TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 8, August 2014, pp. 6181 ~ 6189
DOI: 10.11591/telkomnika.v12i8.5458

Which Representation to Choose for Image Cluster

Haolin Gao*, Gang Chen, Bicheng Li, Yongwei Zhao
Department of Data Processing Engineering, Zhengzhou Information Science and Technology Institute,
Zhengzhou, 450002
*Corresponding author, e-mail: holygao@126.com

Abstract

To choose best representation for image cluster, we analyzed the distinguish ability of three typical image representations, color histogram, Gabor texture and geometric moment by fitting image distance curve, and show different retrieval results of image retrieval for different image representations. To depict the distinguish ability of different image representations, we present a cluster discriminant index called Simplified Overall Cluster Quality, which constitutes cluster compaction and cluster separation, and the experiment results show that the image representation with best distinguish ability also possesses maximum discriminant index value. This index can be used to determine optimal representation for clustering images.

Keywords: *image clustering, image representation, simplified overall cluster quality*

1. Introduction

Clustering, also called cluster analysis or unsupervised classification is a method of creating groups or clusters of objects, so that objects in one cluster are very similar and objects in different clusters are quite distinctable [1]. The application areas of clustering include image segmentation, information retrieval, document classification, associate rule mining and web usage tracking and so on. Generally, clustering problems can be divided into two categories: hard clustering and soft clustering. In hard clustering, a data point belongs to one cluster, while in soft clustering; a data point may belong to two or more clusters with some probabilities. Clustering algorithms can also be classified into two categories: hierarchical algorithms and partitioned algorithms. Hierarchical algorithms include two types: divisive and agglomerative.

Though a large number of clustering methods have been developed, clustering remains a challenging task. A clustering algorithm may behave differently because the different chosen of features of the data set or the parameter values of the algorithm [2]. A typical clustering method includes the following three stages [3]: pattern representation, definition of a pattern proximity measure, grouping data points according to the pattern representation and the proximity measure. During the past decades, many clustering methods 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.

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. Image clustering is one of the important applications of data clustering. Reference [4] classified the image clustering algorithms into two categories, the supervised and the unsupervised methods. The supervised algorithms incorporate a priori knowledge, such as the number of image clusters.

Data representation in these clustering algorithms is still one of the most important factors that influence the performance of the clustering algorithm. If the representation is good, the clusters are 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; the choice of representation must be guided by the domain knowledge.

So, different image features may result in different clustering effects. However, many currently literatures about image clustering directly employed image features to clustering without discussing the clustering effects of different features. This paper try to analysis the different clustering effects of different image features, mainly global features, such as color

Received December 23, 2013; Revised April 18, 2014; Accepted April 30, 2014

histogram, Gabor features and geometric moment, and propose a discriminant index to judge which feature is better for clustering.

To find the variation among different image features, we compared the image distance curves of different features; the curves showed that the waveform 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). 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. As to clustering algorithms, one of the most popular and simple algorithms was K-means. Though it was proposed over 50 years ago and thousands of clustering algorithms have been published since then, K-means is still widely used such as in image cluster. This speaks to the difficulty of designing a general purpose clustering algorithm and the ill-posed problem of clustering[5]. Therefore, in our experiment we use it to cluster images and testify the clustering effects.

Our work focused on how to choose a beat feature for image cluster, which is different from feature selection [6],[7], they selected a subspace from high dimensional space to improve the efficiency of classification. We don't select a subspace vector from a high dimension vector, we just evaluated the clustering effect of image feature vector.

The following of this paper was organized as follows, sector 2 analysis the distinguish ability of three global image features, sector 3 present discriminant index for clustering effects 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 image features and the image clustering.

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 the 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.

HSV Color histogram needs to translate an image into HSV space and quantified the results. The quantification is showed as follows:

$$H=\begin{cases}0 \ \ if \ h\in[316,20]\\ 1 \ \ if \ h\in[21,40]\\ 2 \ \ if \ h\in[416,75]\\ 3 \ \ if \ h\in[76,155]\\ 4 \ \ if \ h\in[156,190]\\ 5 \ \ if \ h\in[191,270]\\ 6 \ \ if \ h\in[271,295]\\ 7 \ \ if \ h\in[296,315]\end{cases} \quad S=\begin{cases}0 \ \ if \ s\in[0,0.2]\\ 1 \ \ if \ s\in[0.2,0.7]\\ 2 \ \ if \ s\in[0.7,1]\end{cases} \quad V=\begin{cases}0 \ \ if \ v\in[0,0.2]\\ 1 \ \ if \ v\in[0.2,0.7]\\ 2 \ \ if \ v\in[0.7,1]\end{cases} \tag{1}$$

Then the quantification results are combined in a single value:

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

Where $x\in[0,71]$, so an image can be represented by a histogram of 72 bins.

