development,2016,53(01):113-122.
[4] zhang jianqing, pan li, wang yinshu, projection measurement: wuhan university press, July 2014
[5] sun xiaoyin. Analysis and research on user behavior of mooc based on k-means [D]. Beijing jiaotong university,2017.
[6] he xiaoqun. Modern statistical analysis methods and applications [M]. Beijing: renmin university of China press, 1998.
[7] Li peipei. Research on fitting, parameterization and shape optimization in curve modeling [D]. Shandong university,2012.
[8] Feng chun. Research on target recognition and positioning based on monocular vision [D]. Nanjing university of aeronautics and astronautics,2013