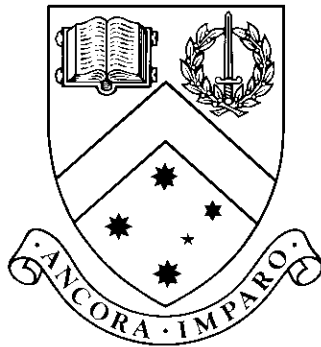


Content-Based Image Retrieval

by

Melissa Anne Yung



Submitted by Melissa Anne Yung (18993729)

in partial fulfillment of the Requirements for the Degree of
Bachelor of Software Engineering with Honours (2770)

Supervisor: Dr. Sid Ray

Clayton School of Information Technology
Monash University

2006

Abstract

There has been a great amount of research work done in the Content-Based Image Retrieval field, the majority of which covers issues such as Feature Representation, User Relevance Feedback and Indexing Structures. There has not been much investigation of the sample size and its effects on the accuracy of a retrieval system. The following report presents a CBIR system that has been designed and developed for this purpose. The effects with varying scope and number of irrelevant semantic classes are illustrated, underlining the role of these factors in CBIR. An evaluation of appropriate performance measures is also covered to enable a consistent representation for CBIR systems. This is essential to enable a true comparison on similar grounds.

Contents

1	Introduction	7
2	Background	9
2.1	Feature Weighting	10
2.2	Relevance Feedback	11
2.3	Sample Size Issue	12
2.3.1	Dimensionality	13
2.3.2	Scope	13
2.3.3	Database Size	13
3	Feature Effectiveness	15
3.1	Colour Features	16
3.2	Shape Features	17
4	Feature Combining	19
4.1	Weighting Schemes	19
4.2	Sum-Result Indexing Algorithm	20
5	Similarity Function	22
5.1	Normalization	22
6	Performance Measures	24
6.1	Contingency Tables	24
6.2	Normalized Average Rank	25
6.3	Improved Performance Measures	26
6.3.1	Appropriate Averaging	26
6.3.2	Adjusted Precision vs Relevant Scope Graphs	26
6.3.3	Precedence Indication Measure	27
7	Experiments and Results	28
7.1	Combining Features	28
7.1.1	Weights Assignments	29
7.1.2	Weights Updating	29
7.2	Sample Size	31

7.3	User Interaction	35
7.3.1	Query Refinement	35
8	Conclusion	38
9	Future Work	39
A	Screen Shot of User Interface	40
B	Experimental Results with Weights	41
C	Experimental Results with Feature Combination	43
	Bibliography	43

List of Figures

2.1	A CBIR System	10
7.1	Feature Combination Graph	29
7.2	Weights Contribution Graph	30
7.3	Query Image A	30
7.4	Retrieved Images with Calculated Weights	31
7.5	Retrieved Images with User-Defined Weights	31
7.6	Adjusted Precision-Relevant Scope Graph for varying Embedding Ratios	32
7.7	Contingency Table with Embedding Ratio = 0.5	33
7.8	Contingency Table with Embedding Ratio = 0.75	33
7.9	Recall-FPr Graph for two Embedding Ratios	34
7.10	Query Image B	35
7.11	Retrieved Images without refinements	36
7.12	Retrieved Images after 1 refinement	36
7.13	Retrieved Images after 2 refinements	37
A.1	Screen Shot of Developed CBIR System	40
B.1	Query Image C	41
B.2	Results without Weights	41
B.3	Results with Weights	42
C.1	Query Image D	43
C.2	Results with 3D-Colour Feature	44
C.3	Results with 7D-Shape Feature	44
C.4	Results with combined 3D-Colour and 7D-Shape using SRI algorithm	45

Declaration of originality

I, Melissa Yung, declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Melissa Yung
7th November 2006

Acknowledgements

I would like to thank Dr. Sid Ray for his help and invaluable patience throughout the year. This thesis would not have been possible without his supervision. I am also grateful to all those who have supported and encouraged me.

Melissa Yung

Chapter 1

Introduction

Content-Based Image Retrieval, commonly referred to as CBIR, is the automatic retrieval of digital images from large databases. This technique makes use of the inherent visual contents of an image to perform a query. As opposed to earlier image retrieval methods which involved the manual textual annotations of images, CBIR systems identify the images by automatically-extracted syntactical features. With the advance in technology, including the ever-increasing popularity of digital cameras and the possibility to manage and store large databases of information, CBIR proves to be much more efficient and practical. It relieves the user from the previous cumbersome, subjective and error-prone task of image description and has therefore dramatically improved the usability of the system.

However, there are obvious limitations in such a metadata-based system. The removal of human interaction results in a number of issues such as the ability to deal with semantic attributes of pictures. Machines are unable to accurately extract all the features perceived by humans. This semantic gap is the driving force behind studies and experiments.

There are numerous possibilities that have been investigated to remedy this situation. In this project, further attempts to improve the accuracy by exploring sample size issues, were made possible. This entailed the design and development of a convenient system which allows users to have their say in the retrieval process. The computational results have been recorded and analysed to provide some guidance for more efficient design of CBIR systems. It is an addition to the wide range of existing systems that explore retrieval methods. A selection including the PicHunter, MARS, Query By Image Content (QBIC), Blobworld, Virage and Chabot were helpful for the completion of this report [1] [3] [7] [18] [27].

As CBIR techniques, tools and algorithms used originate from various fields

such as Pattern Recognition, Statistics and Computer Vision, they attract the interest of many researchers. It is a growing area of research with a number of unsolved challenges. CBIR has vast and diverse application possibilities that range from art galleries and biomedical research to weather forecasting and fabric design.

Chapter 2 provides some background information including various techniques used to improve CBIR accuracy. Chapter 3 introduces the colour and shape features used in the system. Chapter 4 discusses the approach used to combine those descriptors while Chapter 5 describes the similarity function and normalization required. Chapter 6 is an overview of the performance measures used for evaluation as well as some new alternative ones. Chapter 7 presents the experimental results and findings of the project. Chapter 8 and 9 finally give a conclusion and directions for the future. Additional query results are illustrated in the appendices.

