The Improved DBSCAN Algorithm Study on Maize Purity Identification

Pan Wang, Shuangxi Liu, Mingming Liu, Qinxiang Wang, Jinxing Wang, Chunqing Zhang

The Improved DBSCAN Algorithm Study on Maize Purity Identification

Pan Wang [1], Shuangxi Liu [1], Mingming Liu [1], Qinxiang Wang[1], Jinxing Wang [1], Chunqing Zhang [2*]

[1]College of Mechanical and Electronic Engineering, Shandong Agricultural University, Tai'an 271018, China
[2] College of Agricultural , Shandong Agricultural University, Tai'an 271018, China
{sdwangpan, lentree, lmmjx, wqx, jinxingw} @163.com, cqzhang@sdau.edu.cn

Abstract. In order to identify maize purity rapidly and efficiently, the image processing technology and clustering algorithm were studied and explored in depth focused on the maize seed and characteristics of the seed images. An improved DBSCAN on the basis of farthest first traversal algorithm (FFT) adapting to maize seeds purity identification was proposed in the paper. The color features parameters of the RGB, HIS and Lab color models of maize crown core area were extracted, while H, S and B as to be the effective characteristic vector after data analysis. The abnormal points of different density characteristic vector points were separated by FFT. Then clustering results were combined after local density cluster by DBSCAN. According to the result of test, the method plays a great role in improving the accuracy of maize purity identification.

Keywords. Maize, Purity identification, Density, Color, Core area

Introductions

Maize is one of the major crops in China, which has been used extensively in food and feed field. The purity of corn seed directly is related to the crop yield and quality. In recent years, machine vision technology with its characteristics of objective, non-destructive, rapid speed was widely used for crop detection. It is very significant to research machine vision technology instead of manual testing on maize purity identification. At present, at home and abroad, application of machine vision technology to corn seed appearance characteristics were studied. Foreign scholars used shape parameter to analysis contour features of maize for quality testing and damage identification and identified corn color by chromaticity[2-5].Domestic scholars used the image processing technology to study the features of flank shape[6-9], HIS[8-11], RGB[8-12] and area of corn white part (embryo) or yellow part (crown)[9-10] to distinguish the quality, varieties[13] and purity.

The research on morphological characteristics of corn seed had got good identification results while much less study of corn purity identification and much more application about corn side features. This paper proposed that the color features of crown core area had significant function on corn purity after the study on three commonly used maize varieties. The DBSCAN was optimized by farthest first traversal algorithm for the purity identification. The experimental result indicated this method had high classification rate and the higher precision and offered a reference for building accurate purity identification system.

1Feature extraction and analysis of image

1.1 Image acquisition and processing

The Original image of corn crown captured with CCD camera is processed according to the color character. The image was pre-processed that the background was segmented by R from RGB color model and single colored image was taken by contour labeling. A single crown core area was obtained by taking the single corn crown area centroid as the center, 10 pixels as the radius and signed convenient for purity identification[11].

Took Zhengdan958 for example, the maize original crown image was shown in Fig.1.

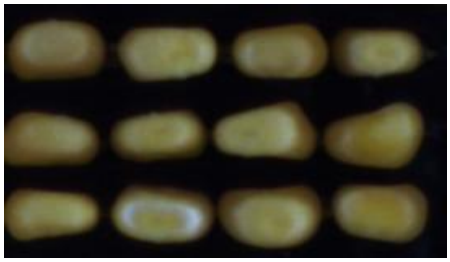
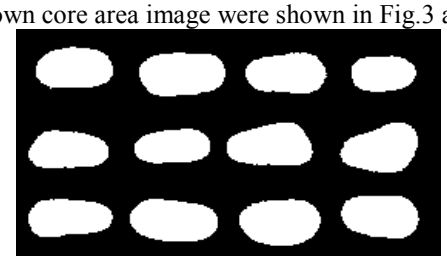
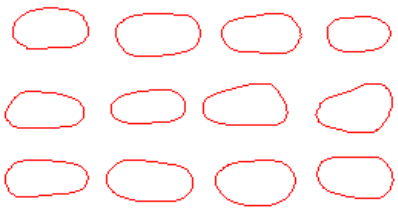


