FEATURE SELECTION FOR
CONTENT-BASED IMAGE RETRIEVAL
USING STATISTICAL DISCRIMINANT ANALYSIS

TEE CHENG SIEW

UNIVERSITI TEKNOLOGI MALAYSIA

ABSTRACT

As we known, the very large repository of digital media arise the challenge of various digital search applications. In order to make use of this huge amount of data, effective tools are required for retrieve multimedia information. An image retrieval system is one of the tools that can be used for searching and retrieving images from a large database of digital images. However, there are several challenges and problems need to be considered when applied image retrieval system such as the gap between high-level semantic concept and low-level visual features. This refers to problem of feature selection, which is critical to really solve the gap problem in CBIR. Recently, the most feasible feature selection method is discriminant analysis. Therefore, in this project, we proposed title feature selection in content-based image retrieval using statistical discriminant analysis. In the project, we intended to enhance performance by improve the feature selection process. Besides, we used fuzzy theory in content-based image retrieval to solve the problem of perspective subjectivity of human in image retrieval. The system would be more depends to the human-like and how to response with relevant images that match the concept of current query is always the research question in this project.

ABSTRAK

Seperti yang kita tahu, layan web membenarkan pengkumpulan maklumat yang berbilang besar. Maka, masalah wujud apabila seseorang pengguna ingin mencari gambar yang berkaitan sahaja dalam database yang begitu besar. Oleh sebab itu, teknik yang berkesan adalah diperlukan untuk pengguna dalam pencarian imej atau gambar. Sistem pencarian imej adalah salah satu teknik yang berkeupayaan melaksana pencarian imej dalam database yang berbilang besar. Walaubagaimanapun, terdapat juga beberapa cabaran perlu difikirkan semasa mengguna sistem pencarian imej seperti masalah sistem mengkenalpasti imej yang diperlukan oleh pengguna. Ini disebabkan sistem hanya boleh mencari imej berdasarkan ciri-ciri yang ada dalam suatu imej. Dengan menjalankan proses pengestrakan imej dan proses memkenalpasti kesamaan imej, sistem papar imej yang mempunyai kesamaan yang banyak. Oleh sebab itu, kami bercadang tajuk pengestrakan imej yang bermakna untuk membuat perbandingan dengan imej di database dalam sistem pencarian imej. Projek ini bercadang meningkatkan ketepatan pencarian imej dengan memilih ciri-ciri yang bermakna sahaja untuk membuat perbandingan imej. Kami juga bercadang menggunakan teori Fuzzy supaya memperbaiki masalah pencarian imej berdasarkan keinginan pengguna. Sistem ini adalah lebih mengikut keperluan pengguna dan bagaimana papar imej yang berkaitan dengan keperluan pengguna merupakan soalan pengkajian dalam projeck ini.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

With the advances in computer technologies and the advent of the World-Wide Web, there has been an explosion in the amount and complexity of digital data being generated, stored, transmitted, analyzed, and accessed. Much of this information is multimedia in nature, including digital images, video, audio, graphics, and text data. In order to make use of this vast amount of data, efficient and effective techniques to retrieve multimedia information based on its content need to be developed. Among the various media types, images are of prime importance. Thus, this study deals with the feature selection investigation in the relevance feedback learning process to improve the retrieval performance of the region-based image retrieval system.

An image retrieval (Wikipedia, access on 2008) system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. The reason behind research on multimedia systems and content-based image retrieval (CBIR) is the fact that multimedia databases deal with text, audio, video and image data which could provide enormous amount of information and which has affected life style of human for the better.

CBIR is the application of computer vision to the image retrieval problem, that is, the problem of searching for digital images in large databases. Content-based image retrieval (Wikipedia, access on 2008) also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR). "Content-based" means that the search will analyze the actual contents of the image. The term 'content' in this context might refer colors, shapes, textures, or any other information that can be derived from the image itself. There are some types of feature used for image retrieval such as colour retrieval, textual retrieval, shape retrieval and so on. Figure 1.1 show diagram fundamental of content-based image retrieval system.

Relevance feedback (RF) learning is a common approach that gathering semantic information from user interaction and attempts to reduce the semantic gap between high-level concepts and low-level features. Based on feedback, the user labels each image returned in the previous query round as relevant or irrelevant. The retrieval scheme is adjusted and the next set of images is presented to the user for labeling. There are two main RF learning approaches have been used which are distance re-weighting and query modification. Query modification changes the representation of the user's query in a form that is closer to the semantic intent of the user. Distance re-weighting changes the calculation of image to image similarity to strengthen the contribution of relevant image components in regard to the current query.