Chapter 2

Background

A typical CBIR system makes use of low-level features such as colour, texture and shape to describe images. Given a database of images, it performs feature extraction on each image and indexes them accordingly. For any subsequent image query from the user, the system compares those features and outputs a list of relevant database images.

Figure 2.1 shows the steps involved in

- populating the database by extracting and storing features for images I_1 to I_N and
- processing an image query by extracting features from the query image Q specified by the user and performing a similarity comparison for retrieval from the database

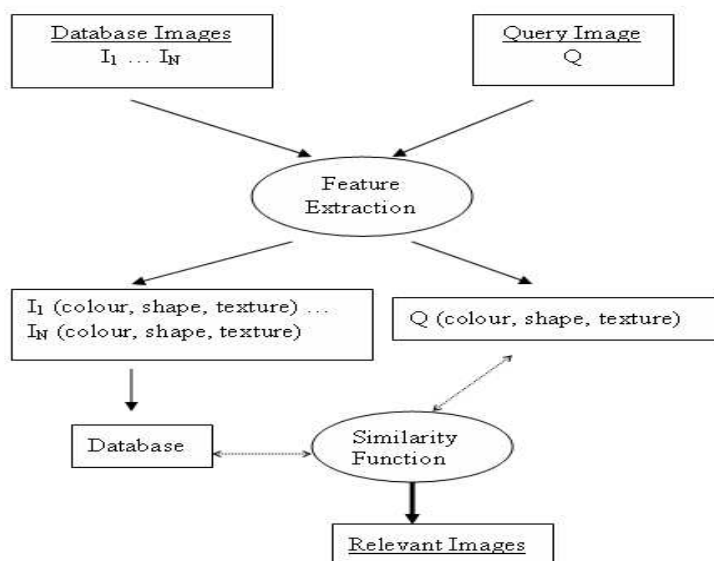


Figure 2.1: A CBIR System

Some improvements to this basic system include the use of

- Feature Weighting [10]
- User Relevance Feedback [24]
- Different number of features also known as Dimensionality with various number of retrieved images known as Scope [2] [6] [19]
- Optimal number of images in the database broadly referred to as the Sample Size [9] [15] [19]

2.1 Feature Weighting

In practice, a typical system makes use of multiple features to represent an image. Due to the individual characteristics of each feature, the separation of classes might differ. Hence, it is justified to apply different weights to different features in accordance with the amount of discriminatory characteristics they carry.

It has been discussed that a good feature component is one with values producing a large σ over the entire database images but a small σ within the relevant images, where σ is the standard deviation or the amount of variation between the values. This means that by making use of this feature, the system will be able to easily distinguish between the different classes of

images, as the feature should appropriately distinguish between the separate categories since all the images from the same category will have a small σ within them.

Hore and Ray [10] have already explored four weighting calculations and concluded that the one below does indeed provide the best performance results.

$$w(x) = \frac{\sigma DF_N(x)}{\sigma DF_{rel}(x)} \quad (2.1)$$

where for a given feature x , $\sigma DF_N(x)$ denotes the standard deviation of the similarity measures over the database with N samples and $\sigma DF_{rel}(x)$ denotes the standard deviation of the similarity measures over the relevant images.

This weight calculation effectively assigns more weight to those features with a large variation in the similarity measures over the entire database and a small variation in the similarity measures over the relevant images present. This scheme requires a set of similarity measures for the query image to be applied before calculating the weight. It thus not only requires the computing of the dissimilarity for each feature but is also dependent on the similarity function used.

2.2 Relevance Feedback

Relevance Feedback(RF) is a technique which incorporates the user in the retrieval process. The two most common ways of using RF in CBIR are to:

- modify the query image based on the true positives returned and
- modify the results by assigning weights showing their relevance

Weights can be calculated from a set of labelled samples with ground truth information and adjusted to reflect a user's specific needs. The initial calculation is computer centric i.e. solely based on computed figures and how their values are distributed. The user-defined weights run the risk of producing adverse effects as what a user might perceive to be prominent in the query might not be the same as what the computer registered. Relevance Feedback remedies this problem by focusing on the high-level features i.e. through the actual image displays, instead of the low-level features. RF thus takes into account the semantics that could not be defined with

the computed low-level features, by incorporating the user’s perception and judgement - a desirable improvement.

Rui and Huang weight updating technique requires the user to mark the images as highly relevant, relevant, no opinion, non-relevant and highly non-relevant [21]. Those five options are a good trade-off between the precision obtained and the burden imposed on the user. They calculate the proportion of relevant images retrieved to the ones defined by the user, for each computed feature. The weights are then readjusted to reflect the importance each feature should really carry when calculating the similarity distances. Their solution is a heuristic-based one which cannot be proven to be the best. For instance, the use of an initial set of no-bias weights could be improved to incorporate Ray and Hore initial weight calculations based on ground truth database. This will of course require some labelled samples to be present as is the case with the developed system.

PicHunter attempts to find a specific target. The precision is based on the number of iterations necessary before a target image is found. Before implementation, experiments were conducted to increase the confidence on the appropriateness of the proposed design. As PicHunter requires the user to denote images as being relevant or not, it is important to ensure that the human similarity judgements can be appropriately reflected. Prior to conducting the experiment, the users were exposed to the database images in order to calibrate their similarity scales and the results showed the possibility of using the distance measure as a similarity measure [5].

PicHunter uses the Bayesian approach to find the probabilities for each database images. With the ability to define a linear function directly connecting the image score (based on how far it is from the query) and the probability of selection for RF, the probability of a particular user’s action can be determined. Therefore for each iteration, the change in probability of a given datum can be computed [3][4]. However PicHunter has not explored this RF technique in other search contexts such as category search which attempts to find a category of relevant images.