Fig. 1. Original image of Zhengdan958

The pre-processed image was shown in Fig.2, The number mark image and Maize crown core area image were shown in Fig.3 and Fig.4



a Background segmentation image



b Contour markers image

Fig. 2. Preprocessed image

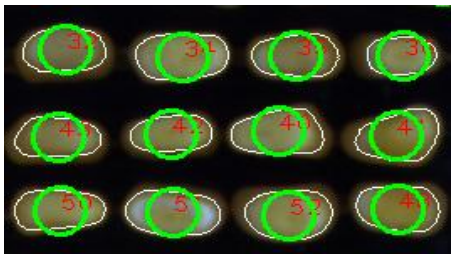


Fig. 3. Number mark image

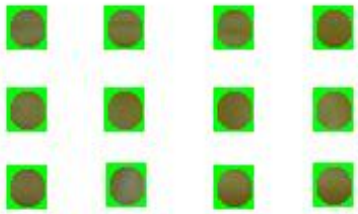


Fig. 4. Maize crown core area image

1.2 Color feature extraction and characteristic vector selection

The main information of corn image is the shape and color features. In a real production process, it is different about the shape features . The study selected three models RGB, HSI and CIE Lab color features to extract nine color R (red), G (green), B (blue), H (hue), S (saturation) , I (brightness), L (luminance), a (red-green), b (yellow-blue) of maize crown core area image. The Eigenvalues were average of all pixels of each corn.

In order to reduce the feature vector dimensions and improve recognition rate, the nine characteristics was studied (combined I and L because of their same meaning).Took Zhengdan958 and Ludan981 for example, the means and standard deviation and coefficient of variation were shown in Tabel1.

Table 1. Color feature parameters statistical data of maize seeds

Varieties	Zhengdan958			Ludan981		
Color features	Means	Standard deviation	Coefficient of variation	Means	Standard deviation	Coefficient of variation
H	64.509	6.106	0.095	70.544	15.660	0.222
S	0.354	0.062	0.176	0.220	0.102	0.464
I	109.820	8.631	0.079	139.477	15.388	0.110
B	71.247	11.167	0.157	110.607	24.0188	0.217
G	130.668	10.742	0.082	154.911	15.3248	0.099
R	126.247	7.013	0.055	151.197	10.1308	0.0679
a	89.043	0.876	0.010	90.662	0.897	0.010
b	174.583	2.208	0.013	169.780	3.805	0.022

Coefficient of variation showed the difference among sample means. For any data set, if the data near the boundary density was relatively discrete, it would have a larger coefficient of variation. Standard deviation declared the between means and eigenvalue, the degree of dispersion between hybrid seeds and normal ones[14]. As can be seen from Table1，H, S, B were contributed to identified , and selected the H, S, B, as identification eigenvectors.

The three-dimensional characteristics of scatter points chat of the three varieties all showed strip fluid-like in figure5(S range of 0 to1, magnified it 100-fold in order to reflect its changing patterns [11])There were individual abnormal points deviating from the sample, and the distribution of feature vectors point had density differences.

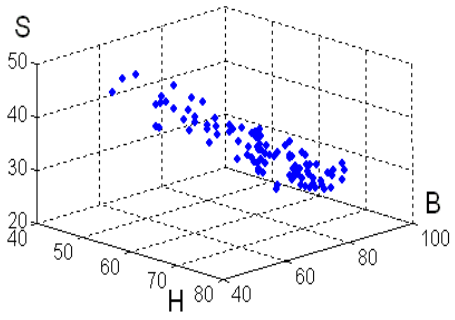


Fig. 5. Three-dimensional diagram of feature vector

2. The corn purity identification based on the farthest first traversal (FFT) optimization DBSCAN

According to the three-dimensional density distribution characteristics of the color

characteristic vector points in the sample, DBSCAN was used for purity identification. In the light of uneven density of the data points, the data was optimized by FFT[15]. The feature vector density difference points in the whole data space were divided into two regions of edge abnormal points that excluded and high density points clustered by DBSCAN, and finally merged cluster results.

This algorithm can roughly be divided into 4 basic steps;

(1) Input the cluster data set S which containing n sample points;

(2)Use the FFT to optimize data. According to the initial clustering centers Z_1 and Z_2, the corresponding dataset is divided into two subsets $Z_1 \in G_1^{(t)}$ and $Z_2 \in G_2^{(t)}$, and then excluded the abnormal points set G_2.

(3) DBSCAN clustering to high density local data set.

(4) Merge the cluster results and get the final cluster result.

2.1 Data optimization

In the process of recognition, it is poor to take a global variable to cluster no uniform density points (H, S, B). Therefore, obviously abnormal data will be excluded before DBSCAN. FFT (Farthest First Traversal Algorithm) can centralize similar samples fleetly and exclude abnormal samples in dataset. The major parameters are the number of clusters (num) and the initial center point (seed). This is two-groups problem, namely, only normal hybrid seed and miscellaneous ones. In order to get accurately results, the mean value was to as the initial center point.

The main algorithm are summarized as following:

(1)Made the num=2，and calculated the mean value point(H, S,B) which was the initial center Z1. $G_1^t$is the dataset whose clustering center was $Z_1^{(t)}$.

