A Cluster Priority Level Decision Method for Image Features

Tianqiang Peng[1] and Haolin Gao[2]

1. Department of Computer Science and Engineering, Henan Institute of Engineering, Zhengzhou 451191
2. Department of Data Engineering, Zhengzhou Information Science and Technology Institute, Zhengzhou, China, 450002
ptq_drumboy@163.com

Abstract

Though clustering Analysis has been developed for many years with many clustering methods come into application, image clustering is still a difficult problem. One of the most fundamental problems is that there are many kinds of image representations, and the distinguish ability of each feature is different, so their cluster effects are also different. To decide cluster priority level of different images features on a specific image dataset, the distinguish ability of three typical image features are analyzed, and a cluster discriminant index is present, which called Simplified Overall Cluster Quality is composed of cluster compaction and cluster separation. Experimental results showed the feature with best distinguish ability also possessed best discriminant index. So this index can be used to decide the priority of features for clustering images or the best feature for image cluster.

Keywords: image clustering, image representation, cluster priority level, Simplified Overall Cluster Quality

1 Introduction

In pattern recognition, learning problem mainly includes supervised learning(classification) and unsupervised learning(clustering), the distinction between them is clear, the first involving only labeled data, while the latter involving only unlabeled data. Clustering is a more difficult and challenging problem than classification. There is a growing interest in a hybrid setting; called semi supervised learning [1]. In semi-supervised classification, the labels of only a small portion of the training data set are available. Clustering (or unsupervised classification) is a method of creating groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct [1]. The application areas of clustering include image segmentation, information retrieval, document classification, associate rule mining, web usage tracking, and transaction analysis and so on.

The goal of clustering is to assign data points with similar properties to the same groups and dissimilar data points to different groups. Generally, clustering problems can be divided into two categories: hard clustering (or crisp clustering) and fuzzy clustering (or soft clustering). In hard clustering, a data point belongs to one and only one cluster, while in fuzzy clustering, a data point may belong to two or more clusters with some probabilities. Though a large number of clustering methods have been developed, clustering remains a challenging task because a clustering algorithm behaves differently depending on the chosen features of the data set and the parameter values of the algorithm [2].

A typical sequence of the clustering activities includes the following three stages [3]:
1. pattern representation, optionally including feature extraction or selection,
2. definition of a pattern proximity measure for the data domain, and
3. clustering or grouping of data points according to the chosen pattern representation

and the proximity measure.

Pattern representation refers to the paradigm for observation and the abstraction of the learning problem. Feature extraction refers to the paradigm for converting the observations of the natural features into a machine understandable format. Feature selection is defined as the task of identifying a set of most representative subset of the natural features to be used by the machine. Pattern representation is considered as the basis of machine learning. Since human accessibility of the patterns is highly dependent on their representation format, an unsuitable pattern representation may result in a failure of producing meaningful clusters. Feature selection and extraction play an important role for abstracting patterns into a machine understandable representation.

During the past decades, many clustering methods including several types were proposed, and all these methods have the same goal, i.e., the maximization of homogeneity within each cluster and the minimization of heterogeneity between different clusters.

Image clustering is one of the important applications of data clustering. In the last few decades, there has been a growing interest in developing effective and fast methods for detecting the different clusters into an input image. Clustering in image processing and computer vision is a procedure for identifying groups of similar image primitives, such as image pixels, local features, segments, objects or even complete images. The general goal in image clustering is to classify the different image objects or patterns in such a way that samples of the same cluster are more similar to one another than samples belonging to different clusters.

Reference [4] classified the mage clustering algorithms into two categories, the supervised and the unsupervised methods. In the supervised image clustering algorithms, the researchers incorporate a priori knowledge, such as the number of image clusters. The main restriction in supervised image clustering is that human intervention is required, that is, image must be labeled before clustering, which is nearly impossible especially in online application. On the other hand, unsupervised methods aim at providing the correct number of image clusters without any priori information.

In image cluster, pattern representation (image feature) is still one of the most important factors that influence the performance of the clustering algorithm. If the representation is good, the clusters are more likely to be compact and isolated, and even a simple clustering algorithm such as K-means will find them. Unfortunately, there is no universally good representation.

