ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT ’09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 274-277

Application Research of k-means Clustering Algorithm in Image Retrieval System

Hong Liu [1], and Xiaohong Yu [2]
[1] College of Computer science and Information Engineering Zhejiang Gongshang University, HangZhou, China
Email: LLH @mail.hzic.edu.cn
[2] College of Computer science and Information Engineering Zhejiang Gongshang University, HangZhou, China
Email: XHYU @mail.zjgsu.edu.cn

***Abstract*—In image retrieval algorithms, retrieval is according to feature similarities with respect to the query, ignoring the similarities among images in database. To use the feature similarities information, this paper presents an application of k-means clustering algorithm to image retrieval system. Combining the low-level visual features and high-level concepts, the proposed approach fully explores the similarities among images in database, using such clustering algorithm and optimizes the relevance results from traditional image retrieval system by firstly clustering the similar images in the images database to improve the efficiency of images retrieval system. The results of experiments on the testing images show that the proposed approach can greatly improve the efficiency and performances of image retrieval, as well as the convergence to user’s retrieval concept.**

***Index Terms*—image retrieval, k-means cluster algorithm, feature extraction**

I. Introduction

With the popularity of internet and rapid development of digital technique, content-based image retrieval (CBIR) has become an important part of information retrieval technology. CBIR technique focus on searching images in database similar to the query image, according to the image features related to content. This technique is based on automatic extraction of image features, retrieves by automatically comparing the features of query image (such as color, shape, texture, etc.). With the corresponding features in image feature library, and finally outputs the best matching images and its corresponding information.

In CBIR, feature vectors extracted from images usually exist in a very high-dimensional space, such high dimensionality of the feature vectors leads to high computational complexity in calculation for similarity retrieval, and inefficiency in indexing and search, and where a parametric characterization of the distribution is often impossible. Due to the high dimensionality, researchers use the similarity measure to measure the degree of similarity between images. But, there still exists a semantic gap, which just reflects the discrepancy between the relatively limited descriptive power of low-level visual features and high-level concepts. The system is based on the similarities between the query image and images in database while ignoring the similarities among images in database. In order to solve this problem, graph theoretic approaches have been used to effectively explore the similarities among images in database, and the problem could be transformed into solving a maximal clique problem in graph theory which is a clustering problem in computer vision area.

Nowadays, there are also some researches about it. For example, “Texts” in natural languages are the main means to convey semantics among human being. Therefore, the status quo of semantic image clustering is to incorporate ‘‘texts”, e.g., captions to facilitate the understanding of images. Gong et al. [5] proposed to integrate the captions of images for semantic clustering; Dai and Cai [6] built a semantic tolerance model which first represents images based on semantic classification. Hai [7] proposes to understand the images through the analysis of semantic links existing among web pages. In this area where the ever-increasing number of images acquired through the digital world, it makes the brute force searching almost impossible. A user’s query interest is often focused on one particular part of the image, i.e., a region in the image that has an obvious semantic meaning. Therefore, rather than viewing each image as a whole, it is more reasonable to view it as a set of semantic regions Of course, for such a problem, some people do some research. In [8, 10], it is proposed that semantic clustering is performed using relevance feedback. These works are based on the whole image instead of image sub-regions/regions. The clustering method in [10] is based on a method called CAST while the one in [8] is based on the Association Rule Hyper-graph Partitioning algorithm, etc.

This paper puts forward a new framework of content-based image retrieval, which integrates semantic cluster classifier with k-means algorithm. And to improve the efficiency, we propose to impose a clustering component in the region-based image retrieval, which makes it possible to only search the clusters that are close to the query target, instead of searching the whole search space. The rest of the paper is organized as follows. The algorithms for k-means clustering are introduced in Section 2. The new image retrieval algorithm framework is described in Section 3. Experiments and results are presented in Section 4. Finally, conclusions and future works are given in Section 5.