As for Li and Yuan, they introduced a novel RF method, where both the update of weights and the query vector change are done simultaneously [16]. This method outperformed the Bayesian approach as well as the techniques from Rui and Huang, mentioned above.

2.3 Sample Size Issue

In CBIR context, the sample size issue covers the following three matters:

- Dimensionality i.e. number of features used for representation
- Scope i.e number of retrieved images for a query
- Database Size i.e. number of samples in the whole database, number of semantic classes in the whole database, number of samples per semantic class

2.3.1 Dimensionality

It is a well documented fact that the accuracy of retrieval results does not necessarily increase as the dimensionality of a feature vector increases. Hughes has demonstrated that, with a finite database size, the accuracy of the results will fall after a maximum peak but then fall after an optimal level is reached [11]. This phenomenon contradicting one's intuitive assumption is known as the Curse of Dimensionality. Performance does not always improve as the number of variables is arbitrarily increased. It may even deteriorate [15].

Das and Ray proposed a compact feature representation and demonstrated that the use of 25-dimensions as opposed to 148-dimensions effectively produced better results [6]. It is important to note that they have experimented with varying scope and that at higher scope, the effect of dimension reduction is more prominent.

2.3.2 Scope

On the other hand, the effects of varying scope have not received as much attention. This is mainly due to the nature of most typical CBIR systems that only display retrieved images to the user. The latter can simply retrieve images until his expectations are met. However, retrieved images can also be used for subsequent tasks. This includes judging the significance through RF and other computationally intensive tasks. The redundant information from those irrelevant images is useless and not desirable. It is thus necessary to illustrate a system's performance under those various conditions.

2.3.3 Database Size

The effects of a varying database size are dependent on the system's machine learning capabilities and the nature of the samples. In a machine learning environment, the CBIR accuracy increases with increasing database size regardless of the additional samples. On the other hand, without machine learning, the nature of the added samples is important.

Hughes [11] showed that the probability of correct classification reduced

as the ratio of sample size to dimensionality decreased. Consequently, a trade-off problem occurred. Kanal and Chandrasekaran [15] pointed out that the number of features to be used depended on the probability structure. As the knowledge of the probability structure increased, more features could be included to increase performance.

Foley [8] compared the actual error rate with the Bayes error rate to investigate the sample size issue in a classifier for a two-class problem. The Bayes error rate being the probability of mis-classification can be obtained by averaging over the distribution x the conditional probability of mis-classifying the object x into its class. The study derived the error rate as a function of the sample size and dimensionality used. Whenever a ratio exceeding three was used, reasonably close error rates were achieved.

Wang, Chan and Tan [29] proposed combining the use of the Euclidean distance with the Support Vector Machines based active learning in CBIR systems to improve accuracy. This did not require much additional computation and yet addressed the small sample size issue. The results were similar to what would have been obtained with an increase to the number of learning samples.

Shahshahani and Landgrebe [25] investigated the small sample size issue in relation to the size of the training sample used for estimations in the classification process. They have remedied the situation of small ratio of training sample size to dimensionality which resulted in poor performance by introducing unlabeled samples. It was observed that this effectively delayed the Hughes phenomenon while decreasing error rate.

In this light, the number of irrelevant semantic classes in a CBIR system is considered for this report. The effects of those samples are overlooked since it is obvious that an increasing number of irrelevant semantic classes can only degrade a system's performance. This is what happens when CBIR systems are implemented in real-life. They inevitably make use of a much larger database with more embeddings i.e. number of irrelevant items. The question of how the performance can be mapped from the research version to a scaled-up version thus arises.

To reflect this expected change in accuracy, the magnitudes of the detrimental effects of increasing the embedding size become relevant. They might be able to predict the performance for a real-life implementation.

Chapter 3

Feature Effectiveness

For the purpose of the present study, the developed CBIR system extracts the colour and shape properties of images as representative features. The experiments are then carried out on two datasets:

Dataset A contains 2000 images with each of the 10 semantic classes below contributing 200 images.

- A.1 Flowers
- A.2 Leaves
- A.3 Faces
- A.4 Views
- A.5 Fish
- A.6 Airplanes
- A.7 Cars
- A.8 Leopards
- A.9 Ships
- A.10 Watches

Dataset B contains 30 images with each of the 6 semantic classes below contributing 5 images.

- B.1 Sunflowers
- B.2 Roses
- B.3 Fish
- B.4 Yellow Cars
- B.5 Grey Cars
- B.6 Watches

All the images have been converted to GIF87 format.

3.1 Colour Features

Colour is one of the major features used in CBIR systems. This popularity is attributed to the ease in implementation and the distinguishing differences between colours. It is a robust feature to changes such as the scene layout or viewing angle. Colour can be represented with different models such as HSI, YIQ, CMY and RGB.

The RGB model is the most widely known one and can be visualized as a cube. One corner of the cube is the origin $L(0, 0, 0)$ and each of the three primary colours Red, Green and Blue are assigned an edge to represent the axis from the origin. Any other individual colour obtained after combining the red, green and blue components in certain proportions then lie in this coordinate space. The origin represents black as it is the point of lowest red, green and blue values. Understandably, the opposite corner with the highest red, green and blue values represents white. The 3D coordinate space is similar to the way our three sets of retinal cones work in our human visual system. The RGB model is nonetheless limited in representing the full human perception which includes details such as the brightness and purity of a colour. Those are however implicit in the coordinate space and the non linear transformation from RGB to HSI is used to capture those additional properties.