Though different image features may result in different clustering effects, many literatures about image clustering directly employed image features into clustering without discussing the clustering effects of different features. This paper try to analysis the cluster priority level of different images features or the clustering effects of different image features. Global features, such as color histogram, Gabor features and geometric moment are discussed. To show the priority level, a discriminant index is proposed to decide which feature is better for clustering.

To find the difference among different image features, the image distance curves of different features are showed, we observed that the change tendency for three curves are obviously different, and the retrieval results for three image feature are also different. To depict the difference, we defined a discriminant index called SOCQ (simplified overall cluster quality) based on OCQ (Overall Clustering Quality). The value of this index can reflect the clustering effect of corresponding image feature. So, we can determine to use which image representation to clustering images according SOCQ. It is different from feature selection, which try to select a subspace from image feature with best distinguish ability[5][6]. As to clustering algorithms, because it is difficult to design a general purpose clustering algorithm and the ill-posed problem of clustering [7], k-means is still widely used. Therefore, in our experiment we use it to cluster images and testify the clustering effects.

The following of this paper was organized as follows, sector 2 analysis the distinguish ability of three global image features, sector 3 present an discriminant index to decide the cluster priority level based on OCQ [8], sector 4 compares the distinguish ability of the three image features, sector 5 concludes the main work of this paper and discuss the cluster priority of image features.

2. The Distinguish Ability of the Image Features

For the convenient of computation and results visualization, we mainly discuss the global features for image clustering. The frequently used global image features include color, texture and shape, respectively represented by HSV color histogram, Gabor texture and geometric moment. The extraction of three features is detailed as follows.

2.1 The Feature Extraction

For different image feature embodies different image information, and each feature is an approximation of image information, so, different image feature may have different similarity for same two images. The three global features include color histogram, Gabor texture and geometric moment.

The Color histogram in HSV space is defined as follows:

$$H = \begin{cases} 0 & if\ h \in [316, 20] \\ 1 & if\ h \in [21, 40] \\ 2 & if\ h \in [16, 75] \\ 3 & if\ h \in [76, 155] \\ 4 & if\ h \in [156, 190] \\ 5 & if\ h \in [191, 270] \\ 6 & if\ h \in [271, 295] \\ 7 & if\ h \in [296, 315] \end{cases}, \quad S = \begin{cases} 0 & if\ s \in [0, 0.2] \\ 1 & if\ s \in [0.2, 0.7], \\ 2 & if\ s \in [0.7, 1] \end{cases} \quad V = \begin{cases} 0 & if\ v \in [0, 0.2] \\ 1 & if\ v \in [0.2, 0.7] \\ 2 & if\ v \in [0.7, 1] \end{cases} \tag{1}$$

$$x = 9H + 3S + V \tag{2}$$

where x is a 72 dimension vector.

Gabor texture:

The geometric or regular moments are projections of the image function onto the monomial $x^p y^q$, where x, y are the image coordinates. The regular moments u_{pq} of order $(p+q)$ of the image function $f(x,y)$ are defined as:

$$u_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{3}$$

where $p, q = 0, 1, 2, \cdots \infty$.

Hu introduced seven nonlinear functions which are translation, scale, and rotation invariant. Hu's seven moment invariants have been widely used in pattern recognition, and their performance has been evaluated under various deformation situations. They are defined as:

$$m_1 = \eta_{20} + \eta_{02}$$
$$m_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$
$$m_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$
$$m_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$
$$m_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2\right] \qquad (4)$$
$$+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2\right]$$
$$m_6 = (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$
$$m_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2\right]$$
$$+ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2\right]$$