(2)(Represented the points (H, S, B) as (S1，S2…Sn), and calculated the distances between the points to the Z1. Euclidean distance was used in this paper.

$$d(S_j, Z_1) = \sqrt{\sum_{j=1}^{n} (S_{ij} - Z_{1i})^2} \quad (1)$$

Where Sij was the i-D coordinate of the jth points, Z_{1i} was the i-D coordinate of the initial center and n is the number of the example. i=1,2；j=1, 2…n

(3)Took the farthest point $\max d(S_j, Z_1)$ far from Z1 as the other center. $G_2^{(t)}$ was the dataset whose clustering center was $Z_2^{(t)}$.

(4) Calculated the distances from the S_j which were remainder points in S to the Z_1 and Z_2, and define the larger one as max(d_{j1}， d_{j2}).All the points in set S were distributed to the most similar group until the end of the cluster.The results was shown in Fig.6

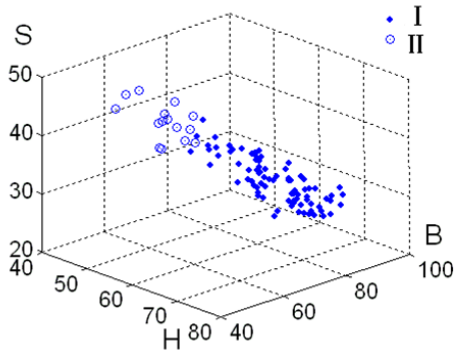


Fig. 6. Data optimization of FFT

2.2 DBSCAN Algorithm

2.2.1About DBSCAN Algorithm

DBSCAN is a clustering algorithm which relies on a density-based notion of clusters. It is designed to discover clusters of arbitrary shape[16]. The key idea in DBSCAN is that for each object of a cluster, the neighborhood of a given radius has to contain at least a minimum number of objects. The procedure for finding a cluster is based on the fact that a cluster is uniquely determined by any of its core objects

(1) Given an arbitrary object S for which the core object condition holds, the set $\{o \mid o > G_{1s}\}$ of all objects o density-reachable from S in D forms a complete cluster C and S∈C.

(2). Given a cluster C and an arbitrary core object S∈C, C in turn equals the set $\{o \mid o > G_{1s}\}$.

2.2.2 Determination of parameters

DBSCAN requires the user to specify the global parameter Eps (The parameter MinPts is fixed to 4 to reduce the computational complexity [17]). In order to determine the Eps, DBSCAN algorithm calculates the distances between all the data objects and found the most adjacent Minpts object. The k-dist diagram was made by the distances sort to confirm the Eps.

The horizontal coordinate of the k-dist diagram represented the number of data objects corresponding to some distance value of the k-dist; the vertical coordinate represented the distance between data objects (H, S and B) and its fourth adjacent objects. The function of k-dist diagram was to determine the most appropriate Eps, namely lower limit density. The distances between the data objects and its kth closest object is ordered by decreasing, so k-dist diagram was also called as sorting k-dist diagram. K-dist diagram was shown in figure 7, while the value of position A in the flat part was set for Eps.

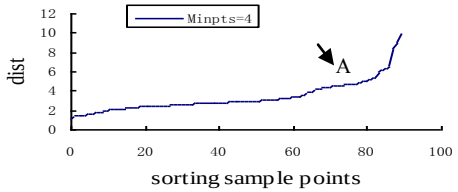


Fig. 7. k-dist diagram

Set(4, 5.2) for (Minpts, Eps), the description of DBSCAN algorithm is as follows:

(1) Choose the object S that belongs to data set G_1 but not any clustering to meeting kernel conditions and create a new clustering.

(2) According to the kernel object of the clustering, collect accessible density of kernel object in a circular until there were no new kernel objects.

(3)The circle will not end until any kernel object does not exist in any clustering, otherwise continue step 1.

In this process, noise points were recorded who's distance from the division boundary were less than Eps while they could be boundary points or the points of some divided cluster.

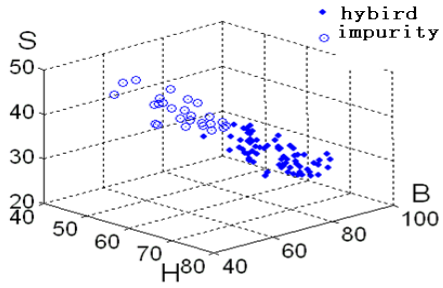


Fig. 8. Clustering results

2.3 Local clustering merger

According to the clustering process, the abnormal scattered points removed by FFT data optimization and noise points or small class points by local clustering had been as false seeds. Take Zhengdan958 for example，there just were small class points. The combining clustering results was as shown in Figure 8.

2.4 Experimental results and analysis

Three common maize varieties Zhengdan958，Nongda108 and Ludan981 were tested to prove the proposed algorithm. The results with the purity identification rate above 93.3% were as shown in Table2 .The basic flow field to meet the. It had met the corn seed purity identified requirements in transaction filed basically.