The HSI model can be visualized as the cone obtained when the RGB cube is viewed from the origin. The three channels of the HSI space are Hue, Saturation and Intensity. The Intensity axis coincides with the black-white diagonal and denotes the brightness of a colour from its vertical position in the cone. The Saturation is computed as the radius from the Intensity axis for any value and reflects the purity of the colour. The Hue component is the value of the angle with respect to the Red axis for any point in the coordinate space and represents the tone of a colour. The following equations are used for converting from RGB to HSI values:

$$H = \begin{cases} \theta & \text{if } B \leq G \\ 360 + \theta & \text{if } B > G \text{ where} \end{cases} \quad (3.1)$$

$$\theta = \cos^{-1} \left\{ \frac{\frac{1}{2}[(R - G) + (R - B)]}{[(R - G)^2 + (R - B)(G - B)]^{\frac{1}{2}}} \right\},$$

$$S = 1 - \left(\frac{3}{R + G + B} \right) [\min(R, G, B)] \quad (3.2)$$

and

$$I = \frac{1}{3}(R + G + B) \quad (3.3)$$

A difficulty with those equations is encountered with images with a low Saturation value i.e. with points close to the Intensity axis. The closer they are, the harder it is to determine the Hue value. The extreme condition being if Saturation is zero and Hue is undefined. To avoid this problem, the developed CBIR system makes use of only the Red component, Green component, Blue component and Intensity component to derive the colour features for the images.

3.2 Shape Features

Shape is a common low-level feature used to describe the geometric characteristics of images. Hu proposed a set of seven Moment Invariants for this purpose. Those formulas have proved to recognise images after rotation, translation and scaling [12].

The image is described as a two-dimensional non-negative discrete function:

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y) dx dy \quad (3.4)$$

For a digital image, Central moments are defined:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (3.5)$$

where

$$\bar{x} = \frac{m_{10}}{m_{00}} \quad (3.6)$$

and

$$\bar{y} = \frac{m_{01}}{m_{00}} \quad (3.7)$$

Central moments are normalized according to:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \quad (3.8)$$

where

$$\gamma = \frac{p+q}{2} + 1 \quad (3.9)$$

for $p+q = 2, 3 \dots$

After computing the following normalized central moments $\eta_{11}, \eta_{20}, \eta_{02}, \eta_{30}, \eta_{03}, \eta_{21}$ and η_{12} , they are combined to define the Moment Invariants as follows:

$$\Phi_1 = \eta_{20} + \eta_{02} \quad (3.10)$$

$$\Phi_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \quad (3.11)$$

$$\Phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (3.12)$$

$$\Phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (3.13)$$

$$\begin{aligned} \Phi_5 = & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\ & (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[(3\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \quad (3.14)$$

$$\Phi_6 = (\eta_{20} + \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \quad (3.15)$$

$$\begin{aligned} \Phi_7 = & (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\ & (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[(3\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \quad (3.16)$$

The CBIR system extracts four colour features and seven shape features to represent the images. The user can select the feature(s) to use as well as the feature combination methods in the case of multiple features. Retrieval can thus be based on:

- i) Individual features
- ii) Multi-dimensional features
- iii) Combined features

Chapter 4

Feature Combining

A good selection of discriminatory features is essential for an accurate CBIR system. However, this selection will vary depending on the feature relevance in a particular query image. This is why a typical system makes use of multiple features and then assigns corresponding weights to denote their respective importance for a retrieval session. Different weights are applied to different features in accordance with the amount of distinguishing characteristics they carry.

A large number of features can be used to describe an image but due to the small sample size problem mentioned, it will be of no help to the system to arbitrarily increase this number. Besides, the extra computations, the additional amount of data and the increased complexity are not desirable.

4.1 Weighting Schemes

The weight calculation from Hore and Ray [10] was adapted and explored in the developed CBIR system. However, instead of computing the dissimilarities, the initial sets of values were used as the measures. This results in a more representative value for the variation over the database and over the relevant images than the previous similarity measures. The similarity function makes use of absolute values and therefore reduces the range by eliminating all negative values to positive ones. The outcome of the reduced variability is a flawed reflection of the feature's importance. The system is thus alleviated from extra computations while calculating more representative weights for a fast and accurate retrieval process.

Furthermore, if we assume the similarity function to be the signed difference between query features and database features, both methods will give the same results as shown below.

The vector $F_N(x)$ represents the set of values for feature x over the whole database of N samples. The standard deviation of $F_N(x)$ is the variation over the initial feature values of the whole database.

$$F_N(x) = \{I_i(x); i = 1, 2, \dots, N\} \quad (4.1)$$

For any query image, computing the similarity measure will result in a set of dissimilarity values denoted by vector $DF_N(x)$.

$$DF_N(x) = \{I_i(x) - Q(x); i = 1, 2, \dots, N\} \quad (4.2)$$

It can be observed that the standard deviation of $F_N(x)$ = standard deviation of $DF_N(x)$.

The same applies when computing the standard deviation over the relevant images. Therefore, computing the weight on initial values with

$$\text{Method1} = \frac{\sigma F_N(x)}{\sigma F_{rel}(x)} \quad (4.3)$$

is equivalent to computing the weight on similarity measures

$$\text{Method2} = \frac{\sigma DF_N(x)}{\sigma DF_{rel}(x)} \quad (4.4)$$

4.2 Sum-Result Indexing Algorithm

The probability of an event A given events x_1, \dots, x_n , where n is the number of features such that $(A = x_1 \wedge \dots \wedge x_n)$, can be determined by

$$Pr(A) = \prod_{i=1}^n p(x_i)(1 - p(x_i)) \quad (4.5)$$

which combines the probabilities for all n features.

A problem occurs when a feature has a probability of zero, due to its absence from the image. This will produce a zero result regardless of how well the remaining features match.

The SRI algorithm overcomes this problem by summing the individual features similarity measures [10]. The SRI vector is thus adopted for the system and represents the combination of how well each feature has performed. This

is illustrated below:

$$SRI = \sum_{i=1}^n DF_N(x_i) \quad (4.6)$$

After sorting the SRI vector, the most relevant images are retrieved. Convincing results have been obtained and the reader is referred to the paper by Hore and Ray for further details [10].

