Acta Technica *62* No. 1B/2017, 187–196

A computer behavioral analysis algorithm based on image color classification statistics[1]

Jia Wang[2]

Abstract. At present, the analysis of computer operation behavior is mainly through the artificial desktop monitoring, recording button content, file operation records and other ways to monitor and analyze the behavior of the computer. In order to study how to use image color classification statistics to analyze computer behavior, a method based on color image clustering analysis and K-means algorithm for statistical analysis of graphics to achieve behavior analysis was proposed in this paper. And according to the type, range and depth of the color, the computer operation images were classified and counted. In this way, the entertainment, work and other operations performed by the computer at a given point of time were determined. The final experimental results show that the algorithm is more practical, and the effect of the deep colored game and movie analysis is remarkable.

Key words. Color image clustering algorithm, K-means algorithm, behavior analysis.

1. Introduction

With the continuous development of science and technology and the continuous expansion of computer applications, the image processing method came into being. The purpose of which is to classify the image and analyze the image information intelligently by using computer equipment. Nowadays, the image processing and recognition have been applied more and more widely. However, as far as the current level is concerned, the computer's perception of the external is still relatively weak, and a lot of manpower and material resources are needed to study the theory and application of digital image processing and recognition. As a result, all walks of life have greater demand for the accuracy and intelligence of digital image processing technology, especially in aerospace, biomedical engineering, industrial testing, robot

[1]This work is supported by Project from The Applied Basic Research Youth Programs of Science and Technology Department of Yunnan Province, Analysis on Screen Soft Proofing based on ICC Profile (No .2015FD039)

[2]Yunnan Open University, Kunming City, Yunnan Province, 650500 China

vision, public security, justice, culture, art and other fields. In this paper, according to the characteristics of digital image processing technology, it was applied to the analysis of universal computer operation behavior, and an innovative function of intelligent analysis of its operation behavior based on the image color appearing in the process of computer operation was realized.

2. State of the art

K-means algorithm is the most common classical algorithm in data clustering analysis, which has been widely used because it has the advantages of short time and simple algorithm when clustering data [1]. However, there are some shortcomings in the algorithm, for example, high-dimensional and non-spherical data are difficult to be clustered, and greatly affected by the selection of initial cluster centers during clustering, which will fall into the local optimal solution during the solution process, so that the clustering result is not good, and the number K of cluster centers needs be determined before clustering analysis [2]. Therefore, many experts and scholars have carried out a more in-depth study of the K-means algorithm, so as to find out the new algorithm to make up for the shortcomings of the K-means algorithm. When some commonly used clustering algorithms are used to deal with some high-dimensional complex data and nonlinear data, there are some problems such as long computing time, low efficiency and low accuracy of clustering. Therefore, a new distance cost function is proposed, which can calculate all kinds of data, including intra class and inter class data distance, and the computing efficiency is higher [3]. The division of data is done by the principle of maximum and minimum distance. After the data is partitioned, the required K value can be automatically determined without the user's prior experience. There is a clustering algorithm, which can determine the optimal number of clusters K_{opt} for the selected data. The specific process is as follows: firstly, the upper and lower boundaries of the selected object are determined, thus effectively reducing the scope of cluster search. Then in the algorithm, the k parameter should be set up, and $K_{\min}$ is usually set to 2. The setting of $K_{\max}$ needs to be determined according to the type of data selected and the data AP, and then the optimal clustering number K_{opt} is determined by the Silhouteet clustering validity index [4]. An improved k-means algorithm is proposed for isolated data, the main idea of which is to determine the initial clustering center based on the average distance method. Firstly, the isolated data is removed from the general data, and then the distance between the remaining data is calculated, and a set of data with the smallest distance in these data is selected. The center of the set of data is used as the initial clustering center. Then the same method is used to obtain the next smallest set of data between the data sets, and the data in the set is also chosen. The distance between the two cluster centers is calculated, and finally whether the distance is less than the average distance between all the data is determined. If it is less than the clustering among all the data, the center is taken as the second clustering center, and this is continued to find K clustering centers [5].

3. Methodology

Color image clustering analysis and K-means algorithm are the keys of this topic. In this paper, through color image clustering analysis and K-means algorithm, the image was transformed into 3 kinds of modules, which were normal image, rendering image and gray image. These 3 types of modules were clustered simultaneously, and its influence color block interval and influence depth interval that can affect behavioral analysis were worked out. Finally, based on the combination of two kinds of intervals, the original image was judged according to the threshold value, and the analysis conclusion of operation behavior was obtained [6].

