

Content-Based Retrieval of Images Using Pre-determined Image Classes

S.R. Subramanya, Sanan Srakaew, Abdou Youssef, Nikitas Alexandridis
Department of Electrical Engineering and Computer Science,
The George Washington University,
Washington, DC 20052.
{subra, srakaew, youssef, alexan}@seas.gwu.edu

Abstract

Image databases have applications in digital libraries, radiological image archives, satellite imagery for earth resources, law enforcement, etc. The huge data sizes of images, the enormous number of images in a typical image database, coupled with the inexact nature, and subjective interpretations have called for content-based retrieval systems. Fast and accurate retrievals are crucial for such systems to be used effectively. In this paper, we propose a scheme to speedup the retrieval of relevant images based on pre-determined image classes. The algorithm has been implemented and the results of retrieval are presented.

Keywords: Image databases, Image classes, Content-based retrievals.

1 Introduction

The increased use of images in various applications have led to the development of image databases [4, 5, 6, 7]. Example applications are quite diverse such as digital libraries, radiological images in medicine, satellite imagery for earth resources, CAD databases, photo journalism, art, law enforcement, and several others. The huge data sizes of images and the enormous number of items in the database, coupled with the inexact and subjective nature of image data have rendered keyword-based queries ineffective and have called for the use of content-based queries [1, 2]. Fast and accurate image retrievals for content-based user queries are crucial for such systems to be useful.

Although the content-based search could search the entire collection of images, it might be beneficial to have a classification of the images, so that, the search can be made to look for images more closely in certain classes than in other classes, in order to improve the search speed and search accuracy. Sometimes, even in a particular application, the images are very diverse and a classification might be beneficial. An automated classification scheme for images was pro-

posed in [9], which is based on *k-means clustering* algorithm. We use that classification algorithm to derive the image classes from the image database.

In this paper, we propose a scheme for efficient and accurate retrievals of images for content-based queries using the pre-determined image classes. The general idea is as follows: for each image class, a representative image is determined. The given query is matched with the representatives of all classes and a ranking of the classes is determined based on the distance of the query from the each of the class representatives. The search is then tuned to look more closely in the classes which are 'closer' to the query than the others. For comparison, we use the image retrieval scheme described in [8] (henceforth referred to as the 'earlier scheme'), where a given query is matched uniformly with all images in the database. The depth of K-tree to which the search can descend, is however, controlled depending on the required retrieval accuracy.

The next section gives a description of the notations and the image data model used in the proposed scheme, Section 3 presents the proposed algorithm. Experimental results are given in Section 4, followed by conclusions.

2 Notations and Image Data Model

In the following algorithms and their descriptions, the following notations are used:

- U : Set of all images. $|U| = N$.
- N : Total number of images in the system.
- I_i : An image i in the collection of images.
- M : Number of image classes.
- C_i : Image class i , $1 \leq i \leq M$.
- R_i : Representative image of class i .
- Q : Query Image.

2.1 Image data model

We use (1) color histograms and (2) texture information of the images as *features* representing the images. The (R, G, B) values of the image pixels are converted to (H, S, V) (hue, saturation, and value components) and then mapped to pre-selected domain of color set [3]. The texture of a region of an image is a 16-dimensional tuple of means obtained by wavelet (subband) decomposition. The color and texture histograms are derived for different resolutions - considering the whole image, then the four quadrants at the next level, then the 16 quadrants at the next lower level, and so on. The histograms are then organized in the K-tree. The global features appear at the top levels of the tree and local, finer details are stored at the lower levels. This structure enables tuning the searches for the required search speed and the search accuracy. It also supports *progressive transmission* which facilitates quick 'effective' retrievals of large amounts of data over networks. The similarity between images is determined by finding the similarity between the features (histograms), which is done by computing distances for each of the features and then a computing a combined distance, based on a weight vector.