Chapter 5

Similarity Function

In order to determine the similarity between two images, the weighted Minkowski-Form Distance commonly known as the weighted Euclidean Distance is used.

It is defined as:

$$D = \sqrt{\sum_{i=1}^n w(x_i) (I(x_i) - Q(x_i))^2} \quad (5.1)$$

where,

x_i represents the i th feature component,

$w(x_i)$ represents the weight for the feature,

n represents the number of features,

$I(x_i)$ represents the database image and

$Q(x_i)$ represents the query image.

A problem arising from this metric is its subjectivity to the magnitude of the feature values. For instance a feature ranging over a larger scale will contribute more to the dissimilarity measure than one with a smaller scale thereby corrupting the representative value. This is solved through normalization steps.

5.1 Normalization

Intra-normalization refers to the readjusting of values within a particular feature vector. This ensures that the values are more appropriately represented, based on the proportional magnitude relative to values in their sets. Normalizing all the feature measurements result in values lying in the same dynamic range [0,1] thus removing the previously bias similarity measure.

The Gaussian normalization is used for the normalization process as follows:

$$f' = \frac{f - \mu}{3\sigma} \quad (5.2)$$

By using the 3rd rule, 99% of the values are conserved and will lie within $[-1,1]$. The remaining values are then either set to -1 or 1. To map the values from $[-1,1]$ to $[0,1]$, the following equation is used:

$$f'' = \frac{f' + 1}{2} \quad (5.3)$$

The same method is used for the intra-normalization of the calculated weights within their respective sets (colour or shape). There is also the need to normalize when combining different features as the number of features used might differ from set to set. In the developed system, an additional inter-normalization of weights is required across the different sets.

The weight vector for all n features in set C is denoted by

$$C = \{w(x_i); i = 1, \dots, n\} \quad (5.4)$$

Each weight is divided by the vector sum for normalization.

$$C' = \left\{ \frac{w(x_i)}{sum}; i = 1, \dots, n \right\} \quad (5.5)$$

where,

$$sum = \sum_{i=1}^n w(x_i) \quad (5.6)$$

Chapter 6

Performance Measures

Precision and Recall values are extensively used as performance measures for CBIR systems.

$$Precision = \frac{Number\ of\ Relevant\ Retrieved\ Images}{Number\ of\ Retrieved\ Images} \quad (6.1)$$

$$Recall = \frac{Number\ of\ Relevant\ Retrieved\ Images}{Number\ of\ Existing\ Relevant\ Images} \quad (6.2)$$

Hyperbolic Precision-Recall graphs can thus be plotted.

6.1 Contingency Tables

In order to gain more insights, the contingency table as an alternative analysis option is explored. The scope value and embedding size are explicitly shown.

	+P	-P	TOTAL
+R	TP	FP	<i>Scope</i>
-R	FN	TN	
TOTAL		Embedding	N

where,

N represents the database size

+R represents all existing Relevant images

-R represents all existing Irrelevant images

+P represents all Retrieved images

-P represents all Non-Retrieved images

TP represents the True Positives i.e. Relevant Retrieved Images
 FP represents the False Positives i.e. Irrelevant but Retrieved Images

FN represents the False Negatives i.e. Relevant but Non-Retrieved Images
 TN represents the True Negatives i.e. Irrelevant and Non-Retrieved Images

Even in the proposed evaluation techniques [22], the embedding effects are not illustrated. The TN value is completely disregarded although it represents the number of Irrelevant and Non-Retrieved Images. The relevance of such measure is trivial when evaluating the accuracy with a constant embedding size but this assumption does not always hold.

Huijsmans and Sebe have recently highlighted this shortcoming of the Precision-Recall graphs and introduced a new GRnP graph (Generality-Recall=nPrecision Graph) which includes generality information [14]. Generality is a measure of the density of relevant images in the search.

The symbol n in GRnP graphs denotes the relevant scope where,

$$RelevantScope = \frac{Scope}{RelevantClassSize} \quad (6.3)$$

When the scope value is coupled with the relevant class size, i.e. $n = 1$, recall and precision values are similar, therefore enabling a 2D-illustration of Generality vs Recall=Precision plane. The reader is referred to Huijsmans and Sebe paper for further details and illustrations.

6.2 Normalized Average Rank

The results of a retrieval session in CBIR are not only a set, but also a sequence of ranked images based on their similarity measures. Precision and Recall measures fail to take into account the ordering of the retrieved images.

The Normalized Average Rank has thus been introduced [22].

$$Rank = \frac{1}{NN_R} \left\{ \sum_{i=1}^{N_R} R_i - \frac{N_R(N_R - 1)}{2} \right\} \quad (6.4)$$

where,

R_i represents the rank of the i th relevant retrieved image,

N represents the database size and

N_R represents the number of relevant images.

6.3 Improved Performance Measures

As illustrated previously, it is necessary for performance measures to be meaningful, reliable and objective. They are required to allow comparison between different systems.

The following is proposed for a full illustration of the system's performance.

6.3.1 Appropriate Averaging

For a true representation of the measures obtained, experiments are conducted over each of the database images and results averaged.

The coefficient of variation is a measure of the dispersion of a distribution. It can be used regardless of the system's classification algorithm and gives a degree of confidence in the results obtained. It is calculated for each sets of results and defined as:

$$C_v = \frac{\textit{StandardDeviation}}{\textit{Mean}} \quad (6.5)$$

Also, Radial Averaging is suggested as the best averaging method compared to precision-box averaging or recall-box averaging for Precision-Recall graphs [13] [14]. The idea of radial averaging is to restrict the values to be averaged to those obtained over a constant relevant scope. Lines radiating from the origin with gradients set to the relevant scopes are graphed and only those values from the Precision-Recall curves lying on those lines, are averaged.

This is taken care of by making use of Precision vs Relevant Scope graphs. All the precision values are thus averaged over constant scopes.