K-means algorithm is a traditional and classic data clustering algorithm, the principle of which is to analyze the similarity between different spatial data by Euclidean distance [7]. The algorithm's data clustering process is: firstly, the data to be clustered is divided, and the K value of the algorithm is set. The data is divided into k families by means of K values, and the principle of similarity is used to make the data in the dataset divided into K-class families, and the partitioned data is similar in the same class. The similarity of data in different classes is small, but the algorithm is used to cluster data [8].

If Q is a finite set $X = \{x_1, x_2, \ \cdots, x_n\}$ of spatial S^Q, the initialization is randomly divided into K, remarked as $C_1, C_2, \cdots, C_K$. If there are n objects in a class, the clustering center of class i is defined as $Z_1, Z_2, \cdots, Z_K$, as is given by the expression

$$Z_i = \frac{1}{n} \sum_{j=1}^{n} x_j, \ j \in [1, K] \,. \tag{1}$$

The defined objective function is given by the expression

$$J = \sum_{i=1}^{K} \sum_{j=1}^{n_j} D^2_{x_j, z_i} \,. \tag{2}$$

$D^2_{x_j, z_i}$ represents the distance from the first j text to the cluster center of class i, that is, the Euclidean distance.

There are three main steps in K-means algorithms running.

Step 1: K objects are selected from the finite set X as the clustering center.

Step 2: the distance between objects are calculated according to the Euclidean distance, and the objects of the same distance are divided into the corresponding clusters.

Step 3: according to formula (2), the cluster center is continued to calculate, and step 2 is repeated until the algorithm converges [9]. The flow chart is shown in Fig. 1.

K-means algorithm is a classical clustering algorithm for data processing, which usually requires two stages of processing when dealing with data. These two processes are: firstly, the data algorithm is set by class, and is equally allocated to these classes. Then, the spatial distances of the data are computed. It can be found that if the distance is shorter, the similarity will be high, the distance between races will

be large, and the similarity will be low. If it is found that the initial clustering center is not good, then the cluster center needs modified [10].

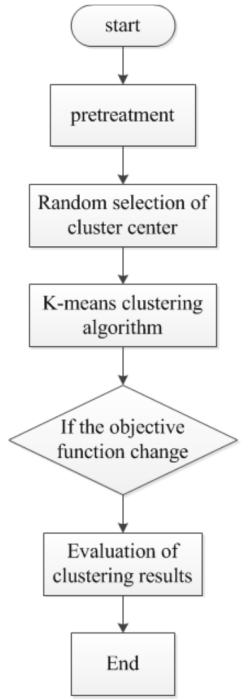

Fig. 1. K-means algorithm flow chart

In order to implement the algorithm in this paper, firstly, the computer image and the size of the current form were obtained. Then an image with the current form as a template and a bitmap Bitmap drawing surface were created. The handle of the form and the handle to the image were obtained. The clods of raster operation code was copied, the API function was called, so as to achieve a form of capture, release the handle, and save the image [11].

In this example, the image results are shown in Fig. 2.

According to the behavior analysis image - custom color of the system, the image was unified into RGB format, the pixels of the image was traversed, the type and proportion of colors in the image and the position of the color system were obtained, and the weight of the color system in which the main colors were located was counted (Sun et al. 2008) [12]. The image content was analyzed according to the custom color model. Based on the RGB format of the color image, the three-dimensional coordinate map was established, among them, R was the X axis, G was the Y axis, B was the Z axis, and the coordinate axis length was 255. According to the slight difference of the approximate color, the three channel pigment was divided into 14 colors - 14 color blocks. After that, according to the temperature and the brightness of the color blocks, the existing operation behavior and image results were matched, and the specific threshold was obtained (Zhang et al. 2011) [13].

Fig. 2. Original drawing obtained by the computer

Figure 3 shows a partial chromatic graph of a partition. According to the color image, all pixels were planned into 14 color blocks. After that, all pixel points were traversed and incorporated into class groups. The pixels of the same class were rendered to the same color.