3 The Proposed Algorithm

A brief description of the proposed algorithm is first given followed by the pseudocode. We assume that all the N images in the database are partitioned into M different classes using the scheme given in [9]. The proposed retrieval scheme uses the images classes to strike a good balance between the search speed and search accuracy. First, a *representative* R_i of each class C_i is determined. The representative could be either (1) the image in the class which is 'equidistant' from the rest of the images in the class, or (2) a 'synthetic' image which is equidistant from all the images in the class. Then, the distance d_i between the given query image Q and each of the representatives R_i , is determined and then the distances are sorted in non-decreasing order. Based on this order and the distances, the classes are grouped into a few 'super-classes'. The classes that are nearer to the given query are considered more promising, as they are likely to contain more images which are closer to the given query. So, the search process is designed to look more closely for matches in the classes which are closer to the query than those which are farther. This is done by descending lower in the K-tree of feature histograms for those images in the classes that are closer to the query. As a specific example, suppose that the classes are arranged in non-decreasing order

of distances from the given query. The top 30% of the classes form super-class 1, the next 30% form super-class 2, and the remaining 40% form super-class 4. The search algorithm descends to level 4 in the trees for all images in super-class 1, to level 3 in the trees for images in super-class 2, and only examines the top level of the trees for super-class 3. The 'earlier algorithm', on the other hand, descends to level 4 of the trees for *all* images.

Algorithm 3.1 RETRIEVE (in: $C_i, 1 \leq i \leq M, Q$; out: L (result list))

1. begin
2. FINDREP(C_i, R_i), $1 \leq i \leq M$.
{For each C_i find a representative R_i .}
3. FINDDISTANCE(Q, R_i, d_i), $1 \leq i \leq M$.
{Find the distances d_i between Q and R_i .}
4. Sort $d_i, 1 \leq i \leq M$ in non-decreasing order.
5. Form Super-classes based on the d_i 's.
6. Prioritize the searches in different classes.
7. Do the search in the classes based on the priorities.
8. Collect the results of search in all classes.
 $L \leftarrow \{\text{best } \beta \text{ matches}\}$. Return L .
9. end

Algorithm 3.2 FINDREP (in: C_m ; out: R_m)
{In: A class of images. Out: Representative.}

1. begin
2. for each $I_i \in C_m$ do
3. Compute $D(i, j) = \sum d(I_i, I_j), \forall j \neq i$
4. endfor
5. Find MIN($D(i, j)$), say $D(k, j)$.
{Find image k such that $D(k, j) \leq D(i, j)$ }
6. I_k is the new representative. $R_m \leftarrow I_k$.
7. end

Algorithm 3.3 FINDDISTANCE(in: I_i, I_j ; out: d)

{Input: two images, I_i, I_j . Output: distance.}

1. begin
{ $w = [w_1 w_2]$ is a feature weights vector determined empirically.}
2. Determine d_c and d_t , the distances between I_i and I_j with respect to
(1) color histogram and (2) texture.
3. Find the distance $d = d_c w_c + d_t w_t$.
4. end

4 Experimental Results

The proposed algorithm has been implemented and used to retrieve images from a database

of about 2,000 images. The database is built using images from Eastman Kodak homepage (URL: <http://www.kodak.com>) and Smithsonian Institute's 'Online Collection of Pictures' (URL: <ftp://photo1.si.edu>).

The images were classified into 20 classes and the retrieval algorithm was used to work on the classes.

Figures 1 and 2 show respectively, the results of retrievals using the earlier scheme [8] and the proposed algorithm which uses the pre-determined classes. The equal weights were assigned to both color and texture distances to derive the overall feature distance. (Work is underway evaluating different combinations of weights for the features).

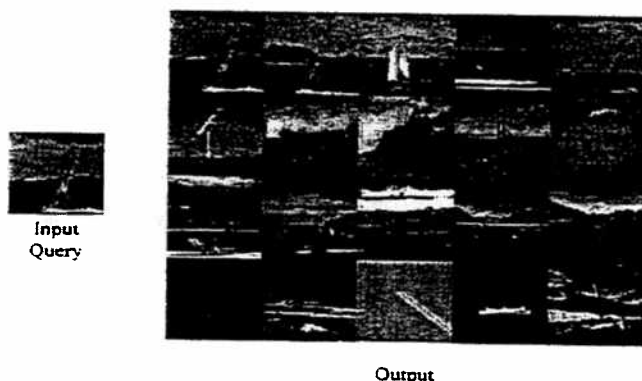


Figure 1: Retrieval results of earlier algorithm.