6.3.2 Adjusted Precision vs Relevant Scope Graphs

Although the GRnP graph described above takes into account the generality measure, an instance of the graph is restricted to one relevant scope value (n). It does not allow for the comparison of precision with varying scopes.

An Adjusted Precision vs Relevant Scope graph is thus used with a number of curves to represent different generality measures.

The Adjusted Precision value is needed for comparing the curves. It attempts to discard the expected change in precision, due to the reduced a priori probability of the relevant class with more embeddings.

Adjusted Precision value is defined as:

$$\text{InitialPrecision} - \text{ExpectedChangeInPrecision} \quad (6.6)$$

Adjusted Precision vs Relevant Scope graphs incorporate the generality measure and average over a constant scope. By clearly indicating the scopes used, the graphs further restrict misleading information on systems' performance. Using a small scope to obtain a high precision value will now be evident.

6.3.3 Precedence Indication Measure

A Precedence Indication measure based on the rank of the retrieved images is used side by side with the Adjusted Precision vs Relevant Scope graph for further discriminating between those closely related curves.

The Precedence Indication Measure is then computed as:

$$\text{PrecedenceIndication} = \sum_{i=1}^r W(I_i) \quad (6.7)$$

and $W(I_i)$, which is the weight assigned to each image based on its rank, is:

$$W(I_i) = 1 - \frac{(i-1)}{r} \quad (6.8)$$

for $i = 1 \dots r$ where,

I_i represents the i th retrieved image and

r represents the number of retrieved images.

As opposed to the Normalized Average Rank, the ranks do not range from 1 to the database size. They all lie within $[0,1]$ therefore limiting the biased measurement in the presence of outliers.

Chapter 7

Experiments and Results

7.1 Combining Features

Aim: To investigate ways of exploiting the features for higher accuracy.

Each database image was used in turn as a query with the scope set as 10, 20, 40, 60, 80, 100 and 200. After performing all the retrievals, the results were averaged. The accuracy for various feature combinations with an increasing scope on Dataset A is illustrated in Figure 7.1.

Results from using the single feature 2D-Shape Feature or 7D-Shape Feature are represented by the two lowest performing curves. There was a clear improvement when extending the dimensionality to combine the 3D-Colour Feature for retrieval. The highest performing curve was obtained with the combination of the 3D RGB Colour Feature and 7D Moment Invariants Shape Feature using the SRI algorithm.

However, increasing the dimensionality with features that poorly discriminate had adverse effects. The single 3D-Colour Feature outperformed the combination of 3D-Colour Feature and 2D-Shape Feature. The feature strength is an important factor and was taken into consideration through weights incorporation.

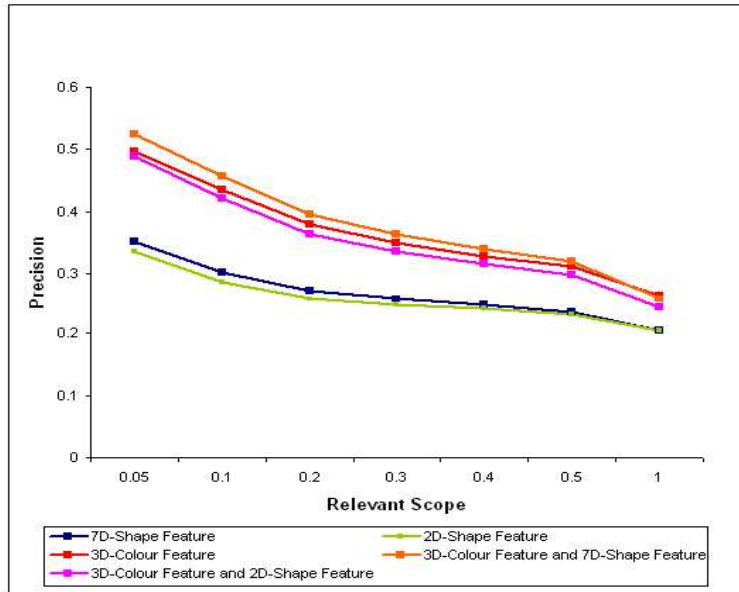


Figure 7.1: Feature Combination Graph

7.1.1 Weights Assignments

To accommodate for those varying features discriminatory capabilities, the two weighting schemes Method 1 and Method 2, as defined in Eq. (4.3) and Eq. (4.4) respectively, were used. Method 2 makes use of the dissimilarity measure, which is the absolute difference in the developed system and therefore the computed weights were expected to be different. Method 1 proved to be the better choice despite the limited magnitude of the precision increase. The expected reduction in variability, as discussed earlier, did not significantly improve the results but since Method 1 required much fewer computations, it was an improvement.

The same experiment was repeated with Dataset B. The Precedence Indication was computed and Method 1 outperformed Method 2 by 0.11, further reinforcing the Method 1 superiority.

Figure 7.2 illustrates the improvement in precision when weights are applied. The weakness of the features used limited the improvement in precision, but it was seen that weights application generates better results.