Feature selection (FS) is the technique used to select and find an optimal subset of relevant features in machine learning. FS technique select only the relevant features and remove most irrelevant or redundant features from the data to improve the performance of learning models. FS technique can be used to improve performance by alleviating the effect of the curse of dimensionality, enhancing generalization capability, speeding up learning process, and also improving model interpretability. In CBIR system, feature selection is a keys step in relevance feedback. Since it ideally to select the optimal feature subset (most important features) and their combinations for describing and querying items in the database to reduce retrieval (time and computational) complexity while maintaining high retrieval performance. In other words, how to select the optimal feature subset to describe a learning system is always regarded as a key technology in domain of machine learning.

Research works had been undertaken in the past decade to design efficient image retrieval techniques from the image or multimedia databases. Although large numbers of indexing and retrieval techniques have been developed, there are still no universally accepted feature extraction, indexing and retrieval techniques available. Despite the shortcomings of current CBIR technology, several image retrieval systems are now available as commercial packages, with demonstration versions of many others available on the Web. Some of the most prominent such as commercial systems and experimental systems. Beside that, a wide range of possible applications for CBIR technology has been identified (Gudivada V N and Raghavan V V, 1995). Some potentially fruitful areas include crime prevention, intellectual property, architectural and engineering design, medical diagnosis, geographical information and remote sensing systems, education and training, web searching.

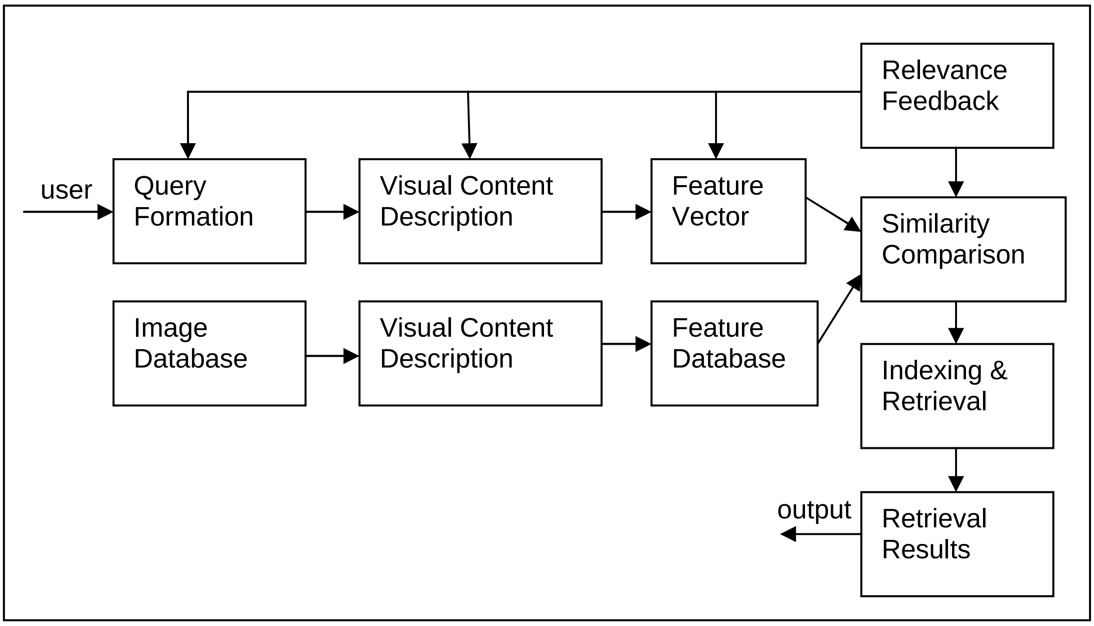

Figure 1.1: Diagram fundamental of content-based image retrieval system (F. Long and H. Zhang)

1.2 Problem Background

Content-based image retrieval (CBIR) has become one of the most active research areas in the past few years. Thus, many visual feature representations have been explored and many CBIR systems have been built. However, there are several problems and challenges need to be consider in attempt to apply CBIR systems.

Firstly, the gap between high-level semantic concept and low-level visual features is great. In the CBIR context, an image is represented by a set of low-level visual features which are the features have no direct correlation with high-level semantic concept. Human prefer to retrieve images according to the “semantic” or “concept” of an image. But, CBIR depends on the absolute distance of image features to retrieve similar images (X. Wang, K. Xie, 2005). Thereby, appear the gap between high-level concepts and low-level features which is the major difficulty that hinders further development of CBIR systems (W. Jiang, G. Er and Q. Dai and J. Gu, 2006). In other sentence, the semantic gap problem is the lack of coincidence between the image representation and the human interpretation for an image (C. Chiu, H. Lin, 2003).