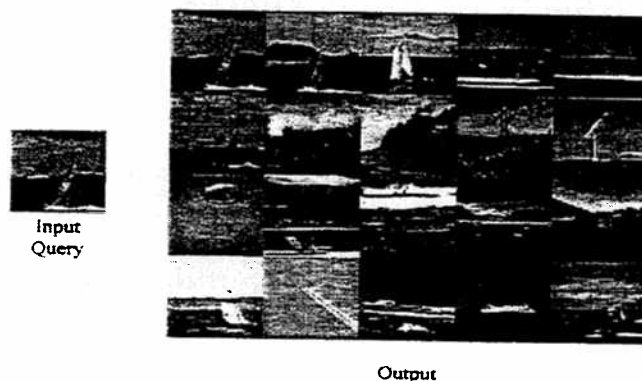


Figure 2: Retrieval results of proposed algorithm.

A comparison of the retrieval times of the two schemes is shown in Figure 3. It is observed that the proposed scheme is faster compared to the earlier scheme and gives comparable retrieval result accuracy.

Note: The visual relevance of the retrievals to the query is better seen in color images, since color is one of the two features used in determining the similarity between images.

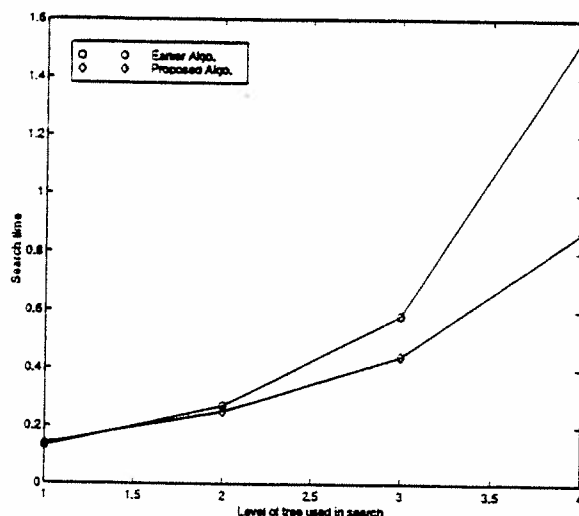


Figure 3: Comparison of search times.

5 Conclusions

Fast and accurate content-based retrievals of images are important in image databases. This paper proposed a scheme for the retrieval of images based on pre-determined image classes. The results show the improvement in speed and accuracy of the retrieval results compared to a scheme which works on the entire image collection uniformly.

References

- [1] Narasimhalu, A.D., ed. Special issue on content-based retrieval. *ACM Multimedia systems*, Vol. 3, No. 1, Feb 1995.
- [2] Gudivada, V. and Raghavan, V. (Eds.) 'Special Issue on Content-Based Image Retrieval Systems', *IEEE Computer*, Vol.28, No.9, September 1995.
- [3] Smith, J.R. and Chang, S-F. 'SaFe: A General Framework for Integrated Spatial and Feature Image Search', *IEEE Workshop on Multimedia Signal Processing*, 1997.
- [4] Kemp, Z. 'Multimedia and Spatial Information Systems', *IEEE Multimedia*, Vol.2, No.4, 1995.
- [5] *SPIE: Image Storage and Retrieval Systems*, SPIE, San Jose, CA. Feb. 1992.
- [6] Tamura, H. and Yokoya, N. 'Image Database Systems: A Survey', *Pattern Recognition*, Vol. 17, No. 1, 1984, pp29-43.
- [7] Chang, S-K. 'Image Information Systems', *Proc. IEEE*, Vol. 73, No. 4, April 1995, pp 754-764.
- [8] Piamsa-nga, P, et. al. 'A Unified Model for Multimedia Retrieval by Content', *Int'l. Conf. on Computers And Their Applications*, Hawaii, March 1998.
- [9] Subramanya, S.R., et. al. 'A Scheme for Automated Classification of Images', Tech. Report, GWU IIST.