7.1.2 Weights Updating

The developed CBIR system also allows for adjustment of weights according to one's judgement of significant features. As observed earlier, these could

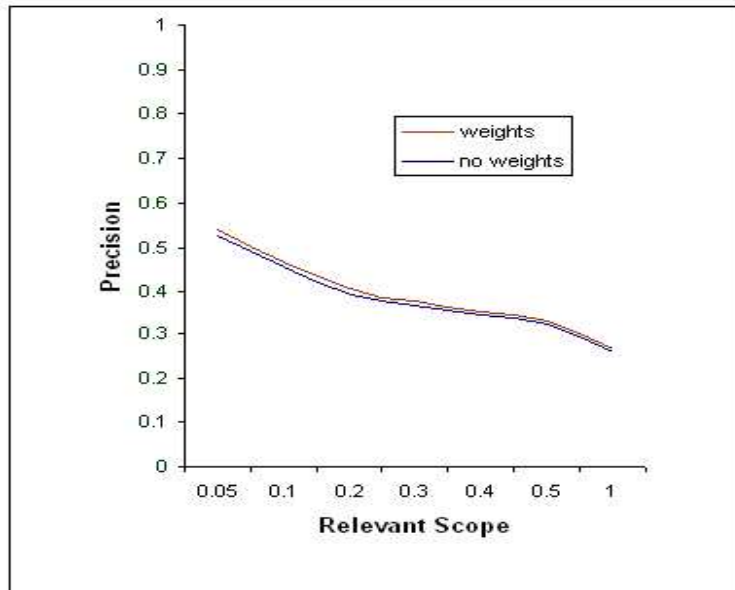


Figure 7.2: Weights Contribution Graph

have adverse effects. But through combining them with calculated weights, it has proved successful.

For the particular Query Image A shown in Figure 7.3, the weights were re-adjusted to give more importance to the Green Component. Those user-defined weights were then used in conjunction with the pre-defined weights to refine the search. Figure 7.4 shows the results with only the pre-defined weights while Figure 7.5 shows the refined search incorporating the adjustments by the user. The system retrieved images according to the user's preferences. This attribute is useful to find particular kinds of images, for which no exact query can be found. Searching can be performed on an image that is somewhat similar.



Figure 7.3: Query Image A

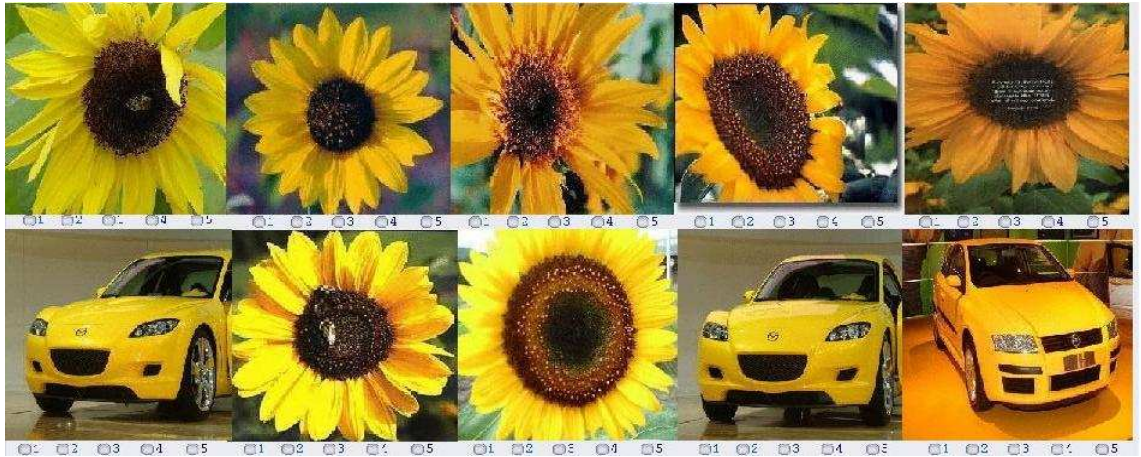


Figure 7.4: Retrieved Images with Calculated Weights



Figure 7.5: Retrieved Images with User-Defined Weights

7.2 Sample Size

Aim: To investigate the increase in scope and number of irrelevant semantic classes (embedding sizes) and their effects on accuracy.

Each database image was used in turn as a query, with the scope set as 10, 20, 40, 60, 80, 100 and 200. In addition, the number of irrelevant semantic classes present was increased from 1 to 9. This is represented as the embedding ratio e . After performing all the retrievals with the best feature combination, the results were averaged.

To illustrate the change in accuracy with varying scope across the embed-

ding sizes, it was necessary to discard the accuracy change due to different a priori probabilities. The intuitive approach defined for adjustment, refer to Eq. (6.6), was not appropriate. This was due to the weakness of the features discrimination relative to the difference in a priori probabilities. At low generality values, the expected change in a priori probabilities was so large that it exceeded the actual accuracy change. This is illustrated in Figure 7.6. There was an unexpected increase in the adjusted precision when the generality increased from 0.50 to 0.67. The increase in precision with a bigger embedding size is wrong and only occurred because the difference in a priori probabilities is large. It is not a true representation of the system's performance and precision could not be related to e as a defined function.

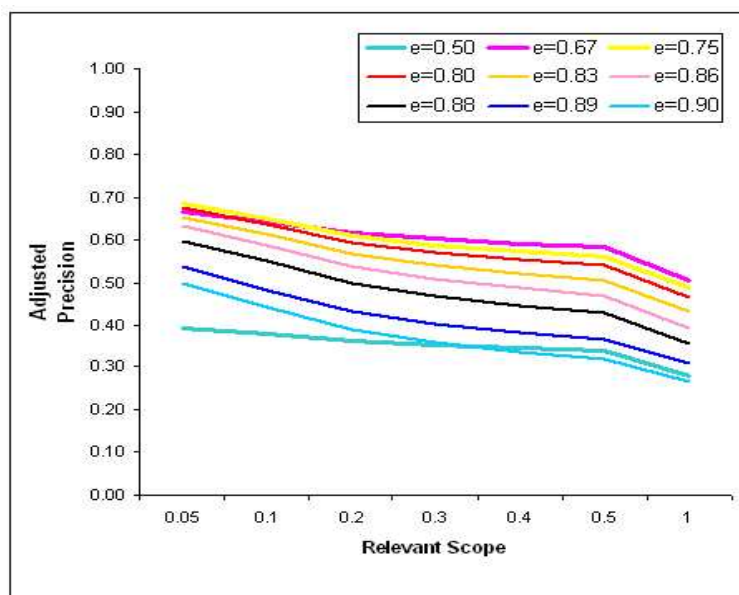


Figure 7.6: Adjusted Precision-Relevant Scope Graph for varying Embedding Ratios

However, it was possible to note that the change in precision was more prominent at lower scopes. Although the change in scope from 100 to 200 was by a magnitude of 100, the change in precision was still lower than for a change of 10 at lower scope. The increase in precision was considered significant until the relevant scope reached 0.2. Images retrieved are therefore rather similar around this scope since changes generated somewhat similar precision values. It was concluded that the ideal relevant scope is 0.125 as it will lie halfway between the value with the most prominent change i.e. 0.1 and the value where saturation was observed i.e. 0.2. The relevant images are more closely clustered in the feature space around the query and thus lead to a higher increase in the number of relevant samples. There are

no benefits from retrieving more samples if the accuracy does not differ by much. This rate of change in accuracy indicated a possible cut-off point, beyond which, no significant improvement is noted.

