A Feature Selection Framework for Small Sampling Data in Content-based Image Retrieval System

Kien-Ping Chung, Chun Che Fung and Kok Wai Wong
School of Information Technology
Murdoch University
Perth, Australia
k.chung, l.fung, k.wong@murdoch.edu.au

Abstract— **Content-based image retrieval (CBIR) systems have drawn interest from many researchers in recent years. Over the last few years, kernel-based approach has been a popular choice for the implementation of the relevance feedback based CBIR system. This is largely due to its ability to classify patterns with limited sample data. A long flat vector has been a popular choice for the input configuration. The reasons are because it is relatively easy to implement and more importantly, because it preserve the information of identifying the target images via different combination of image features. However, one of the biggest weaknesses of such configuration is the "curse of dimensionality". This paper introduces a relevance feedback framework via the use of statistical discriminant analysis method to select only relevant feature for next image retrieval cycle. Hence, minimize the dimensionality of the feature vector. This approach has been tested with four sets of images labelled with different themes. Each set contains 500 images, 50 labelled as positive while the rest are negative. The test showed an improvement from the previous flat input vector configuration when the training samples are relatively small.**

Keywords—Content-based image retrieval, Relevance Feedback, Statistical Discriminant Analysis, Feature Selection

I. INTRODUCTION

Content-based image retrieval system (CBIR) has been one of the most active areas of research in recent years. One of the main reasons is due to the explosion in the number of images available in digital format. Combining this factor along with the ease of transferring media through the Internet, digital images become very easy to be obtained by any person for personal, industrial or commercial purposes. In many industrial applications, an effective image retrieval system becomes essential for the access of the most appropriate images from the database. Examples of such systems are the automatic diagnostic system for medical applications, and, face detection for security surveillance system. CBIR is a system that retrieves an image based on the semantic or visual content of the query which can either be in the forms of key-words or image. The images in these systems are often indexed via key-words, image feature vectors or a combination of both.

Recently, relevance feedback in CBIR system has gained much attention from the research community. It is a strategy that invites interactive inputs from the user to refine the query for subsequent retrieval. This approach generally starts from prompting users to search the system via keywords, image examples or a combination of both. The system then prompts the user to select the relevant images from the search results. After the user selected the images, the system will refine the original query by analyzing the common features among the selected images. This process is continued iteratively until the target is found.

In a generic CBIR system, it is impossible to know what feature model/s can be used to capture the unique identity of certain groups of images. Hence, one idea is to employ as many image features as possible in hope that at least one has captured the unique feature of the targeted images. Such idea introduces problems if the image features are treated as a cascade of one flat vector. Such arrangement may increase the chances of "polluting" the feature element that uniquely identifies the selected image group. This is also known as the curse of dimensionality. This arises when the training samples are smaller than the feature vector. It has been an issue for a lot of the pattern recognition methods. Thus, a lot of the retrieval systems are restricted in using only a handful of features in an attempt to avoid the dimensionality problem.

This paper proposes a feature selection framework for the statistical discriminant analysis in content-based image retrieval system. `The idea is to calculate the discriminant ability of each image feature by using the training samples gather through the relevancy feedback cycle. The discriminant ability of each image feature is calculated by the ratio factor similar to the discriminant factor as proposed by Wang, Chan and Xue [1]. After this feature selection cycle, the important features are then used to analyze the rest of the image database and discriminant analysis is used to adjust the weight of each of the selected feature element.

This paper will first provide a brief description on the background of the theory, and follow by the description of the proposed framework. The paper will then look at the experiment conducted on the proposed method and lastly, conclusion will be drawn based on the experiment finding.

