Automatic Identification Of Construction Dust Based On Improved K-Means Algorithm

Fei Lei, Xiaohe Ma, Xueying Dong
Beijing University of Technology, Beijing 100124, China
S201861346@emails.bjut.edu.cn

Abstract. At present, the construction dust has caused great losses to people's healthy life and national economic development. In order to solve the shortcomings of the existing real-time sensor network detection of construction dust such as poor accuracy, this paper proposes an automatic identification of construction dust based on computer vision and improved K-Means clustering algorithm. We extract the saturation of HSV color model of each image to form a text data set. We determine the median value of the data set as the initial centroid of clustering, reduce the number of iterations and get the global optimal solution. Mahalanobis distance is used as similarity measure to cluster, which reduces the difference between different feature measures, improves the accuracy, and realizes the automatic recognition of construction dust based on computer vision. For the improved K-Means algorithm, the precision, recall rate and harmonic mean value are used to analyze the clustering results. The experimental results show that the improved K-Means algorithm has good robustness and high accuracy, and the automatic recognition rate can reach 89.33%.

1. Introduction
With the rapid development of economy, the process of urbanization is accelerating, and the construction of urban infrastructure is more and more. However, due to the influence of traditional production mode, the management mode of construction site is more bad. In the process of construction, a lot of dust will be produced, which will bring serious damage to the environment of the city. From the point of view of environmental pollution, Hui Yan[1] et.al showed that the concentrations of suspended particulates, PM2.5 and PM10 in construction sites were 42.24%, 19.76% and 16.27% higher than those in the surrounding environment, respectively. Construction dust is also one of the main factors of haze formation. Smaoui Nour[2] et.al have shown that dust emission is also a major factor threatening the health of construction workers and managers. The concentration of dust in the environment is directly proportional to the incidence rate of respiratory system. The concentration of dust in the air increases and the chance of illness increases. Dust reduces the visibility of the surrounding air, increases the pressure of urban traffic and reduces the visual beauty of people's life. At the same time, dust contains a variety of heavy metal elements, dust with the accumulation of heavy metal elements on the surface of plant leaves, hindering photosynthesis and respiration of plants, affecting the growth of plants. The dust from the construction site enters the equipment, which shortens the service life of the machine and increases the construction cost[3].

In view of the above hazards of construction dust, it is great significant for the urban environment and the health of construction workers to detect the dust in the construction site and take timely measures to reduce the dust. In order to reduce the emission of construction dust, the government has promulgated various impact standards for construction sites and construction dust detection method[3].
At present, weight method is the only method with legal effect in China. The weight method is to calculate the weight difference of filter membrane before and after sampling, and obtain the concentration of construction dust through conversion relationship. The process of this method is complicated, the workload is large, and the degree of automation is relatively low. On construction sites, real-time sensor networks are usually used to detect the emission of construction dust. The real-time sensor network monitors the emission of construction dust. If the emission of construction dust reaches a certain concentration, it will send intelligent instructions to the dust reduction facilities to reduce the concentration of construction dust. However, the maintenance cost of real-time sensor network detection equipment is high and the accuracy is poor[4-6].

With the development of computer vision and large capacity digital image memory, the use of computer vision and machine learning for image classification and recognition is becoming more and more popular[7-8]. In view of the shortcomings of the existing dust detection technology in sensor networks, this paper proposes an improved K-Means algorithm based on computer vision to realize the automatic identification of construction dust. This economic and efficient method of construction dust detection greatly facilitates the monitoring personnel.

The section one of this paper describes the hazards of construction dust and the limitations of the current research status. The section two discusses the overall design scheme. The section three describes the extraction of image features and the improved experimental algorithm. The section four mainly analyzes the results of our improved algorithm and the traditional method. The section five summarizes the actual contribution of this study, as well as the future research prospects.

2. Scheme Overview

The principle of K-Means clustering algorithm is simple and easy to understand, which it is suitable for clustering experiments of a large number image data. In outdoor vision system, a large number of images data need to be cluster analyzed to automatically recognition the images with construction dust. Therefore, this paper uses K-Means clustering algorithm to realize the images with construction dust. The scheme flow chart is shown in Fig.1.