where $\eta_{pq} = u_{pq} / (u_{00})^\gamma$, and for $p + q = 2, 3, \ldots$, $\gamma = 1 + (p+q)/2$. The seven moments form a feature vector $f = (m_1, m_2, \cdots, m_7)$, which can be used to represent an image.
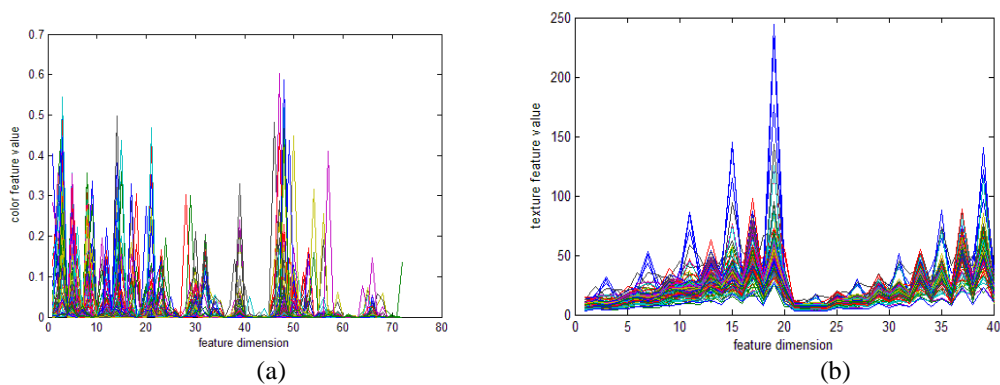
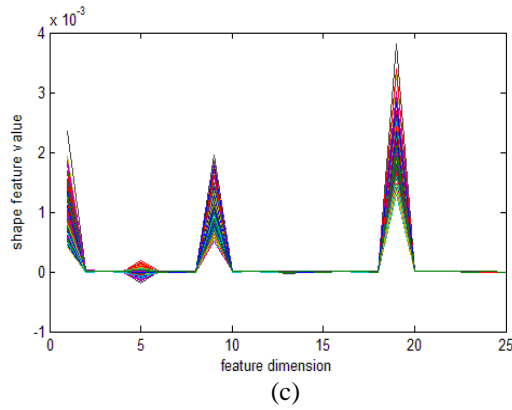## 2.2 The Distinguish Ability of Features

To show the different distinguish ability of three image features, we select 4 categories 100 images from Caltech256 database, each category 25 images. These images we called image set 1. Parts of them are showed in figure 1.

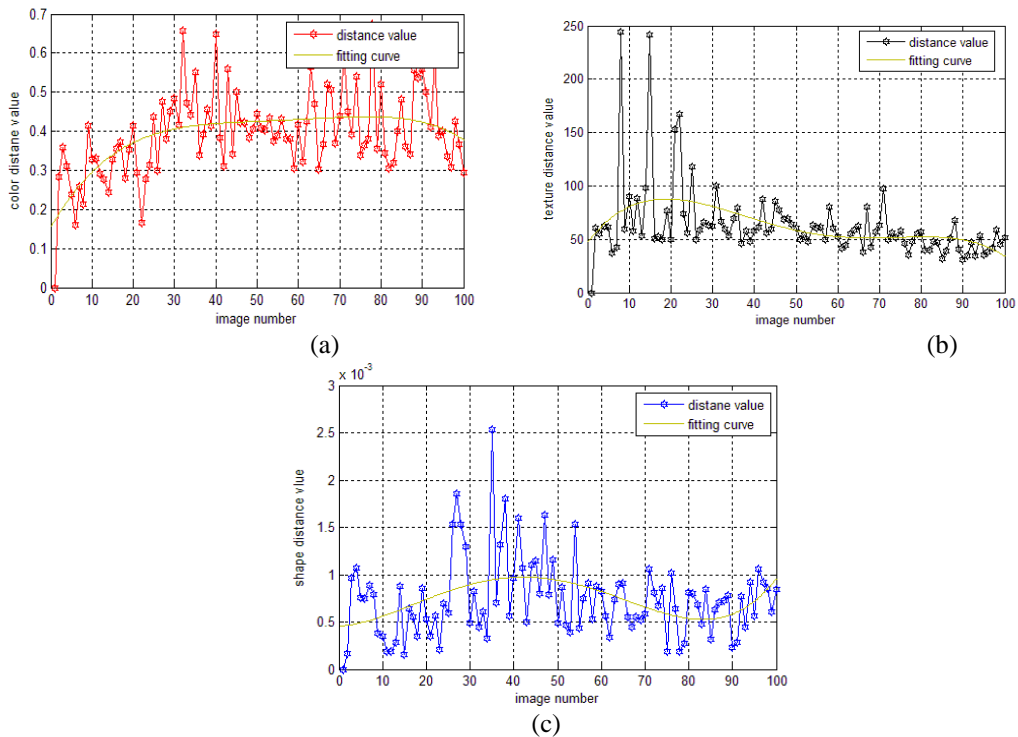**Figure 1. Parts of Images in Image Set 1**

The three global image features are showed in figure 2.