II. K-means Clustering

Clustering algorithm has been widely used in computer vision such as image segmentation and database organization. The purpose of clustering is to group images whose feature vectors are similar by similarity judgment standard; meanwhile to separate the dissimilar images. Clustering algorithms can be broadly divided into two groups: hierarchical and partitional. Hierarchical clustering algorithms recursively find nested clusters either in agglomerativemode (starting with each data point in its own cluster and merging the most similar pair of clusters successively to form a cluster hierarchy) or in divisive (top-down) mode (starting with all the data points in one cluster and recursively dividing each cluster into smaller clusters). Compared to hierarchical clustering algorithms, partitional clustering algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure. Input to a hierarchical algorithm is an n*n similarity matrix, where n is the number of objects to be clustered. On the other hand, a partitional algorithm can use either an n*d pattern matrix, where n objects are embedded in a d-dimensional feature space, or an n*n similarity matrix. Note that a similarity matrix can be easily derived from a pattern matrix, but ordination methods such as multi-dimensional scaling (MDS) are needed to derive a pattern matrix from a similarity matrix.

The most well-known hierarchical algorithms are single-link and complete-link; the most popular and the simplest partitional algorithm is K-means. Since partitional algorithms are preferred in pattern recognition due to the nature of available data, K-means has a rich and diverse history as it was independently discovered in different scientific fields, it is one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity. In the paper, we apply k-means algorithm to analysis images similarities in the database.

Let X={x_i},i=1,…..n be the set of n d-dimensional points to be clustered into a set of k clusters, C={c_k,k=1,….k}; k-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let u_k be the mean of cluster c_k. The squared error between u_k and the points in cluster c_k is defined as

$$J(c_k) = \sum_{x_i \in c_k} \left\| x_i - u_k \right\|^2 . \qquad (1)$$

The goal of K-means is to minimize the sum of the squared error over all K clusters,

$$J(c_k) = \sum_{k=1}^{k} \sum_{x_i \in c_k} \left\| x_i - u_k \right\|^2 . \qquad (2)$$

Minimizing this objective function is known to be an NP-hard problem (even for k =2). Thus K-means, which is a greedy algorithm, can only converge to a local minimum, even though recent study has shown with a large probability k-means could converge to the global optimum when clusters are well separated (Meila, 2006). k-means starts with an initial partition with k clusters and assign patterns to clusters so as to reduce the squared error. Since the squared error always decrease with an increase in the number of clusters k (with J(C)= 0 when k= n), it can be minimized only for a fixed number of clusters. The main steps of k-means algorithm are as follows:

(1) Select an initial partition with K=k clusters; repeat steps (2) and (3) until cluster membership stabilizes.

(2) Generate a new partition by assigning each pattern to its closest cluster center.

(3) Compute new cluster centers.

III. FRAMEWORK OF CONTENT-BASED IMAGE RETRIEVAL COMBINED WITH K-MEANS CLUSTERING ALGORITHM

Since image retrieval is according to the similarities between the query image and images in image database, ignoring the similarities between images in image database. The paper applies the clustering algorithm to further explore the similarities between images in image database for reducing the image retrieval space.

In CBIR, Image feature vectors can be represented by a real matrix G, each row gi of which represents the feature vector of an image in database, and each column represents one kind of feature value. Element G(i,k) represents the feature value of the ith image under the kth feature. The relationship among images can be represented by affinity A ¼ (a_{ij}), where aij represents the similarity between the ith image and the jth image. The similarity could be measured by Euclidian distance or other metrics. The image retrieval system combing clustering algorithm is shown as Fig. 1. The system could be any real value symmetric image retrieval system. First, extract the image features of each image in image database and apply the clustering algorithm to analysis the similarities of images in the database for constructing the images clustering database, then, input the query image, extracting its features and comparing the similarities between features of it and those of images in image clustering database, and output the best matching results.