II. BACKGROUND

A. Previous Work

Over the past ten years, relevance feedback has evolved from a simple machine learning problem such as frequency of

occurrence [2], Bayesian classification [3], and now, to the more popular kernel based approach [4-7]. Kernel technique has long been used in the statistic pattern recognition applications. Its recent popularity gain in CBIR applications is mostly due to its ability to analyze small sample data and the classification of non-linear data. Tong and Chang [5] used support vector machine (SVM) as a machine learning tool for classifying images. The integration of kernel based approach with discriminant analysis is somehow similar to the SVM approach in that they both applied a kernel matrix for transforming the data into another feature space for ease of classification. The kernel bias discriminant analysis (KBDA) approach as reported by Zhou and Huang [7] have shown improvement from the SVM approach and other earlier proposals.

To the authors' knowledge, most of the reported relevance feedback CBIR systems, with the exception of MARS [2], have treated the input as a cascade of a flat vector. In this configuration, it is rather difficult to determine the discriminant ability of each extracted feature as they all are being treated in the same manner. As for the MARS system, each feature is configured in a hierarchical manner, and the similarity measure is performed at the inter and intra feature level. The similarity is first computed within each feature. This information is then used to calculate the similarity at the feature level. This framework is later extended by Chung and Fung [8] via KBDA to improve similarity measure at the inter and intra feature level. Such configuration excels if the targeted group of images can be identify by one of the included feature. However, it assumes each feature is independent from each other. Thus, it is not as effective if the targeted group of images can only be captured by a combination of different features.

The following sub-sections will provide proposal of a feature selection framework which is a modification of the existing MARS framework. It will first provide a brief description of statistical discriminant analysis and its development trend in the content-based image retrieval system. It will then discuss how the statistical discriminant analysis approach can be used for feature selection.

B. *Statistical Discriminant Analysis*

Statistical discriminant analysis is a pattern recognition approach that attempts to maximize the distances between different labelled data samples. This is achieved by calculating the scatter matrix of the inter- and intra-classes of the different data samples. The scatter matrix is represented in the form of a covariance matrix. The goal of the discriminant analysis is to find a *weight matrix* such that the distances between the two scatter class matrixes are maximized. The problem can be expressed as:

$$W_{opt} = \arg\max_{w} \frac{\left\| W^T S_y W \right\|}{\left\| W^T S_x W \right\|} \qquad (1)$$

One can view the above expression as a problem of generalized eigen-analysis where the optimal eigenvectors associated with the largest eigen-values are the weight factor for the new feature space. By knowing the value of the weights, one can project the new input pattern z onto the new space:

$$new\_z = w^T z \qquad (2)$$

The inner-class matrix, it is often expressed as:

$$S_x = \sum_{i=1}^{Ny} (x_i - m_x)(x_i - m_x)^T \qquad (3)$$

where $\{x_i = 1,\ldots,N_x\}$ denotes the positive examples, and, m represents the mean vector of the positive samples. As for the inter-class matrix, different types of discriminant analysis approach will have their own corresponding inter-class scatter matrix. There have been numerous discriminant approaches that have been proposed in the literatures over the years. This paper will only provide the definition of the inter-class matrix for the *nonparametric discriminant analysis* (NDA) and the *kernel bias discriminant analysis* (KBDA), simply because they have shown to be effective in CBIR systems.

C. *Kernel Bias Discrminant Analysis (KBDA)*

KBDA [7] is the kernel version of the BDA [1]. It is based on the *(1 + x)* class biased learning problem implying there are unknown number of possible classes, however, the system is only interested in the positive labelled class. It is assumed that the positive labelled training samples are all visually related and they can be classified into one image group. Under this assumption, the objective of the intra covariant matrix of BDA is to maximise the distances of the negative training samples from the positive centroid. The intra covariant matrix of BDA can be express as:

$$S_y = \sum_{i=1}^{Ny} (y_i - m_x)(y_i - m_x)^T \qquad (4)$$

where $\{y_i = 1,\ldots,N_y\}$ denotes the negative examples, and again, m represents the mean vector of the positive samples.