(a)   (b)

(c)

Figure 2. The Global Feature Image Set 1 (a) HSV Color Histogram (b) Gabor Texture (c) Geometric Moment

Because image distance is important for cluster algorithm and many algorithm group images based on inter-distance of images. And to intuitively see the distance of different features, we just calculate the distance between the first image and all the other images, the distance is showed in figure 3. To see the change trendency of curve, we also plot the fitting curve of each distance curve.

(a)

(b)

(c)

Figure 3. The Distance of the First Image and other Images in Image Set 1 for Global Feature (a) HSV Color Histogram (b) Gabor Texture (c) Geometric Moment

We can see that three fitting curves are different, which means that even same pictures have different similarities to the first image, and the relative similarity is also different, for example, the distance of image 99 to image 1 is bigger than image 100 for color feature, but is smaller for texture. It can be intuitively seen in figure 4. The figure (a) is the retrieval result of color feature, with the first image as the query image. Figure b is the result of texture feature. We can see that the retrieval results are very different. In the two figures, only the first 30 images are showed.

(a)

(b)

Figure 4 The retrieval result for different feature (a) HSV color histogram (b) Gabor texture

Since the relative similarity may be different for different feature, so the clustering results also may be different. Then how to decide the cluster priority level? We seek to cluster validity analysis for clustering algorithms.

3 Cluster Validation and the Overall Cluster Quality

3.1 Cluster Validity

Though a large number of clustering methods have been developed, clustering is still a challenging task because a clustering algorithm behaves differently depending on the chosen features of the data set and the parameter values of the algorithm [10]. And from the viewpoint of data distribution, one of the important reasons for clustering is difficult to do well is that, normally there is no priori information about the structure of the data. Hence, the accuracy of clustering entirely depends upon the data and the extent the clustering algorithm capture the structure in the data. To simply the clustering algorithm, sometimes specific structure was imposed to depict a dataset. For example, we can assume that the data is a mixture of Gaussian distributions, which can be solved in a very powerful way with the EM algorithm. But if the assumption about the data distribution is wrong, the results may be very poor.

Therefore, it is important to have some objective measures to evaluate the clustering quality in a quantitative manner; this problem is also called cluster validity analysis. Cluster validity analysis is a well known research community in the area of pattern recognition, which aims to discovering the distribution of a data set, identifying the clustering paradigm that is most suitable for a problem domain, and deciding the optimal parameters for a specific clustering method.

A number of efforts have been made in Cluster validity. However, the issue of cluster validity is rather under-addressed in the area of databases and data mining applications, even though recognized as important. In general terms, there are three approaches to investigate cluster validity. The first is based on external criteria. It implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set. The second approach is based on internal criteria. We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves. The third approach of clustering validity is based on relative criteria. Here the basic idea is the evaluation of a clustering structure by comparing it with other clustering schemes, resulting by the same algorithm but with different parameter values.

The two first approaches are based on statistical tests and their major drawback is their high computational cost. Moreover, the indices related to these approaches aim at

measuring the degree to which a data set confirms a priori specified scheme.

3.2 Renyi Entropy and Overall Cluster Quality (OCQ)

A good clustering result should fit the distribution of dataset at most. As to the data point in the dataset, we expect that the assignment of the data samples to which cluster will not violate the structure inherent in the data. This idea can be captured mathematically using concepts from information theory. Entropy is used to measure the uncertainty, which is a good starting point to measure the effectiveness of clustering. A very closely related measure is the Information Potential derived from Renyi's definition of entrop. The definition of Renyi's entropy is

$$H_\alpha(x) = \frac{1}{1-\alpha} \log \sum_k p_k^\alpha \quad \alpha > 0, \alpha \neq 1 \tag{5}$$

where p_k is the probability mass function for the discrete data. When α=2 we obtain Renyi's quadratic entropy. When the Parzen Window method is used to estimate the required probability function we obtain:

$$\begin{cases} H_2(x) = -\log V(x) \\ V(X) = \frac{1}{N^2} \sum_i \sum_j Gauss(x_i - x_j, 2\sigma^2) \end{cases} \tag{6}$$

where $Gauss()$ is a spherically symmetric Gaussian function, and N is the number of samples. $V(x)$ is an exponentially decaying function proportional to the square distance between samples, and it is always positive.