IV. EXPERIMENT RESULT

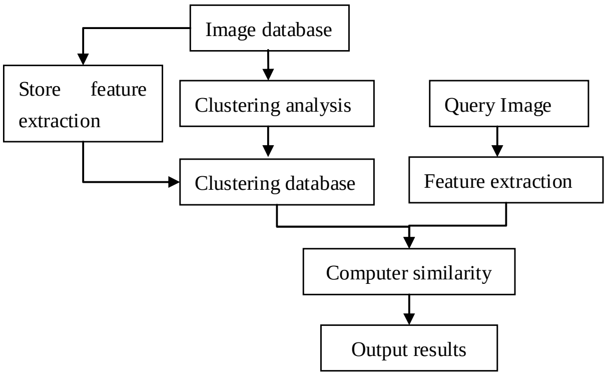

Figure1. Framework of image retrieval system combined with clustering algorithm.

We performed experiments mainly on three images testing set, such as flowers, flags and winter, using color feature only and color and shape feature for retrieval, combining the k-means clustering algorithm with images retrieval.

In order to show k-means clustering algorithm performance in image retrieval system, we design a series of test on the clustering performance. Fig. 2 and Fig.3

show the retrieval results with and without k-means clustering algorithm. The upper left image is the query image. The right part is the retrieved images. Fig. 2 displays the retrieval results without k-means clustering algorithm, while Fig.3 displays the results with k-means clustering algorithm that images similar to each other and to query. In Fig. 2, the query image is a flag and feature vector composed of color and texture. We observe that application k-means clustering algorithm in the images retrieval can throw away some images that are visually irrelevant to the query image for reducing images retrieval space. A leaf image is displayed in the firstly retrieval results in Fig. 2, because of low-level color features. While through doing k-means clustering analysis for image retrieval, such leaf image isn’t displayed in the results in Fig.3.

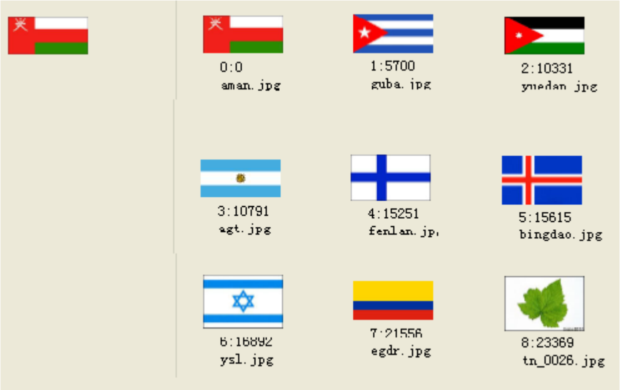

Figure 2. Retrieval results for example query flags image without k-means clustering algorithm

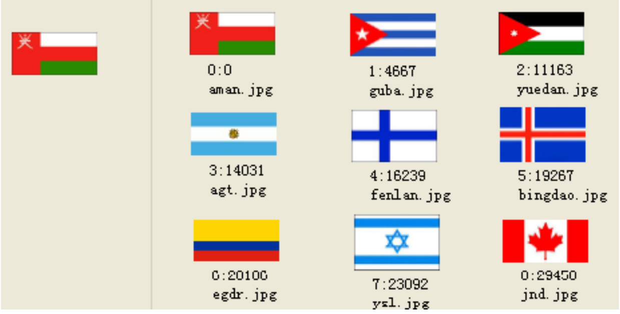

Figure 3. Retrieval results for example query flags image using k-means clustering algorithm before image retrieval.

More query examples are given in Fig. 4 and Fig.5. Fig.4 shows flowers retrieval results without k-means clustering algorithm, and Fig. 5 shows retrieval results using such k-means clustering algorithm to reducing image retrieval space.

V. Conclusion

Image retrieval algorithms always use the similarity between the query image and images in image database. However, they ignore the similarities between images in image database. In this paper we addressed this problem by introducing a graph-theoretic approach for image retrieval post-processing step by finding image similarity clustering to reduce the images retrieving space.