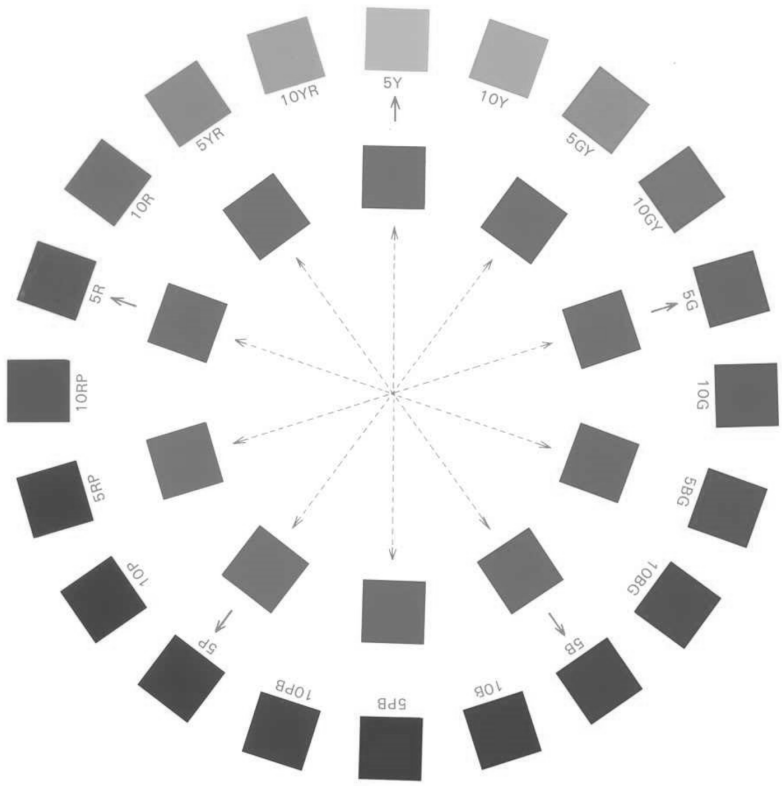

Fig. 3. Part divided color map

Figure 4 shows the rendering. Through repeated test experience accumulation, the six color areas affected by the color proportion in the color block were analyzed according to the lightness, temperature, hue and brightness of the color. An influence color interval was constructed, so as to store the pixels in the six color zones that can ultimately affect the results of the behavioral analysis. And then these six color areas were counted. When the proportion of pixels in a color area exceeded the specified threshold (this threshold was a range of behavior color distinction boundary summed up after repeated trials, pictures, analyses, calculations), the proportion was saved in the influence color interval (Zhang et al. 2011) [14]. Figure 5 shows the proportion of dominant colors.

Fig. 4. Clustering renderings

Fig. 5. Scale diagram of main color system

In this research, the clustering analysis of gray image is mainly based on K-means algorithm. The advantages of the adopted K-means clustering algorithm are mainly concentrated in: the algorithm is fast and simple; for large data sets, it is more efficient and scalable; the time complexity is near linear, and it is suitable

for mining large-scale datasets [15]. The time complexity of the K-means clustering algorithm is O (nKt), among them, n represents the number of objects in the dataset, t represents the number of iterations of the algorithm, and K represents the number of clusters.

The specific process of K-means algorithm consists of the following steps:

The first step is to randomly select K clusters: $\mu_1, \mu_2, \cdots, \mu_K \in R^n$.

The second step is to repeat the following process until it converges.

For each sample i, the class that should belong to is calculated, as shown in the following expression

$$c^{(i)} = \arg\min_j \left\| x^{(i)} - \mu_j \right\|^2 . \tag{3}$$

For each class j, the centroid of the class is recalculated, as shown in the formula

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}} . \tag{4}$$

In the K-means algorithm, the selection of K is manually specified, and the selection of the K value is very difficult to estimate. The summary was made on the basis of a large number of raw data experimental results in this paper, and it was concluded that when the K value was selected to be 5, the operation performance analysis had the best precision.

In the K-means algorithm, firstly, an initial partition is determined according to the initial clustering center, and then the initial partition is optimized. The choice of the initial cluster center has a great impact on the clustering results. Once the initial value is not selected properly, the clustering results may not be valid. In this study, the initial clustering center was set according to the center of the position of the dominant color system in which the color image was the largest.

4. Result analysis and discussion

The K-means algorithm was used in this paper, and through Table 1 and Table 2, further analysis was carried out.

Table 1. The result analysis table of League of Legends

	Entertainment	Office learning	Other
The proportion	79 %	8 %	13 %

Table 2. The result analysis table of word

	Entertainment	Office learning	Other
The proportion	7 %	69 %	24 %

The entertainment games, movies, office software and learning software which were frequently occurred in the market were tested in this study, the K-means algorithm was more practical. The analysis of deep colored movies and games was very effective, such as the League of Legends shown in Fig. 1.

For light colored games, the analysis was disturbed in a particular scene, such as JX Online Version Three, referred to as the JX Three. The image color range of the game was fluctuated after a halo treatment of the picture. In view of the phenomenon, a lot of picture testing was carried out, and the game image was summarized, thereby providing a very good data correction for the future algorithm modification.

In addition, in traditional K-means algorithms, Euclidean distance is used as similarity measure. From the characteristics of Euclidean distance, the Euclidean distance can effectively measure the similarity between the spherical data with uniform distribution, while the Euclidean distance cannot effectively measure the similarity relation between inhomogeneous and non-spherical data. In other words, for each attribute of each data, Euclidean distance treats it the same, and the weights are same. However, in practical problems, the different attributes of data have different effects on the results of data. Therefore, one of the main drawbacks of using Euclidean distance as a similarity measure is that the traditional K-means algorithm is not suitable for non-spherical data sets with uneven distribution.