Based on relative criteria, Ji He et al proposed a new clustering evaluation in virtue of Renyi Entropy [8]. In their approach, there are two criteria proposed for clustering evaluation and selection of an optimal clustering scheme [9]:i) Compactness, the members of each cluster should be as close to each other as possible, ii) Separation, the clusters themselves to be widely spaced. Since the nature of clustering is to reorganize the input samples such that data points in the same cluster are more similar to each other than to points in a different cluster, it is a natural way to evaluate the intracluster homogeneity and the intercluster separation of the clustering output in a global fashion.

The cluster compactness measure is based on the generalized definition of the deviation of a data set given by

$$dev(X) = \sqrt{\frac{1}{n} \sum_{i=1}^{N} d^2(x_i, \bar{x})} \tag{7}$$

where $d(x_i, x_j)$ is a distance metric between two vectors x_i and x_j that reflects their dissimilarity, N is the number of members in X, and $\bar{x} = (1/N)\sum_i x_i$ is the mean of X. A smaller deviation indicates a higher homogeneity of the vectors in the dataset, in terms of the distance measure $d()$. In particular, when X is one-dimensional and $d()$ is the Euclidean distance, $dev(X)$ becomes the standard deviation of the data set . The cluster compactness for the output clusters $C_1, C_2, \cdots, C_k$ generated by a system is then defined as

$$Cmp(X) = \frac{1}{k} \sum_{i=1}^{k} \frac{dev(C_i)}{dev(X)} \tag{8}$$

where k is the number of clusters generated on the data set X, $dev(Ci)$ is the deviation of the cluster , and $dev(X)$ is the deviation of the data set X.

The cluster separation measure used here borrows the idea in[10] and the clustering evaluation function introduced by[11]. The cluster separation of a clustering system's output is defined as

$$Sep(X) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} \exp(-\frac{d^2(c_i, c_j)}{2\sigma^2}) \tag{9}$$

where σ is a Gaussian constant, to simplify the computation $2\sigma^2 = 1$, k is the

number of clusters, c_i is the centroid of the cluster C_i, and $d(c_i, c_j)$ is the distance between the centroid of c_i and the centroid of c_j. Following a similar practice in [10], Ji He combined the cluster compactness and cluster separation measures into one for the ease of evaluating the overall performance of a clustering system. The combination, named overall cluster quality (OCQ), is defined as

$$Ocq(X) = \beta \square Cmp(X) + (1 - \beta) \square Sep(X) \qquad (10)$$

where $\beta \in [0,1]$ is the weight that balances cluster compactness and cluster separation. For example, Ocq (0.5) gives equal weights to the two measures.

3.3 Simplified Overall Cluster Quality (SOCQ)

From the formula 10 and 11, we can see that, the cluster compactness is good when the value of $Cmp(X)$ is small, and so is the cluster separation. Cluster is well separated if the value of $Sep(X)$ is small. And well separated clusters indicate large inter-distance of cluster centers, which result in a small value of $Sep(X)$ due to the monotonous decrease of Gauss function. So, in the whole, the OCQ decreases monotonously with the increase of cluster compactness and cluster separation.

But it is well known that, the exponent function in $Sep(X)$ has a high computation complexity. And, the $Sep(X)$ is design to reflect the effects of inter-cluster distance, simultaneously with a same monotonous property as $Cmp(X)$. So, we can design a new OCQ index, called Simplified OCQ or SOCQ, which discards the exponent function, and just reserve the $d^2(c_i, c_j)$, then the $Sep(X)$ is

$$Sep(X) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} d^2(c_i, c_j) \qquad (11)$$

Then, the $Sep(X)$ will increases monotonously with the increase of inter-cluster distance, i.e. the sum of inter-distance of cluster centers. So, to not destroy the whole monotonous property of SOCQ, which means its two parts have the same monotonous property, we define the $SOCQ(X)$ as follows,

$$Ocq(X) = -\beta \square Cmp(X) + (1 - \beta) \square Sep(X) \qquad (12)$$

where $\beta = 0.5$ yet. Therefore, SOCQ increases monotonously with the increase of cluster compactness and cluster separation, which has an opposite monotonous property. It is noted that, the SOCQ may be a negative value.

4. Experiment

We have noted that the distinguish ability of three features for image set 1 are not well, and the retrieval results are bad either. This illustrates that the distinctness between these image categories is not notable. So, we choose some other images from TRECVID dataset. We choose four categories and 75 images total, called image set 2, parts of images are showed in Figure 5, the five categories includes 'compere', 'compere2', 'singer', 'rice' and 'sports'.