Figure 1. Scheme flow chart
3. Feature extraction and image recognition

3.1. Data sources
At the beginning of the experiment, we went to the Beijing construction site to collect images from time to time, as shown in Fig.2, each image was taken at different time, angle and place. It ensures the comprehensiveness and coverage of the data set. We collected 150 images with construction dust and 150 images without construction dust. The resolution of each image is 500 * 500.

3.2. Image processing and feature extraction
Color feature is one of the most basic features of image. At the same time, color feature is the most widely used feature in image classification, which is easy to understand and extract. Compared with the surrounding environment, the construction dust has a great color difference. HSV is a color model that can clearly express human visual perception and characteristics. The hue, saturation and value of HSV color model are consistent with human visual sense\[9-10\].

The images collected in the early stage are RGB color images. We need to convert the RGB color model of the images into HSV color models. This paper compares the image features of HSV color model and HSV color model separated the three channels from the sample images. We found that the saturation of the images with construction dust was significantly lower than that without construction dust. As shown in the comparison in Fig.3. Therefore, the mean value and variance of saturation components are selected as feature to realize the images with construction dust.
According to the transformation relationship between RGB model and HSV model, the saturation component of the image is extracted as shown in formula (1)\cite{10}. The mean value $A$ and variance $B$ of image saturation are used as features to form coordinate points $(A_N, B_N)$ ($N$ represents the total number of samples) representing the features of each image. The coordinate points of all images in the sample constitute a data set $C$.

$$S = \begin{cases} 
1 - \frac{\max(R,G,B) - \min(R,G,B)}{\max(R,G,B)} & \max(R,G,B) \neq 0 \\
0 & \max(R,G,B) = 0
\end{cases} \quad (1)$$

3.3. K-Means clustering algorithm recognition

3.3.1. K-Means clustering algorithm

K-Means is an unsupervised iterative clustering algorithm. K-Means algorithm is based on the distance between the sample data point and the centroid point, and divides the sample into the nearest class. The standard measure function $M$ is calculated as formula (2). The smaller the value of the standard measure function $M$, the smaller the distance between the sample and the cluster centroid, the better the clustering effect. The $M$ value of two adjacent times is unchanged, and the clustering results are output.
In the formula (2), $D_\tau (\tau = 1, 2, 3, \ldots K)$ represents the class $\tau$ after clustering, $X$ represents the sample data points in class $D_\tau$, and $\mu_\tau$ represents the mean value of sample data points in class $D_\tau$.

### 3.3.2. Improved K-Means clustering algorithm

The stability of K-Means algorithm is affected by the selection of initial centroid. Random selection of the initial centroid may increase the number of iterations and fall into the local minimum solution. The optimal solution cannot be reached\(^{[11]}\). In view of the above problems, our improved K-Means clustering algorithm selects the initial centroid.

We use the median of the sample data set as the initial centroid of clustering, and we arrange the coordinate points of the data set in ascending order as formula (3). The initial centroid clustering of two classes are centroid1 (0.282410, 0.037153) and centroid2 (0.287985, 0.037225).

$$\overline{\mu} = \begin{cases} X_{(n+1)/2} & \text{n is odd} \\ \frac{X_{n/2} + X_{(n+1)/2}}{2} & \text{n is even} \end{cases}$$

In the formula (3), the median of the data set $\overline{\mu}$, $X$ represents the sample data points, and $N$ represents the total number of samples.

The traditional K-Means clustering algorithm uses Euclidean distance $d_1$ such as formula (4) as the measurement of similarity. The contribution of different dimensions of each coordinate point to Euclidean distance is the same, which leads to the difference between different feature metrics, which leads to judgment errors and seriously affects the accuracy of the experiment\(^{[12]}\). In this paper, Mahalanobis distance $d_2$, such as formula (5), is used as the similarity measurement method to normalize the variance of each dimension data of coordinate points. The feature relationship is in line with the actual situation to improve the accuracy of clustering results\(^{[13]}\).

$$d_1 = \sqrt{(x_{ix} - \mu_{tx})^2 + (x_{iy} - \mu_{ty})^2}$$

$$d_2 = \sqrt{(X - \mu_{x})^T V^{-1} (X - \mu_{x})}$$

In formula (4), $x_{ix}$ and $x_{iy}$ represent the abscissa and ordinate of data point $X$. $\mu_{tx}$, $\mu_{ty}$ represent the abscissa and ordinate of the class where the data point $X$ is located. In formula (5), $X$ is the sample data point, $\mu_{x}$ is the centroid of the class where $X$ belongs, and $V$ is the covariance matrix of all sample data points in the class where data point $X$ belongs.