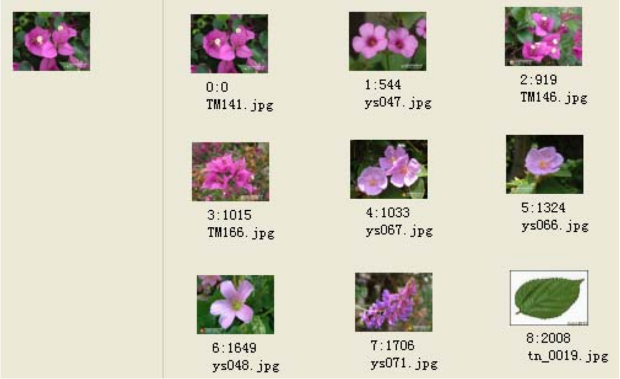

Figure 4. Retrieval results for example query flags image without k-means clustering algorithm

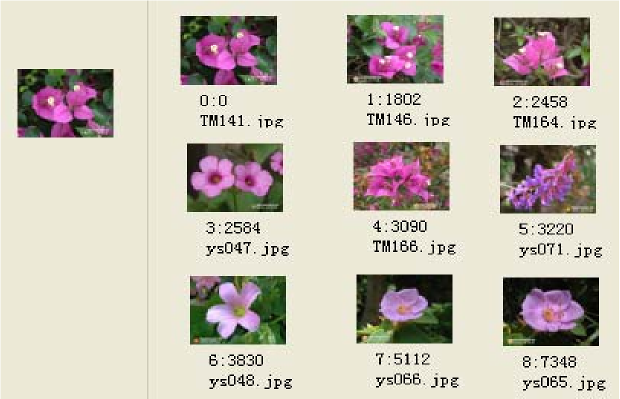

Figure 5. Retrieval results for example query flowers image using k-means clustering algorithm before image retrieval.

Experiment results show that the efficiency and effectiveness of k-means algorithm in analyzing image clustering, which also can improve the efficiency of image retrieving and evidently promote retrieval precision. This k-means algorithm independent on the feature extraction algorithm is used as a post-processing step in retrieval. The improvements in selecting neighborhood vertices of the retrieval results from tradition image retrieval system in image feature space could also improve the recall rate. Since image features is another problem in image retrieval, finding suitable features is important. Thus, feature selection is a problem for our future work. For k-means clustering algorithm, machine learning and pattern recognition communities need to address a number of issues to improve our understanding of data clustering. Below is about research directions that are worth focusing about such algorithm applications in the image retrieval.

(1) Regardless of the principle (or objective), most clustering methods are eventually cast into combinatorial optimization problems that aim to find the partitioning of data that optimizes the objective. As a result, computational issue becomes critical when the application involves large scale data. For instance, finding the global optimal solution for K-means is NP-hard. Hence, it is important to choose clustering principles that lead to computationally efficient solutions.

(2) A fundamental issue related to clustering is its stability or consistency. A good clustering principle should result in a data partitioning that is stable with respect to perturbations in the data. We need to develop clustering methods that lead to stable solutions.

(3) Choosing clustering principles according to their satisfiability of the stated axioms. Despite Kleinberg's impossibility theorem, several studies have shown that it can be overcome by relaxing some of the axioms. Thus, maybe one way to evaluate a clustering principle is to determine to what degree it satisfies the axioms.

(4) Given the inherent difficulty of clustering, it makes more sense to develop semi-supervised clustering techniques in which the labeled data and (user specified) pair-wise constraints can be used to decide both (i) data representation and (ii) appropriate objective function for data clustering.

ACKNOWLEDGMENT

The work of this paper was supported by the Zhejiang provincial natural science foundation as general science and technology research project (No. Y1080565).

REFERENCES

[1] Shyi-Chyi Cheng , Tian-Luu Wu. Fast indexing method for image retrieval using k nearest neighbors searches by principal axis analysis. S.-C. Cheng, T.-L. Wu / J. Vis. Commun. Image R. 17 (2006) 42–56.