Gabor texture is based Gabor function and wavelet. Given an image *I*(*x*, *y*), its Gabor wavelet transform is then defined to be:

$$W_{mn}(x,y)=\int I(x_1,y_1)g_{mn}^{*}(x-x_1,y-y_1)dx_1y_1 \tag{3}$$

Where * indicates the complex conjugate. It is assumed that the local texture regions are spatially homogeneous, and the mean μ_{mn} and the standard deviation σ_{mn} of the magnitude of the transform coefficients are used to represent the region:

$$\mu_{mn} = \iint \left|W_{mn}(x,y)\right| dxdy,\ \ \sigma_{mn} = \iint \left(\left|W_{mn}(x,y)\right| - \mu_{mn}\right)^2 dxdy \tag{4}$$

A feature vector is now constructed using μ_{mn} and σ_{mn}, as feature components. We use scales *m* = 5 and orientations *n* = 4, resulting in a feature vector.

$$\overline{f} = \left[\mu_{00}\sigma_{00} \cdots \mu_{34}\sigma_{34}\right] \tag{5}$$

Which is a 40 dimension vector.

The shape feature is depicted by geometric moments, which are projections of the image function *f*(*x*,*y*) onto the monomial $x^p y^q$. The geometric moments u_{pq} of order (*p*+*q*) of the image function *f*(*x*,*y*) are defined as:

$$u_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x,y) \tag{6}$$

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:

$$\begin{aligned}
m_1 &= \eta_{20} + \eta_{02} \\
m_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
m_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
m_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
m_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2\right] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2\right] \\
m_6 &= (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
m_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2\right] \\
&\quad + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2\right]
\end{aligned} \tag{7}$$

Where $\eta_{pq} = u_{pq} / \left(u_{00}\right)^{\gamma}$, 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. We use the 25 dimension vector in this paper.

$$\overline{f} = (\eta_{02}\eta_{03}\eta_{04} \ldots m_1 m_2 \ldots m_7) \tag{8}$$

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, called image set 1. Parts of them are showed in Figure 1.

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 2.

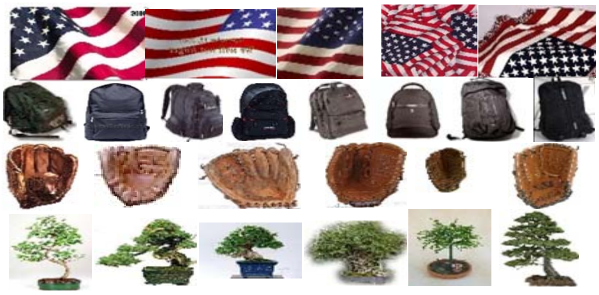
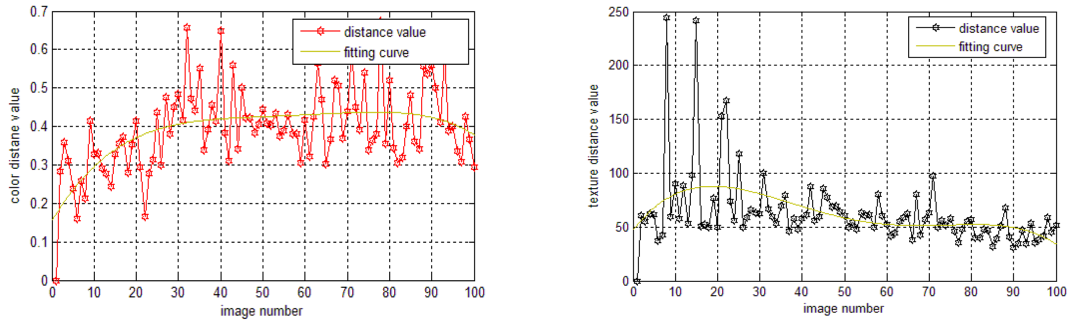


