ACQ: An Automatic Clustering and Querying Approach for Large Image Databases

Dantong Yu and Aidong Zhang
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260

Abstract

Large image collections such as web-based image databases are being built in various locations. Because of the diversity of such image data collections, clustering images becomes an important and non-trivial problem. Such clustering tries to find the densely populated regions in the feature space to be used for efficient image retrieval. In this paper, we present an automatic clustering and querying (ACQ) approach for large image databases. Our approach can efficiently detect clusters of arbitrary shape. It does not require the number of clusters to be known a priori and is insensitive to the noise (outliers) and the order of input data. Based on this clustering approach, efficient image querying is supported. Experiments demonstrate the effectiveness and efficiency of the approach.

1 Clustering in High Dimensional Feature Space

Images are represented by high-dimensional feature vectors. Clustering and content-based image retrieval are actually applied on the feature vectors. In the high dimensional feature space, the actual dense areas are very small compared to the whole space. A hash table can be an effective method to exclusively deal with this kind of situation. In [1], we present a novel clustering algorithm, termed ACQ, and provide an indepth analysis of the data structure used, the problems associated in adapting this type of data structure and how we solve them. Given a set of feature vectors \( \mathbf{v} = \{v_i, 1 \leq i \leq N\} \), the goal of the algorithm is to detect clusters and assign labels to the vectors based on the cluster they belong to. We use wavelet transform as a tool to facilitate clustering. The new idea here is to efficiently represent high dimensional data and perform wavelet transform as well as connected component analysis on this representation. The outline of the algorithm is given below:

Algorithm 1

Input: \( N \) image feature vectors
Output: The set \( \mathcal{F} \) of all detected clusters

1. Quantize feature space, aggregate feature vectors into cells and construct a hash table \( H \).
2. Apply wavelet transform on \( H \) and save the transformed data in a new hash table \( H' \).
3. Find the connected components (clusters) in \( H' \).
4. Assign labels to the cells and make the lookup table.

2 Querying

To retrieve database images which are similar to a query image, the distance between the query image and database images will be computed. We use Euclidean distance \( D \) to calculate the distance between the feature vector of the query image and the feature vectors of database images. The images with the distance smaller than a given threshold will be retrieved. We can speed up retrieval by searching only the related cluster instead of the whole database. Cluster selection can be achieved by checking the lookup table. Linear search within a cluster is straightforward, but is not efficient when a cluster is large. A novel index structure was built to help narrow down the search scope.

3 Demonstration

We evaluated the performance of ACQ and demonstrate its effectiveness and efficiency on image data sets. Tests were done on Photodisc data sets. We present the experiments on 128-dimension feature vectors of color histogram of images.

References