### 4. Feature extraction and image recognition

#### 4.1. Experimental evaluation method

We have done four clustering experiments to determine whether the initial centroid of clustering and different metric distances. The experimental results were evaluated by precision, recall rate and harmonic mean value $F1$\(^{[14]}\).

$$Right(x_i, x_j) = \begin{cases} 1 & L_{(c_1)-D_{(c_2)}} = D_{(c_1)} \\ 0 & \text{others} \end{cases}$$
In the formula, \( X = \{X_1, X_2, X_3, \ldots, X_n\} \) represents all sample data, where \( n \) is the number of class data sample points, \( X_j \) (1\( \leq j \leq n \)) is the determined category \( L(X_j) \), \( D(X_j) \) is the category after \( X_j \) clustering, where \( X_i, X_j \) (1\( \leq i, j \leq n, i \neq j \)). \( F_1 \) represents the harmonic mean value of evaluation index.

### 4.2. Experimental results

The comparison of our four experimental results is shown in Table 1. In this paper, experiments 1 and 3 are using Mahalanobis distance and Euclidean distance respectively under the initial centroid of random selection. Experiment 2 and experiment 4, we use Mahalanobis distance and Euclidean distance to measure the classification, and the median value of the data set is taken as the initial clustering centroid. The evaluation data show that the precision, recall rate and harmonic average value of the improved K-Means clustering algorithm are better.

| Experiment number | Evaluation data |
|-------------------|-----------------|
|                   | Precision | Recall rate | Harmonic mean value |
| 1                 | 76.28%     | 79.33%     | 0.7778               |
| 2                 | 91.28%     | 89.33%     | 0.9029               |
| 3                 | 75.34%     | 73.33%     | 0.7432               |
| 4                 | 78.94%     | 80.00%     | 0.7947               |

Compared with experiments 1 and 2, experiments 3 and 4, the results show that the recognition rate is higher when the initial centroid is selected. The reason for the improvement of recognition rate is that the determination of initial centroid eliminates the local minimum solution and achieves the global optimal solution. At the same time, compared with experiments 1 and 3, experiments 2 and 4, Mahalanobis distance has higher recognition rate than Euclidean distance. The reason is that Mahalanobis distance normalizes the variance of each dimension of coordinate points, reduces the difference between different feature quantity standards, eliminates the interference of correlation between variables, and improves the accuracy of clustering algorithm.

The precision of clustering evaluation shows that the algorithm can distinguish images without construction dust. The higher the precision, the stronger the discrimination ability. Recall rate is the ability to identify the image with construction dust. The higher the recall rate, the higher the recognition ability of the algorithm. The larger the harmonic mean value, the more robust the algorithm is\(^{[15]}\). The table data show that the improved K-Means algorithm has more accuracy recognition rate and better robustness.
According to the evaluation results, the best clustering effect is obtained by Mahalanobis distance when the initial centroid of clustering is determined in Experiment 2. The accuracy of automatic identification of construction dust is 89.33%. Fig.4 shows the distribution map of sample coordinate data before clustering, and Fig.5 shows the distribution map of coordinate points of sample data after clustering, which determines the initial centroid and measures clustering by Mahalanobis distance. In Fig.4 and Fig.5, the green coordinate points represent the mean value and variance of the saturation of the image with construction dust, and the red coordinate point represents the mean value and variance of the saturation of the image without construction dust.