The reason for choose these images is that they are in highly accordance with the goal of clustering, i.e. high intra-class similarity and low inter-class similarity. It is clear that this image set is not an idea dataset for testifying the performance of clustering algorithms. However, this is not our target fortunately. Similarly, we also perform image retrieval with a same query image, and the results are showed respectively in figure 6.
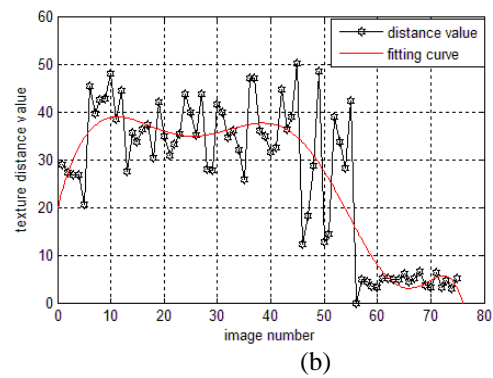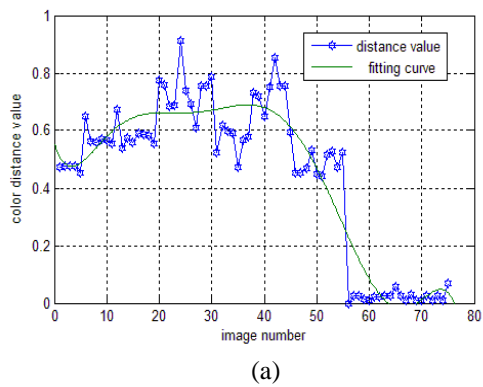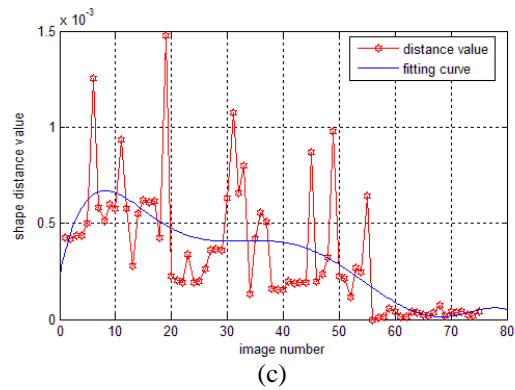
Figure 5. Parts of Images in Image Set 2

(a)

(b)

Figure 6. The Retrieval Result of Texture Feature (a) HSV Color Histogram (b) Gabor Texture


(a)

(b)

(c)

Figure 7. The Distance between the 56th Image and other Images for Global Features (a) HSV Color Histogram (b) Gabor Texture (c) Geometric Moment

We can see explicitly that the color feature do better than texture from the two retrieval results, this illustrates that color feature have a strong distinguish ability than texture feature for the specific query image, i.e. the first image in two figures.

We also plot the distance between the first image and all the other images, and the fitting curve is also showed in the figure 7. We can see that there are obvious step in all three fitting curve, which identifies the inter-class distance is high. Contrastingly, the steps in fitting curves of image set 1 are quite unobvious in figure 3.

From these three distance figures, that the step of color feature or texture feature is higher than shape feature, which means that the inter-class distance of color and texture is bigger than shape. To calculate the OCQ index value of clustering results, we clustering image set 2 with the general clustering algorithm, k-means, and get the values of all three features for 20 times. The results are showed in figure 8. It can be seen that, the values of shape is biggest and color is the smallest, which mean that the quality of color clustering is best, and that of shape is worst.
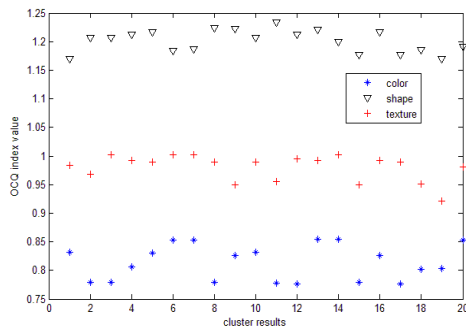
Figure 8. The OCQ Index Value of Image Set 2
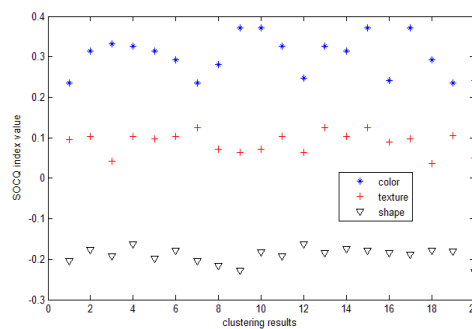
Figure 9 The SOCQ Index Value of Image Set 2