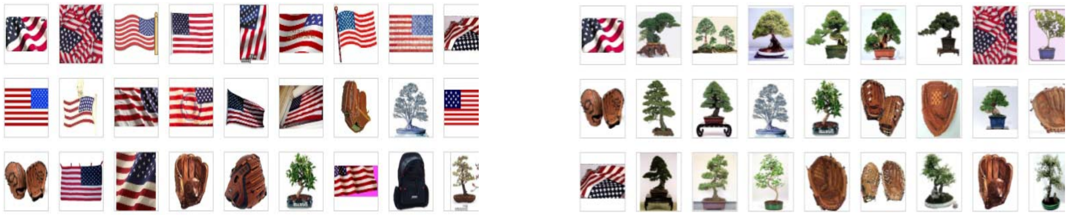
Figure 1. Parts of Images in Image Set 1



(a) (b)

Figure 2. The Distance of the First Image and other Images in Image Set 1 (a) HSV color histogram, (b) Gabor texture

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 the distance of image 100 to image 1 for color feature, but is smaller for texture. This can be intuitively seen in Figure 3. The Figure 3(a) is the retrieval result of color feature, with the first image as the query image. Figure 3(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 3. 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 which feature is better for clustering? We seek to cluster validity analysis for clustering algorithms.

3. Cluster Validation and the Overall Cluster Quality

3.1. Cluster Validity

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, normally there is no priori information about the structure of the data. Hence, the accuracy of clustering entirely depends on the data and the extent the clustering algorithm capture the structure. To simply the clustering algorithm, sometimes specific structure was imposed to a dataset. But if the assumption about the data distribution is wrong, the results may be very poor. Therefore, it is important to evaluate the clustering quality in quantitative manner, and this problem is also called cluster validity analysis, 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 general terms, there are three approaches to investigate cluster validity. The first is based on external criteria. It evaluates 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. It evaluates 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. It evaluates a clustering structure by comparing it with other clustering schemes.

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. That is to say 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, and entropy is used to measure the uncertainty. A well known entropy is Renyi's entrop. The definition is:

$$H_{\alpha}(x) = \frac{1}{1-\alpha}\log\sum_{k} p_k^{\alpha} \qquad \alpha > 0, \alpha \neq 1 \tag{9}$$

Where p_k is the probability mass function for the discrete data. When α=2 we obtain Renyi's quadratic entropy.

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, compactness and separation [9].

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

$$dev(\boldsymbol{X}) = \sqrt{\frac{1}{n}\sum_{i=1}^{N} d^2(x_i, \overline{x})} \tag{10}$$

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 $\overline{x} = (1/N)\sum_i x_i$ is the mean of ***X***. A smaller deviation indicates a higher homogeneity of the vectors in the dataset. In 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 cluster results C_1, C_2,…,C_k is defined as:

$$Cmp(\boldsymbol{X}) = \frac{1}{k}\sum_{i=1}^{k}\frac{dev(C_i)}{dev(\boldsymbol{X})} \tag{11}$$

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 is defined as:

$$Sep(\boldsymbol{X}) = \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}\exp(-\frac{d^2(c_i,c_j)}{2\sigma^2}) \qquad (12)$$

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 c_i and c_j. Similar to [10], Ji He combined the cluster compactness and cluster separation measures into one for the ease of evaluating the overall performance. The combination, named overall cluster quality(OCQ), is defined as:

$$Ocq(\boldsymbol{X}) = \beta\Box Cmp(\boldsymbol{X}) + (1-\beta)\Box Sep(\boldsymbol{X}) \qquad (13)$$

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)

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:

$$Sep(\boldsymbol{X}) = \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k} d^2(c_i,c_j) \qquad (14)$$

Then, the *Sep*(***X***) will increases monotonously with the increase of the square of inter-cluster distance. 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:

$$Ocq(\boldsymbol{X}) = -\beta\Box Cmp(\boldsymbol{X}) + (1-\beta)\Box Sep(\boldsymbol{X}) \qquad (15)$$

Where $\beta = 0.5$. Therefore, SOCQ increases monotonously with the increase of cluster compactness and cluster separation. It is noted that SOCQ may be a negative value.