There are two major drawbacks on the proposed approach. Firstly, the regularization approach as used by [7] for avoiding matrix singularity problem is often unstable. Secondly, the parameters used in the kernel function require to be manually tuned for maximum retrieval accuracy. To solve the first issue, Tao and Tang [9] reported a full rank null-space method for calculating the eigenvalues and vectors of the inter and intra covariant scatter matrix. As for the second issue, Wang, Chan and Xue [1] have very recently suggested that the accuracy performance of the kernel approach can be optimised by maximising the discriminant ratio of inter and intra covariant matrix. The optimization of the ratio is solved by applying Broyden-Fletcher-Goldfarb-Shanno (BFGS) Quasi-Newton method. This is simply a method of using the gradient information for locating the turning point of a curve. The proposed approach has shown to be very close to the

maximum accuracy KBDA can possibly get by manually tuning the kernel parameter. However, this comes with the additional overhead cost of optimising the discriminant ratio.

Alternatively, Tao and Tang [9] have suggested the used of NDA approach. Kernel transformation is not required in this approach. However, this approach can only barely match the accuracy performance of KBDA. The following section provides a brief description of the NDA approach.

D. Nonparametric Discriminant Analysis (NDA)

In the nonparametric discriminant analysis approach, the inter-class scatter covariant matrix is derived from the distances obtained from the vectors pointing to the centroid of *another class* of the sample data. The main advantage of NDA over other statistical discriminant analysis such as the linear discriminant analysis (LDA) [10] and bias discriminant analysis (BDA) [7] is that NDA does not require all positive samples to be based on a single Gaussian distribution. Hence, NDA does not require an additional kernel transformation matrix to transform the non-linearly related data to a new feature space for analysis. It therefore eliminates the use of extra parameters. This scatter inter-class matrix for NDA is normally expressed as:

$$S_y = \sum_{i=1}^{N_y} \left(y_i - m_{yi}^{kx}\right)\left(y_i - m_{yi}^{kx}\right)^T + \sum_{i=1}^{N_x} \left(x_i - m_{yi}^{ky}\right)\left(x_i - m_{yi}^{ky}\right)^T \quad (5)$$

The inner-class scatter matrix of the NDA is very similar to the other discriminant analysis methods as expressed in (3). It is expressed as:

$$S_x = \sum_{i=1}^{N_y} \left(x_i - m_x^{kx}\right)\left(x_i - m_x^{kx}\right)^T \quad (6)$$

where k is the k^{th} nearest sample to the input x_i.

E. Proposed Selection Approach

Fig. 1 illustrates the abstract architectural model of the proposed method. It is essentially a simple statistical discriminant framework, except the authors of this paper have added the feature selection process through the relevance feedback from the users. After the relevance feedback, the discriminant ability of each feature is analyzed individually. Only the selected features will be used for performing the similarity measurement in the next retrieval cycle. The discriminant ability of each feature can be analyzed separately using the ratio as follow:

$$ratio = \frac{\left(\sum_{n=1}^{N_y} d'_n \Big/ N_y\right)}{\max(dp)} \quad (7)$$

$$d'_n = \begin{cases} \max(d_p)\, d_n > \max(d_p) \\ d_n, d_n \le \max(d_p) \end{cases} \quad (8)$$

where d_p and d_n are the calculated distances of the respective positive and negative label sample from the positive centroid. N_y is the total number of negative samples. The equations are formulated to provide the framework with a clear indication of the overlapping of the data. Unlike the ratio as used in (1), a ratio of 1 in (7) implies there is no overlap between the two training samples, vice versa, a ratio less than 0.5 indicates the analyse feature is unable to capture unique identity of the selected positive labelled images. With the ratio as used in (1), a ratio value of larger than 1 only implies the *average* of the distances of the negative samples are larger than the distances of the positive samples. It provides no clear indication if the training samples are still overlapping from each other.

With the used of the ratio as calculated from (7), the feature selection process becomes rather simple. Using the training samples, the system first calculates the ratio. The selection process is then based on the threshold value pre-set by the system. A feature is selected when its calculated ratio value is over this threshold value. After the selection, the selected features are then once again cascaded into a flat feature vector ready for another retrieval cycle.