Furthermore, each person has his or her own subjective when thinking about something. So, we know that different user normally may have different subjective perception for same image (C. Chiu, H. Lin, S. Yang, 2003). In such case, the semantic concept and semantic view of same image which user desire to retrieve image may different in perception of user. The question should be considered in this problem is (X. Wang, K. Xie, 2005):

How to reflect subjective perception of user in image retrieval is rather few?

Beside that, the feature extraction of images costs much time due to the large amount of images is the other frequently in current CBIR systems. This results in the slowness of image retrieval (H. Xuelong and Y. Li, 2007). Moreover, CBIR learning has several challenges (W. Jiang, G. Er and Q. Dai and J. Gu, 2006)(F. Li,

Q. Dai, W. Xu, 2006)(W. Jiang. M. Li, H. Zhang, J. Gu, 2004)(J. Yu, Y. Lu, N. Sebe, Q. Tian, 2007) shown as below:

i. Small size of the training set.

   The training samples are the labeled images from the user during each query session, which are very few compared with the feature dimensionality and the size of the database. The CBIR learning algorithm usually encounters severe problems due to the curse of dimensionality.

ii. Intrinsic asymmetry.

   The images labeled to be “relevant” during a query session share some common semantic cues, while the “irrelevant” images are different from the “relevant” ones in different ways. Thus, the “relevant” images are more important for the system to grasp the query concept. This asymmetry requirement makes it necessary to treat the “relevant” and “irrelevant” sets unequally with an emphasis on the “relevant” one.

iii. High Dimensionality.

   In many data analysis application, the observed data have very high dimensionality. Specifically the images in CBIR are represented by image feature vector whose dimensionality ranges from tens to hundreds in most cases.

To really face the fundamental problem of CBIR, two effective learning mechanisms have been proposed which are relevance feedback and region-based image retrieval (RBIR). In general, the RBIR approaches (W. Jiang, G. Er and Q. Dai and J. Gu, 2006) segment images into several regions and retrieve images with low-level features extracted based on these regions. Since region-based features represent images in the object level and better fit the human perception than the global low-level features extracted from the entire image. But, in the RBIR context (F. Li, Q. Dai, W. Xu, 2006), it is difficult to represent the images of the database in a uniform feature space, because the number of regions in different images may not be the same. Thus, not much work has been done to deal with the problem of online feature selection in RBIR systems. Beside, the relevance feedback mechanism is considered as an iterative supervised learning process (W. Jiang, G. Er and Q. Dai and J. Gu, 2006)(F. Li, Q. Dai, W. Xu, 2006) and widely used to improve the performance of CBIR. A CBIR system with relevance feedback can refine the query by users' feedback to improve the retrieval accuracy and performance (C. Chiu, H. Lin, S. Yang, 2003). Query-point moving and weight updating approach are the most two research approaches used in relevance feedback.

Feature selection is a key step in relevance feedback. In other word, how to select a subset of features from a large-scale feature pool and to construct a suitable dissimilarity measure are the key steps in RF (D. Tao, X. Tang, 2004). Thus, problem of feature selection is critical to really solve the fundamental problem in CBIR. There are many existing feature selection techniques such as distribution-based approaches, Kullback-Leibler divergence (K-LD), boosting manner, discriminant analysis (DA) method and others. However, these feature selection techniques remains a challenging problem for image retrieval.

In recent year, there are a lot of discriminant analysis method had been proposed and used as a feature selection method to improve relevance feedback. These methods included multiple discriminant analysis (MDA), biased discriminant analysis (BDA), kernel-biased discriminant analysis (KBDA) and nonparametric Discriminant analysis (NDA). The goal of discriminant analysis is to find a weight matrix such that the distances between the two scatter class matrixes are maximized. However, these methods have their own drawback that must be solved to improve the performance of CBIR. Basic single Gaussian assumption which proposed by MDA and BDA usually doesn't hold, since the few training samples are always scattered in the high dimensional feature space, and their effectiveness will be suffer (W. Jiang. M. Li, H. Zhang, J. Gu, 2004). Moreover, single Gaussian distribution means all positive samples should be similar with similar view angle and similar illumination, which are not the case for CBIR. To overcome the problem of single Gaussian distribution assumption, KBDA had been introduced. But, this kernel-based method has two major drawbacks which is regularization approach is often unstable and it is rely on parameter tuning. Then, NDA had been proposed to solve the problem in MDA, BDA and also KBDA. This approach can only barely match the accuracy performance of KBDA. As a conclusion, many feature selection methods can not satisfy the requirements in CBIR even though there are many method has been apply in content-based image retrieval.