4. Experiment

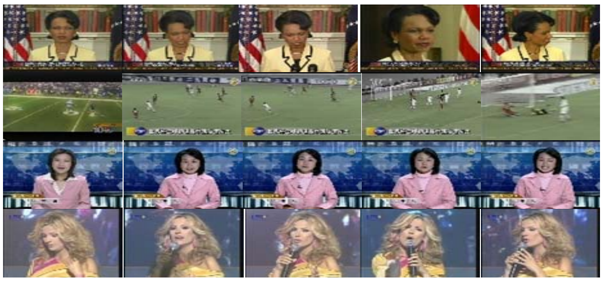
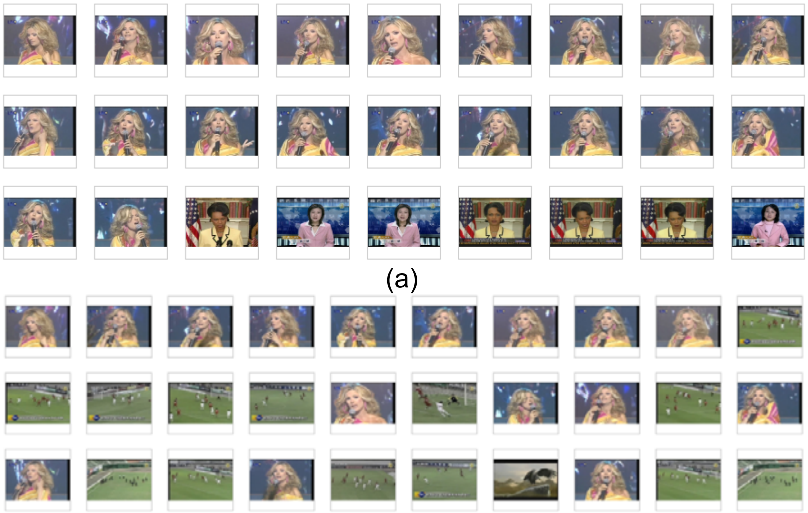


Figure 4. Parts of Images in Image Set 2

We choose four categories and 75 images total, called image set 2 for experiments, parts of images are showed in Figure 4, the 4 categories includes ‘compere’, ‘singer’, ‘rice’ and ’sports’

The reason for choosing these images is that they have high intra-class similarity and low inter-class similarity. It is very suitable to cluster and test the distinct ability of image feature. We perform image retrieval with a same query image, and the results are showed respectively in Figure 5.



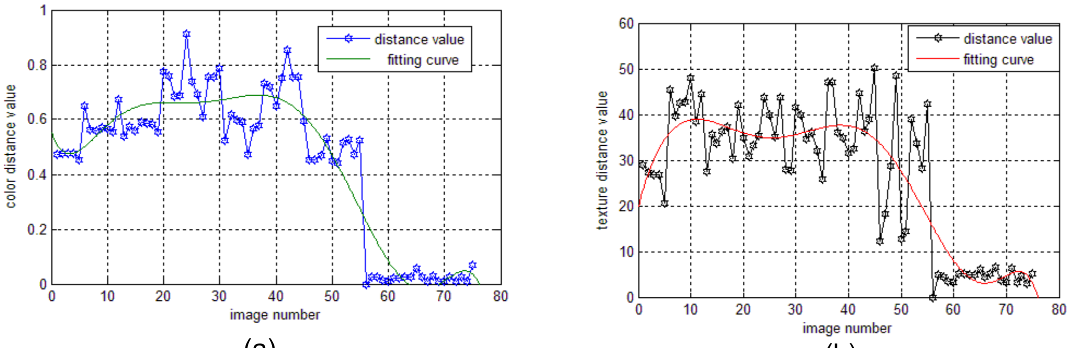
(a)

(b)

Figure 5. The Retrieval Result of Image Set 2; (a) HSV color histogram, (b) Gabor texture

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 6. 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 2.