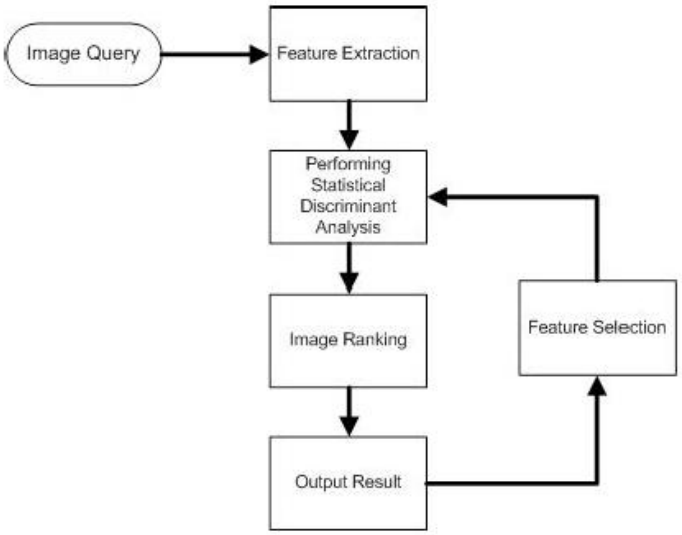

Figure 1 Proposed Feature Selection Model

The main task of the feature selection module is to determine the discriminant ability of the feature given the training samples. Since different statistical discriminant methods are based on different assumptions and they all yield different result. The logical choice of the analysis tool for the feature selection module, therefore, will be the same as the one used in calculating the similarity of the images.

III. EXPERIMENT

A. Prototype System

To evaluate the performance of the proposed approach, the authors have designed and implemented a prototype CBIR system using Matlab. This prototype system provides the user with the ability to query the image database with an image sample. After the first retrieval iteration, users can select the relevant images while ignoring the non-relevant images. The system will label the selected images as positive while treating the ignored images as negative. The retrieval procedure of the prototype system is as follows:

1. User inputs a query image.
2. The visual features of the query are extracted by the system.
3. All images in the database are sorted in ascending order based on the distance of dissimilarity.
4. Display the top 20 images with the highest rank.
5. User selects the positive images and the rest will be automatically labelled as negatives.
6. Use the labelled images, perform the proposed feature selection process and select the features accordingly.
7. Query the images in the database and project to a transformed space based on the nominated statistical discriminant approach with the selected features.
8. Rank and display the top 20 images that have not been labelled by the user.
9. Go back to step 5 for the next retrieval cycle.

B. Test Environment

This paper has implemented the NDA and KBDA methods to test the proposed feature selection framework. Each of the statistical approaches will be compared with their respective feature selection framework. To ensure the test environment for the proposed framework and the normal approach is as close as possible. The images features, the generalised eigenvector calculation method have to be the same and the kernel transformation algorithm for KBDA all have to be the same. The purpose is to evaluate the accuracy performance of the two systems under the same testing environment.

For this experiment, the authors have selected the water-filling edge histogram [11], HSV histogram, global edge direction histogram [12], HSV colour moments [13], RGB histogram and colour intensity histogram. Altogether, there are six features. Each feature comprises a number of elements. A total number of eighty-three feature elements have been used for this testing. Secondly, the calculation of the generalised eigenvalues are performed by the full rank method as reported by Tao and Tang [9]. The full rank method tends to provide a more stable solution then regularization approach as adopted in [1, 7]. Lastly, RBF is selected as the kernel transformation matrix for the KBDA approach. Reason being, literatures [1, 7] have both reported that such RBF yields the best accuracy performance out of all the other kernel transformation approach.