1.3 Problem Statement

In all CBIR systems, the learning process must tackle a fundamental problem which is the gap between high-level semantics and low-level visual features in content-based image retrieval. To really overcome the problem, a problem that must be solved is:

Which features are more representative and suitable for explaining the current query concept?

This question can be refers to the problem of feature selection. By now, the most feasible feature selection method is discriminant analysis. The latest problems faced by statistical discriminant analysis are single Gaussian distribution problem and parameter tuning problem.

In a generic CBIR system, images are represented by image feature vector whose dimensionality ranges are high, which from tens to hundreds in most cases. Also, it is impossible to know what features 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 query image. But, this idea introduces problems because its arrangement may increase the chances of "polluting" the feature element that uniquely identifies the selected image group. This also known as curse of dimensionality and reflect on high dimensionality. To minimize the high dimensionality, a problem must be solved is:

How to select only relevant features to minimize the dimensionality of feature vector?

Beside that, human perception subjectivity is a problem in CBIR which mentioned that different user may has different perception of same images. Thus, the real problem in this point is:

How to reflect the different perception of different user of same image in image retrieval?

This question can be refers to the problem of human thinking. By using relevance feedback in CBIR, we can grasp the current query concept of user.

Therefore, in this study, the improved feature selection method in CBIR system has been proposed to covers the issues with below question:

i. How to reflect human-like and enhance the performance of CBIR system.
ii. Which is the suitable feature selection method can be used to select optimal features in relevance feedback of image retrieval?

Due to current feature selection methods through relevance feedback could not fully representative the query concept from user which make the performance of CBIR is not satisfied. Hence, a better feature selection method is required and the issue of features selection through relevance feedback in CBIR would be highlight or focus in this study.

1.4 Project Aim

The project aims to improve relevance feedback learning process of image retrieval through the feature selection method.

1.5 Objectives

In order to accomplish the purpose of the study, objectives of study have been identified as states below.

1. To investigate the feature selection methods in content-based image retrieval.
2. To enhance the performance of content-based image retrieval by modifying the statistical discriminant analysis feature selection method.
3. To evaluate the performance of improved feature selection method by using standard information retrieval formula such as precision, recall and F1.

1.6 Project Scope

The main focus of this study is on the improved retrieval performance of similarity-based feature selection in content-based image retrieval system. The scopes of this project as follows:

i. The samples consist of 10 images categories which are bus, horse, flower, human, building, food, beach, mountain, dinosaur and elephant.
ii. Both the color feature extraction and shape feature extraction are using in content-based image retrieval.
iii. For color feature extraction, colorspace HSV have been used to extract the color contents of an image.
iv. For shape feature extraction, canny operator has been used to extract the shape contents of an image.
v. Statistic discriminant analysis feature selection methods are used to select optimal features subset.

1.7 Significance of Project

i. The project can significantly improve the performance of feature selection process and the effect of feature selection towards the performance of CBIR system can be shown.

ii. Besides that, this project is important to identify the effectiveness of feature selection in content-based image retrieval. This is because features are more representative for explaining the current query concept than the others.

iii. Furthermore, the effectiveness and efficiency of project is important for future study and can be implemented in real world application.

1.8 Organization of Project

This report consists of five chapters. Chapter 1 presents the introduction of the study, problem background, the hypothesis, project aims, objectives, project scope and significance of project. Next, chapter 2 gives literature reviews on the content-based image retrieval, feature extraction methods, feature selection methods, similarity measurement, relevance feedback, theory of fuzzy logic. Methodology of the study is discusses in chapter 3. Besides, chapter 4 discuss about the experimental result. Lastly, the conclusion and suggestion for future work are explained in chapter 5.

1.9 Summary

In this chapter, the main subjects of the work have been introduced including the problem background, problem statement, objectives, project aims, and scopes, significance of project and organization of project.

REFERENCES

C. Chiu, H. Lin, S. Yang. (2003). A Fuzzy Logic CBIR System. The IEEE International Conference on Fuzzy Systems. pp. 1171- 1176.