[2] Muhammad Atif Tahir,, Ahmed Bouridane, Fatih Kurugollu. Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. M.A. Tahir et al./Pattern Recognition Letters 28 (2007) 438–446.

[3] Ying Liua, Dengsheng Zhang, Guojun Lu,Wei-Ying Ma. Asurvey of content-based image retrievalwith high-level semantics. Y. Liu et al. / Pattern Recognition 40 (2007) 262 – 282.

[4] Hsin-Chang Yang,Chung-Hong Lee. Image semantics discovery from web pages for semantic-based image retrieval using self-organizing maps. H.-C. Yang, C.-H. Lee / Expert Systems with Applications 34 (2008) 266–279.

[5] Z. Gong, L. Hou U, C.W. Cheang, Web image semantic clustering, in:Proceedings of ODBASE, 2005, pp. 1416–1431.

[6] Y. Dai, D. Cai, Image clustering using semantic tolerance relation model, in:Proceedings on European Internet and Multimedia Systems and Applications, 2007

[7] Z. Hai, Retrieve images by understanding semantic links and clustering image fragments, Journal of Systems and Software 73(2004) 455–466.

[8] L. Duan, Y. Chen, W. Gao, Learning semantic cluster for image retrieval using association rule hyper graph partitioning, in: Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing and the Fourth Pacific RimConference on Multimedia, 2003, pp. 1581–1585

[9] C. Zhang, X. Chen, M. Chen, S.-C. Chen, M.-L. Shyu, A multiple instance learning approach for content based image retrieval using one-class support vector machine, in: Proceedings of the IEEE International conference on Multimedia & Expo (ICME), Amsterdam, The Netherlands, 2005, pp. 1142–1145.

[10] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, W.-Y. Ma, IGroup: a web image search engine with semantic clustering of search results, in: Proceedings of ACM MM Demo, 2006.

[11] Ying Liu, Xin Chen, Chengcui Zhang, Alan Sprague. Semantic clustering for region-based image retrieval Journal of Visual Communication and Image Representation, Volume 20, Issue 2, February 2009, Pages 157-166.

[12] Shyi-Chyi Cheng, Tzu-Chuan Chou, Chao-Lung Yang, Hung-Yi Chang.A semantic learning for content-based image retrieval using analytical hierarchy process. Expert Systems with Applications, Volume 28, Issue 3, April 2005, Pages 495-505.

[13] Hsin-Chang Yang, Chung-Hong Lee. Image semantics discovery from web pages for semantic-based image retrieval using self-organizing maps. Expert Systems with Applications, Volume 34, Issue 1, January 2008, Pages 266-279.

[14] Hai Jin, Xiaomin Ning, Weijia Jia, Hao Wu, Guilin Lu.Combining weights with fuzziness for intelligent semantic web search. Knowledge-Based Systems, Volume 21, Issue 7, October 2008, Pages 655-665.

[15] Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma.A survey of content-based image retrieval with high-level semantics. Pattern Recognition, Volume 40, Issue 1, January 2007, Pages 262-282.

[16] Muhammad Atif Tahir, Ahmed Bouridane, Fatih Kurugollu. Imultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. Pattern Recognition Letters, Volume 28, Issue 4, 1 March 2007, Pages 438-446.

[17] Sarbast Rasheed, Daniel Stashuk, Mohamed Kamel.Adaptive fuzzy k-NN classifier for EMG signals decomposition. Medical Engineering & Physics, Volume 28, Issue 7, September 2006, Pages 694-709.

[18] J. Amores, N. Sebe, P. Radeva.Boosting the distance estimation: Application to the K-Nearest Neighbor Classifier. Pattern Recognition Letters, Volume 27, Issue 3, February 2006, Pages 201-209.

[19] Man Wang, Zheng-Lin Ye, Yue Wang, Shu-Xun Wang. Dominant sets clustering for image retrieval. M. Wang et al. / Signal Processing 88 (2008) 2843–2849.