The threshold ratio of the feature selection process is set to 0.8. This implies that any feature that has a ratio of greater that the threshold value will be selected for the next retrieval cycle. This threshold value is used to allow a degree of data overlapping from the two labelled training samples. Reason being, it is often the case that certain feature can only uniquely capture the common characteristic of a sub-group of the positive labelled training samples, and a combination of features are needed to fully discriminate the whole training sets. Thus, the allowance of data overlap allows the system to retain such features.

C. Test Results and Discussion

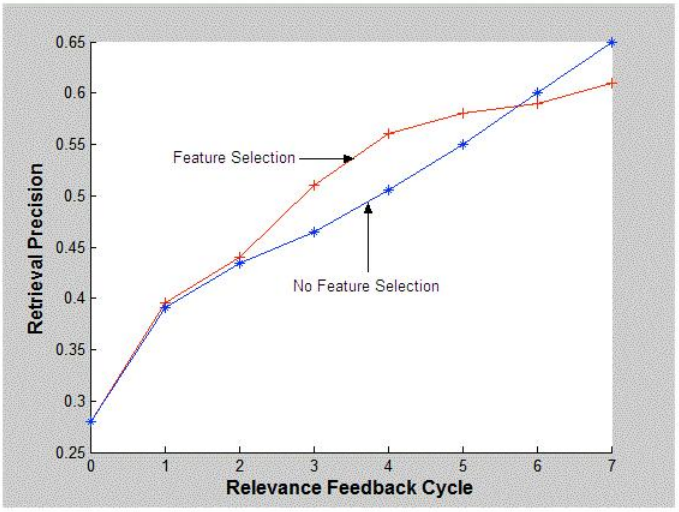

Figure 2 Retrieval Result for KBDA with and without feature selection.

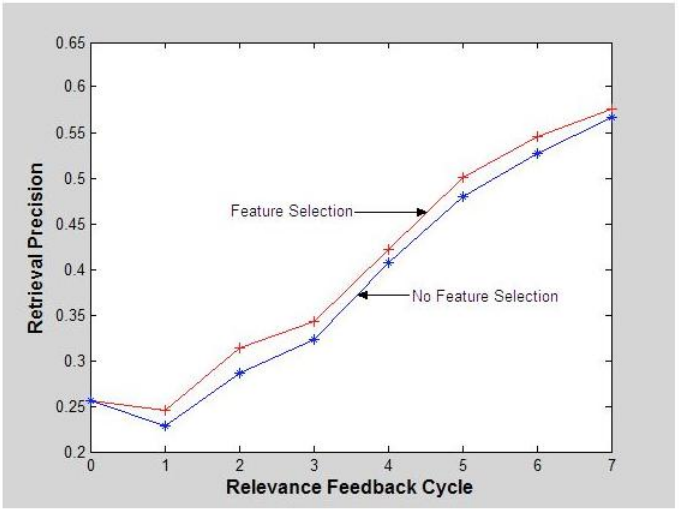

Figure 3 Retrieval Result for NDA with and without feature selection.

Figure 2 and 3 show the test result of the KBDA and NDA with and without the feature selection framework. The retrieval results are generated from 2000 images of the Corel image database. The images were retrieved and classified under four different themes. Each set of images contains 500 images with 50 of them labelled as positive while negative with the rest. The test was carried out by using each positive

labelled image as an input entry point to the system. The retrieval precision is then recorded for the first seven relevance feedback cycles. The result shown is the average retrieval precision of all 200 inputs of the first seven feedback cycles.

The above results have shown that the feature selection approach is more accurate than the non selection approach during the first few feedback cycles. However, the non selection approach of the KBDA becomes more accurate than its respective feature selection approach on the sixth cycle onward. This illustrates that feature selection approach is more superior when the training samples are small. However, as the system gathers more training samples, statistical discriminant approach such as KBDA may becomes more stable and any data filtering will result in lost of potential valuable information and thus, the accuracy.