Similarly, we calculate the SOCQ index value of clustering results. We clustering image set 2 also with k-means, and then get the index values of all three features for 20 clustering. The results are showed in figure 9, the relative relationship the three features is opposite to that of figure 8, but they mean same clustering effects. They both indicate that the color clustering achieves the best effect, the shape clustering achieves the worst results.

5. Conclusion

This paper introduces SOCQ to decide the cluster priority level of image features on a specific image dataset. Experimental results showed that the image representation with higher value is better for clustering, because it is more distinguishable than others. The index SOCQ is different from OCQ not only in computation complexity, but also it can work before clustering, which is different from cluster validation methods. So, it is convenient for determining a best image representation to cluster. If several features are used to cluster all, they can be arranged in the order of SOCQ value.

The limitation of SOCQ is that class labels need to be provided, that is to say the image set has to be labeled first. In fact, if only parts of an image set are labeled, SOCQ can also work, and a better representation can be chose on this small subset, then the result can be used to cluster the whole image set. This is because that the inner structure of an image set is stable generally.

The distinguished ability of image features is highly dependant on the inner structure of image dataset and image type. Even one image feature is more distinguishable for an image dataset; it may be less distinguishable for another image dataset.

Acknowledgement

References

[1] O. Chapelle, B. Scholkopf and A. Zien, "Semi supervised learning. MIT Press, Cambridge", MA, **(2006)**.
[2] G. Guojun, M. Chaoqun and W. Jianhong, "Data Clustering: Theory, Algorithms, and Applications", Proceedings of the Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, **(2007)**.
[3] M. Halkidi and M. Vazirgiannis, "A data set oriented approach for clustering algorithm selection", Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, **(2001)**.
[4] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: A review", ACM Computing Surveys, vol. 31, no. 3, **(1999)**, pp. 264-323.
[5] S. Krinidis, M. Krinidis and V. Chatzis, "An Unsupervised Image Clustering Method Based on EEMD Image Histogram", Journal of Information Hiding and Multimedia Signal Processing, vol. 3, no. 2, **(2012)**, pp. 151-163.
[6] Q. Song, J. Ni and G. Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, **(2013)**, pp. 1-14.
[7] T. T. Nguyen, Z. Li, T. Silander and T. Y. Leong, "Online Feature Selection for Model-based Reinforcement Learning", Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, **(2013)**.
[8] A. K. Jain, "Data Clustering: 50 Years Beyond K-Means", Pattern Recognition Letters, vol. 31, no. 8, **(2010)**, pp. 651-666.
[9] J. He, A. H. Tan and C. L. Tan, "Modified ART 2A Growing Network Capable of Generating a Fixed Number of Nodes", IEEE Transactions on Neural Networks, vol. 15, no. 3, **(2004)**, pp. 728-737.
[10] M. J. A. Berry and Gordon Linoff, "Data Mining Techniques For marketing, Sales and Customer Support", John Willey & Sons, Inc., **(1996)**.
[11] M. Halkidi, M. Vazirgiannis and I. Batistakis, "Quality scheme assessment in the clustering process", Proceedings of the 4th Eur. Conf. Principles and Practice of Knowledge Discovery in Databases, **(2000)**, pp. 65–276.
[12] E. Gokcay and J. Principe, "A new clustering evaluation function using Renyi's information potential", Proceedings of the. Int. Conf. Acoust., Speech, Signal Processing, **(2000)**.

Authors

TianQiang Peng, He was born in SuiZhou, China in 1978. He received his master's and PHD degree from Zhengzhou Information Science and Technology Institute in 2003 and 2008 respectively. Now he is an associate professor in Department of Computer Science and Engineering, Henan Institute of Engineering. His current research interests are image retrieval and pattern recognition.

Haolin Gao, He was born in Zhengzhou, China in 1979. He received his bachelor's degree from Navy Aviation Engineering Academic in 2002, and received his master's and PHD degree from Zhengzhou Information Science and Technology Institute in 2008 and 2014 respectively. His current research interests are image retrieval and object recognition.