The features of crown core area were affected to some extent in result of that the crown core area was set by experiences. Artificial placing had some effect on the height of corn crown. The balance of the image acquisition system lighting was also important to the identification results.

Table 2. Maize purity identification result

Varieties	Number	Impurity number in fact	Impurity number by test	identification rate /%
Zhengdan 958	100	31	30	96.8%
Nongda108	100	30	29	96.7%
Ludan981	100	30	28	93.3%

3. Conclusions

(1) The core area of crown was chosen as a research object, in which the color characteristics were studied and H, S, B, was determined as eigenvectors for identified corn seed purity.

(2) Three-dimensional space model of eigenvectors were established. Aiming at the existence of density distribution differences the DBSCAN corn purity recognition algorithm which based on the FFTA was proposed. The algorithm optimized data by FFT, excluded the edge abnormal scatter, local clustered high density areas through DBSCAN, merged cluster region, and finally obtained the results of purity recognition.

(3) The test for purity Identification of three selected maize varieties Zhengdan958，Nongda108 and Ludan981 proved that the method can achieve an average recognition rate of 93.3% or more.

Ackonwledgemants

The authors would like to thank Project supported by the Shandong Province Innovation Fund for Post-doctoral (200903031).

Reference

1. Liu Yande, Ying Yibin, Cheng Fang. Research of Machine Vision in Purity Inspection of Seed [J]. Transactions of the CSAM, 2003, 34(5): 161~163.
2. Liao K, Paulsen M R, Reid J F, Ni B C, Bonifacio-Magirang E P. Corn kernel breakage classification by machine vision using a neural network classifier [J]. Transactions of the ASAE, 1993, 36 (6):1949~1953.
3. Ni B, Paulsen M R, Reid J F. Corn kernel crown shape identification using image processing [J]. Transactions of the ASAE, 1996, 40(3): 833~838.
4. Paliwal J, Shashidhar N S, Jayas D S. Grain kernel identification using kernel signature [J]. Transactions of the ASAE, 1999, 42(6):1921~1924.
5. Liu J, Paulsen M R. Corn whiteness measurement and classification using machine vision [J]. Transactions of the ASAE, 2000,43(3):757~763.
6. Ning Jifeng, He Dongjian, Yang Shuqin. Identification of tip cap and germ surface of corn kernel using computer vision [J].Transactions of the CSAE, 2004, 20(3):117~119.
7. Hao Jianping, Yang Jinzhong, Du Tianqing, Cui Fuzhu, Sang Suping. A Study on Basic Morphologic Information and Classification of Maize Cultivars Based on Seed Image Process[J]. Scientia Agricultura Sinica, 2008, 41(4): 994~1002.
8. Cheng Hong, Shi Zhixing, Me Wei, Wang Lei, Pang Lixin. Corn Breed Recognition Based on Support Vector Machine [J]. Transactions of the CSAM, 2009, 40(3): 180~183.
9. Wang Yuliang, Liu Xianxi, Su Qingtang, Wang Zhaona. Maize seeds varieties identification based on multi-object feature extraction and optimized neural network[J]. Transactions of the CSAE, 2010, 26(6): 199~204.
10. Shi Zhixing, Cheng Hong, Li Jiangtao, Feng Juan. Characteristic parameters to identify varieties of corn seeds by image processing [J].Transactions of the CSAE, 2008, 24(6): 193~195.
11. Yan Xiaomei, Liu Shuangxi, Zhang Chunqing, Wang jinxing. Purity identification of maize seed based on color characteristics [J].Transactions of the CSAE, 2010, 26(Supp.1): 46~50.
12. Song Peng, Wu Kebin, Zhang Junxiong, Li Wei, Fang Xianfa. Sorting System of Maize Haploid Kernels Based on Computer Vision [J].Transactions of the CSAM, 2010, 41(Supp):249~252.
13. Quan Longzhe, Zhu Rongxin, Lei Pu, Han Bao. Recognition Method of Maize Cultivars Based on K-L Transform and LS-SVM [J]. Transactions of the CSAM, 2010, 41(4): 168~172.
14. Xue Lixiang, Qiu Baozhi. Boundary Points Detection Algorithm Based on Coefficient of Variation [J]. PR＆AT, 2009, 22(5): 799~802.
15. Gonzalez, T. F. Clustering to minimize the maximum intercluster distance [J]. Theoretical Computer Science, 1985, 38: 293~306.

16. Feng Shaorong, Xiao Wenjun. An Improved DBSCAN Clustering Algorithm [J]. Journal of China University of Mining&Technology, 2008, 37(1): 106~111.
17. Yue S H, Li P, Guo J D, Zhou S G. A statistical information based clustering approach in distance space [J].Journal of Zhejiang University Science, 2005, 6A(1):71 –78.