We also compared the running time of four experiments. The running time of the improved K-Means algorithm is shorter than the traditional K-Means algorithm. The improved K-Means algorithm with median as the initial centroid can reduce the number of iterations. Mahalanobis distance eliminates the interference of variable correlation, and its running time is faster than that of European distance. K-Means algorithm, which determines the initial centroid and use Mahalanobis distance as the similarity measurement, also has the advantage in running speed. As shown in Fig.6.

![Figure 4. Data coordinate point](image1)

![Figure 5. Results of Experiment 2](image2)

![Figure 6. Running time comparison of four experiments](image3)

5. Conclusion and outlook

At present, there are few researches on the application of image processing and computer vision for the detection of construction dust. This paper proposes K-Means clustering algorithm to determine the initial clustering centroid and use Mahalanobis distance as the similarity measurement. The improved K-Means clustering algorithm can automatically identify and cluster the images of construction dust. However, the extraction of saturation as a feature of image recognition classification has poor recognition ability for low concentration construction dust. We can increase the amount of data
samples, and have a more detailed clustering analysis on the clustering of construction dust, and divide the construction dust into high concentration and low concentration, so as to promote the development of intelligent detection of construction dust. It lays the foundation for designing an easy to operate construction dust detection system by using camera directly.

References
[1] Yan H, Ding G, Li H, et al. Field Evaluation of the Dust Impacts from Construction Sites on Surrounding Areas: A City Case Study in China[J]. Sustainability, 2019, 11(7).
[2] Smaoui N, Kim K, Gnawali O, et al. Respirable Dust Monitoring in Construction Sites and Visualization in Building Information Modeling Using Real-time Sensor Data[J]. Sensors and materials, 2018, 30(8(1)):1775-1786.
[3] Xing, Jinding, Ye, et al. Control Dust Pollution on Construction Sites: What Governments Do in China?[J]. Sustainability, 2018.
[4] Bekaroo G, Bokhoree C, Pattinson C. Towards Green IT Organisations: A Framework for Energy Consumption and Reduction[J]. International Journal of Technology Knowledge & Society, 2012(3):23-36.
[5] Fan S, Wong Y W, Shen L, et al. The effectiveness of DustBubbles on dust control in the process of concrete drilling[J]. Safety Science, 2012, 50(5):1284-1289.
[6] Wong M S, Mok E, Wang T, et al. Development of an Integrated Micro-Environmental Monitoring System for Construction Sites[J]. Procedia Environmental Sciences, 2016, 36:207-214.
[7] Zhang S, Wu Y, Chang J. Survey of Image Recognition Algorithms[C]/ 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2020.
[8] Kachouri, Djemal, Maaref, et al. Content description and classification for Image recognition system[C]/ International Conference on Information & Communication Technologies: from Theory to Applications. IEEE, 2008.
[9] Hsu M C, Lo Y S, Lin C H. Retinex image enhancement based on exposure fusion[C]/ International Conference on Intelligent Green Building & Smart Grid. 2018.
[10] Wu S, Chen H, Zhao Z, et al. An Improved Remote Sensing Image Classification Based on K-Means Using HSV Color Feature[C]/ Tenth International Conference on Computational Intelligence & Security. IEEE, 2015.
[11] Mahmud M S, Rahman M M, Akhtar M N. Improvement of K-means clustering algorithm with better initial centroids based on weighted average[C]/ International Conference on Electrical & Computer Engineering. 2012.
[12] J. Gan, A. Li, Q. Lei, H. Ren and Y. Yang, "K-means based on active learning for support vector machine," 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, 2017, pp. 727-731, doi: 10.1109/ICIS.2017.7960089.
[13] Bhamre T, Zhao Z, Singer A. Mahalanobis distance for class averaging of cryo-EM images[C]/ IEEE International Symposium on Biomedical Imaging. IEEE, 2017.
[14] Olukanmi P O, Twala B. Sensitivity analysis of an outlier-aware k-means clustering algorithm[C]/ 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech). 2017.
[15] Yu M, Zhang J, Wang J, et al. The Research of Spam Web Page Detection Method Based on Web Page Differentiation and Concrete Cluster Centers[M]/ Wireless Algorithms, Systems, and Applications. 2018.