At the same time, through the analysis of the K-means algorithm and its privacy leakage problem, it can be seen that the key point of privacy leakage is the cluster center point. The clustering center is obtained by dividing the sum of the data points in the cluster by the data points, and detailed data point information is not needed when clustering data sets into a data set. As a result, only by publishing the approximate values of each cluster center point can the data privacy be protected, and meanwhile, the accuracy of the clustering results will not be affected.

In the massive data processing algorithms of this paper and the term, most of them adopt the tree index structure. The biggest disadvantage of this algorithm is the curse of dimensionality. As the amount of data increases, the tree structure expands exponentially, and the memory space becomes larger and larger. Especially for high dimensional data in the image, it is virtually impossible to put it into memory completely. In order to solve this problem, the external storage of hard disk storage is introduced. Although it can solve the problem of storage space, the search efficiency is reduced. There is no doubt that this is a way to exchange space for time, which can greatly reduce the user experience.

5. Conclusion

The purpose of this paper is to test the current entertainment games, movies, office software and learning software. In this paper, the color image clustering algorithm was used, and the color of the image was classified and counted. The influence color interval was constructed and matched, and the interval threshold was demarcated. At the same time, K-means algorithm was introduced to process the image. Through experimental data, as well as empirical selection, K values were selected to

suit behavioral analysis calculations. In addition, the depth influence interval was constructed, and the color depth of the image was classified. Finally, combining the influence of color interval, the boundaries of behavior analysis were delineated, and the innovation of intelligent analysis of computer operation behavior was realized. Research showed that the accuracy rate of the combination of the image abstract semantic features extracted by K-means algorithm and the image low-level visual features for image classification and retrieval was higher than traditional method. The search results were further checked and the results of erroneous retrieval were eliminated, so that higher precision can be achieved. Then, the abstract semantic features were fused with the image color features to compensate for the insensitivity of the CNN algorithm to the color information. However, there are still some problems in this study, for example, how to make the analysis of light color images more effective needs further study.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk: *SLIC superpixels compared to state-of-the-art superpixel methods.* IEEE Transactions on Pattern Analysis and Machine Intelligence *34* (2012), No. 11, 2274–2282.

[2] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu: *An efficient k-means clustering algorithm: Analysis and implementation.* IEEE Transactions on Pattern Analysis and Machine Intelligence *24* (2002), No. 7, 881–892.

[3] L. H. Juang, M. V. Wu: *MRI brain lesion image detection based on color-converted K-means clustering segmentation.* Measurement *43* (2010), No. 7, 941–949.

[4] K. Chen, Y. Zhu: *A summary of machine learning and related algorithms.* Statistics & Information Forum (2007), No. 05.

[5] A. K. Jain: *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters *31* (2010), No. 8, 651–666.

[6] S. L. Yang, Y. S. Li, X. X. Hu, Y. R. Pan: *Optimization study on k value of K-means algorithm.* Systems Engineering-theory & Practice *26* (2006), No. 2, 97–101.

[7] L. Grady: *Random walks for image segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence *28* (2006), No. 11, 1768–1783.

[8] H. D. Zhu, Y. Zhong, X. H. Zhao: *An optimization initial center K-means algorithm for text clustering.* Journal of Zhengzhou University(Natural Science Edition) *41* (2009), No. 02.

[9] K. R. Žalik: *An efficient k'-means clustering algorithm.* Journal Pattern Recognition Letters *29* (2008), No. 9, 1385–1391.

[10] M. Mignotte: *A de-texturing and spatially constrained K-means approach for image segmentation.* Pattern Recognition Letters *32* (2011), No. 2, 359–367.

[11] M. Mignotte: *Segmentation by fusion of histogram-based K-means clusters in different color spaces.* IEEE Transactions on Image Processing *17* (2008), No. 5, 780–787.

[12] J. Fan, M. Han, J. Wang: *Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation.* Pattern Recognition *42* (2009), No. 11, 2527–2540.

[13] A. K. Jain, F. Farrokhnia: *Unsupervised texture segmentation using Gabor filters.* Pattern Recognition *24* (1991), No. 12, 1167–1186.

[14] C. Fowlkes, S. Belongie, F. Chung, J. Malik: *Spectral grouping using the nystrom method.* IEEE Transactions on Pattern Analysis and Machine Intelligence *26* (2004), No. 2, 214–225.

[15] D. L. Pham, C. Xu, J. L. Prince: *Current methods in medical image segmentation.* Annual review of biomedical engineering *2* (2000), 315–337.

Received May 7, 2017