In addition to the accuracy testing, the retrieval time of the prototype has also been recorded. Table 1 are the average time of each input samples for the relevance feedback cycles. The table recorded the average time for the KBDA and NDA with and without the feature selection framework, and the calculation time for the distance measure for each of the input images after the nominated statistical discriminant analysis approach is performed. From the table, it can be seen that the feature selection framework is more expensive than the normal approach. However, judging from the huge differences in computation time between the KBDA and NDA, it is obvious that the additional computation cost of the feature selection module is largely dependent on the nominated statistical discriminant method. The feature selection framework is approximately double the computational cost of the normal approach. Having said this, since the computation time is in seconds, the additional overhead cost is still acceptable.

Table 1. The average time, in seconds, of each input samples for the relevance feedback cycles.

Cycle	Image Ranking	KBDA	KBDA with Selection	NDA	NDA with Selection
1	0.1	0.4	0.8	0.1	0.3
2	0.1	0.3	0.4	0.1	0.1
3	0.1	0.4	0.8	0.1	0.2
4	0.1	0.3	0.8	0.1	0.2
5	0.1	0.6	1.5	0.1	0.2
6	0.1	0.7	1.7	0.1	0.2
7	0.1	0.9	2.0	0.1	0.3

IV. CONCLUSION

A feature selection framework for the relevance feedback content-based image retrieval system has been introduced in this paper. This feature selection framework is primary designed for small training sample problem. The experiment result has been shown to support this goal.

This paper has used a rather naïve way of selecting the appropriate features for similarity measurement. The selection decision is solely dependent on the discriminant factor of the training samples. Quite often, this can be misleading. Thus, one of the immediate research goals is to explore alternative ways of performing feature selection.

REFERENCE

[1] L. Wang, K. L. Chan, and P. Xue, "A Crterion for Optimizing Kernel Parameters in KBDA for Image Retrieval," *IEEE Transactions on Systems, Man, and Cypernetics*, vol. 35, pp. 556 - 562, 2005.

[2] Y. Rui, T. S. Huang, and S. Mehrotra, "Content-Based Image Retrieval with Relevance Feedback in MARS," International Conference on Image Processing, Washington, DC, October 26-29 1997.

[3] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos, "PicHunter: Bayesian Relevance Feedback for Image Retrieval," Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, August 25-29 1996.

[4] D. Tao and X. Tang, "A Direct Method to Solve the Biased Discriminant Analysis in Kernel Feature Space for Content BAsed Image Retrieval," Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing, Montreal Canada, May 19 2004.

[5] S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval," Proceedings of 9th ACM International Multimedia Conference, Ottawa, Cananda, September 30 - October 05 2001.

[6] G.-D. Guo, A. K. Jain, W.-Y. Ma, and H. J. Zhang, "Learning Similarity Measure for Natural Image Retrieval With Relevance Feedback," *IEEE Transactions on Neural Networks*, vol. 13, pp. 811-820, 2002.

[7] X. S. Zhou and T. S. Huang, "Small Sample Learning During Multimedia Retrieval using BiasMap," IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, United States, December 2001.

[8] K.-P. Chung and C. C. Fung, "Multiple Layer Kernel-Based Appraoch in Relevance Feedback Content-based Image Retrieval System," The Fourth International Conference on Machine Learning Cybernetics, Guangzhou, China, August 18 - 21 2005.

[9] D. Tao and X. Tang, "Nonparametric Discriminant Analysis in Relevance for Content-based Image Retrieval," Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, United Kingdom, August 23 - 26 2004.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed: Academic Press, INC, 1990.

[11] X. S. Zhou and T. S. Huang, "Edge-based Structural Features for Content-based Image Retrieval," *Pattern Recognition Letters*, vol. 22, pp. 457 - 468, 2001.

[12] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient Use of Local Edge Histogram Descriptor," Proceedings of the 2000 ACM workshops on Multimedia, Los Angles, Calfornia, United States 2000.

[13] M. Stricker and M. Orengo, "Similarity of Color Images," Storage and Retrieval for Image and Video Databases III, San Diego/La Jolla, CA, USA, February 5-10 1995.