D. Tao, X. Tang. (August 2004). Nonparametric Discriminant Analysis is Relevance Feedback for Content-based Image Retrieval. IEEE. Proceedings of the 17th International Conference on Pattern Recognition. pp. 310-314.

Esin Guldogan, Moncef Gabbouj. (2006). Feature Selection For Content-Based Image Retrieval. SIVip. Springer.

F. Li, Q. Dai, W. Xu. (2006). Improved Similarity-Based Online Feature Selection In Region-Based Image Retrieval. IEEE. ICME. pp. 349-352

F. Long, H. Zhang, David D. Feng. Fundamentals of Content-based Image Retrieval.

Gudivada V N, Raghavan V V. (1995a). Content-based Image Retrieval Systems IEEE Computer 28(9), 18-22

H. Xuelong, Y. Li, Z. Gensheng. (2007). Content-based Image Retrieval Using Fuzzy Hamming Distance. IEEE. The Eighth International Conference on Electronic Measurement and Instruments. 2-826-2-830.

H. Zhang. (2003). Learning Semantics in Content-Based Image Retrieval. Proccedings of the 3rd International Symposium on Image and Signal Processing and Analysis. pp. 284-288.

J. Yu, Y. Lu, N. Sebe, Q. Tian. (2007). Integrating Relevance Feedback In Boosting For Content-Based Image Retrieval. IEEE. ICASSP. pp. I-965-I-968

K. Chung , C. Chun, W. Kok. (2005). A Feature Selection Framework For Small Sampling Data In Content-based Image Retrieval System. School of Information Technology Murdoch University Perth, Australia. IEEE. pp. 310-314

L. Wang, K. Chan, P. Xue. (2005). A Criterion for Optimizing Kernel Parameters in KBDA for Image Retrieval. IEEE Transaction on systems, Man, and cybernetics. pp. 556-562.

Maytham Safar, Cyrus Shahabi, X. Sun. (2000). Image Retrieval By Shape: A Comparative Study. IEEE. pp. 141-144.

Remco C. Veltkamp, Mirela Tanase. (October 2002). Content-based Image Retrieval Systems: A Survey. pp. 1-62

Ricardo da Silva Torres, Alexandre Xavier Falcao. (2006). Content-based Image Retrieval: Theory and Applications. RITA. Volume XIII. Numero 2. pp. 165-189

S. Santini and R. Jain. (September 1999). Similarity Measures. IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 9, pp. 871–883.

Wikipedia, the free encyclopedia. <http://en.wikipedia.org/wiki/Image_retrieval> (Accessed on 1 February 2008)

Wikipedia, the free encyclopedia. <http://en.wikipedia.org/wiki/CBIR> (Accessed on 1 February 2008)

Wikipedia, the free encyclopedia. <http://en.wikipedia.org/wiki/Segmentation_(image_processing)> (accessed on 20 May 2008)

W. Jiang, G. Er, Q. Dai and J. Gu. (2006). Similarity-Based Online Feature Selection In Content-Based Image Retrieval. IEEE Trans. Image Processing, 15 (3), pp. 702-712.

W. Jiang. M. Li, H. Zhang, J. Gu. (2004). Online feature Selection based on Generalized Feature Contrast Model. IEEE International Conference on Multimedia and Expo(ICME). pp. 1995-1998

X. Wang, K. Xie. (April 2005). Application of the Fuzzy Logic in Content-based Image Retrieval. 5(1). pp. 19-24.

X. zhou, Thomas S. Huang. (December 2001). Small Sample Learning during Multimedia Retrieval using BiasMap. IEEE Conference on computer Vision and Pattern Recognition, Hawaii, United States. pp. I-11-I-17.

Y. Chen, James Z. Wang. (September 2002). A Region-based Fuzzy Feature Matching Approach To Content-based Image retrieval. IEEE Transaction On Pattern Analysis and Machine Intelligence. Vol. 24. No. 9. pp. 1252-1267.

Y. Choi, D. Kim, and R. Krishnapuram. (2000). Relevance Feedback for Content-based Image Retrieval Using the Choquet Integral. Proc. IEEE Int'l Conf. Multimedia and Expo, vol. 2, pp. 1207-1210.

Y. Huang, T. Chang, C. Huang. (27-29 July 2003). A Fuzzy Feature Clustering with Relevance Feedback Approach to Content-based Image Retrieval. VECIMS 2003-International Symposium on Virtual Environments, Human-Computer Interfaces, and Measurement Systems Lugano, Switzerland. IEEE. pp. 57-62