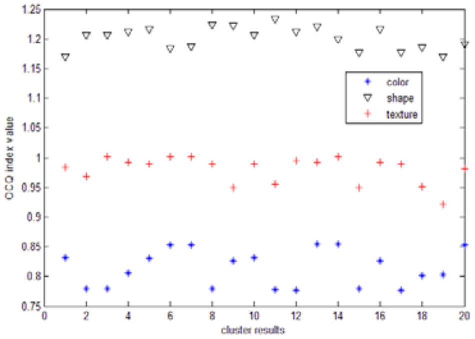
(a)

(b)

Figure 6. The Distance between the 56th Image and other Images; (a) HSV color histogram, (b) Gabor texture

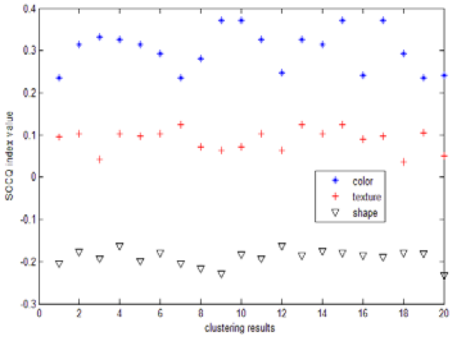
From these three distance figures, we can say that the step of color feature or texture feature is higher than shape feature, which means that the inter-class distance of color and texture is bigger than shape. To calculate the OCQ index value of clustering results, we clustering image set 2 with the general clustering algorithm, k-means, and get the values of all three features for 20 times. The results are showed in Figure 7. 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.

Similarly, we calculate the SOCQ index value of clustering results, we clustering image set 2 also with k-means, and get the index values of all three features for 20 clustering. The results are showed in Figure 8. It can be seen that, the relative relationship in numbers of the three features is opposition to that of Figure 7, but they mean same clustering effects, the color clustering achieves the best effect, the shape clustering the worst.



(a)

Figure 7. The OCQ Index Value of Image Set 2



(b)

Figure 8. The SOCQ Index Value of Image Set 2

5. Conclusion

This paper introduces SOCQ to tackle the problem of choosing which image representation to cluster. The image representation with higher value is better for clustering, because it is more distinguishable than others. 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. Though sometimes several features are used to cluster together, we can weight them 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. If only parts of an image set are labeled, SOCQ can also work, it can choose a better representation on this small subset, then use the result to cluster the whole image set. This is because the inner structure of an image set is stable.

Acknowledgements

This work was supported by Nature Science Foundation of China No. 60872142.

References

[1] Guojun G, Chaoqun M, Jianhong W, Data Clustering: Theory, Algorithms, and Applications. Society for Industrial and Applied Mathematics. Philadelphia, Pennsylvania. 2007.
[2] Halkidi M, Vazirgiannis M. *A data set oriented approach for clustering algorithm selection*. Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery. 2001.
[3] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. *ACM Computing Surveys*. 1999; 31(3): 264-323.
[4] Krinidis S, Krinidis M, Chatzis V. An Unsupervised Image Clustering Method Based on EEMD Image Histogram. *Journal of Information Hiding and Multimedia Signal Processing*. 2012: 3(2): 151-163.
[5] Jain AK. Data Clustering: 50 Years beyond K-Means. *Pattern Recognition*. 2009.

[6] Zhu SX, Hu B. Hybrid Feature Selection Based on Improved Genetic Algorithm. *Indonesian Journal of Electrical Engineering*. 2013; 11(4): 1725-1730.

[7] Li JZ, Meng XR, Wen J. An Improved Method of SVM-BPSO Feature Selection Based on Cloud Model. *Indonesian Journal of Electrical Engineering.* 2014; 12(5): 3979-3986.

[8] He J, Tan AH, Chew-Lim Tan. Modified ART 2A Growing Network Capable of Generating a Fixed Number of Nodes. *IEEE Transactions on Neural Networks*. 2004; (15)3: 728-737.

[9] Michael JA Berry, Linoff G. Data Mining Techniques For marketing, Sales and Customer Support. John Willey & Sons, Inc. 1996.

[10] Halkidi M, Vazirgiannis M, Batistakis I. Quality scheme assessment in the clustering process. *Proc. 4th Eur. Conf. Principles and Practice of Knowledge Discovery in Databases*. 2000: 65–276.

[11] Gokcay E, Principe J. A new clustering evaluation function using Renyi's information potentain. *Proc. Int. Conf. Acoust., Speech, Signal Processing*. 2000.