Alternative ROC analysis were also introduced, where

$$Recall = \frac{TruePositives}{RelevantClassSize} \quad (7.1)$$

and

$$FalsePositivesRate(FPr) = \frac{FalsePositives}{Embeddings} \quad (7.2)$$

Those measures are useful as they reflect the embeddings and enable comparison with varying scopes and number of irrelevant semantic classes.

Consider the following subset of results obtained with varying embedding sizes each with varying scopes, in Figure 7.7, Figure 7.8 and corresponding Table 7.1 and Table 7.2 respectively. All the experiments were conducted on the full database size and the results averaged. The coefficient of variation indicated that, the reliability of the results decreased as more samples were retrieved and as more irrelevant classes were added. This is to be expected as the risks of mixed results increases with more data. They were all however reasonable values.

	+P	-P	TOTAL			+P	-P	TOTAL			+P	-P	TOTAL
+R	9	1	10		+R	18	2	20		+R	34	6	40
-R	191	199	390		-R	182	198	380		-R	166	194	360
TOTAL	200	<u>200</u>	400		TOTAL	200	<u>200</u>	400		TOTAL	200	<u>200</u>	400

Figure 7.7: Contingency Table with Embedding Ratio = 0.5

	+P	-P	TOTAL			+P	-P	TOTAL			+P	-P	TOTAL
+R	8	2	10		+R	15	5	20		+R	28	12	40
-R	192	598	790		-R	185	595	780		-R	172	588	760
TOTAL	200	<u>600</u>	800		TOTAL	200	<u>600</u>	800		TOTAL	200	<u>600</u>	800

Figure 7.8: Contingency Table with Embedding Ratio = 0.75

Table 7.1: Embedding Ratio = 0.5

Scope	Recall	FPr	C_v
10	0.045	0.005	0.052581
20	0.091	0.01	0.062983
30	0.17	0.03	0.07415

Table 7.2: Embedding Ratio = 0.75

Scope	Recall	FPr	C_v
10	0.04	0.0033	0.0695022
20	0.075	0.0083	0.0829299
30	0.14	0.02	0.0990383

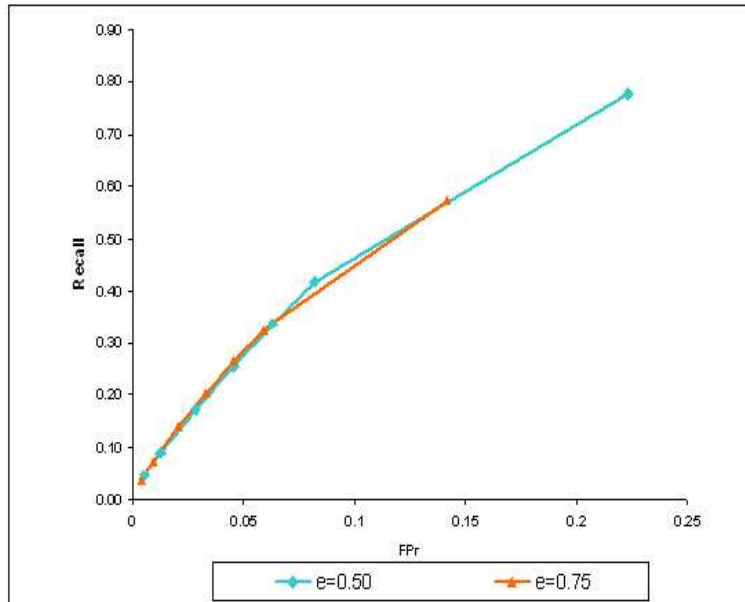


Figure 7.9: Recall-FPr Graph for two Embedding Ratios

The area under each curve in Figure 7.9 is a measure of the ability to correctly classify the samples. The intersection points of the curves indicate at which scope point, the same Recall rate can be achieved. As a larger number of irrelevant classes are used, a larger scope value is required before reaching a similar rate with fewer irrelevant classes.

For a given embedding size, as the scope increases, both the Recall and FPr increased. This is expected as the number of positive and negative images can only remain constant or increase. The trade-off between the two values is governed by the scope used.

7.3 User Interaction

In a further attempt to improve the retrieval accuracy, the developed CBIR system allows the user to provide feedback.

7.3.1 Query Refinement

The developed CBIR system enables refining search results. The user can rate the retrieved images on a scale of 1 to 5 where,

1: Highly Irrelevant

2: Irrelevant

3: Neutral

4: Relevant

5: Highly Relevant

The query is then refined to incorporate the features from the relevant images labelled by the user.



Figure 7.10: Query Image B

The search refinement for the Query Image B in Figure 7.10, only retrieved additional relevant images after the first few user feedback as shown in Figure 7.12 and Figure 7.13. The subsequent refinements had a negative impact on the retrieval process. This statistical phenomenon can be explained by the presence of subclasses within semantic classes.

Although the high-level features will represent a semantic class as one cluster, the low-level features will not. For instance, flowers are made up of different sub-groups and form multiple clusters in the feature space.

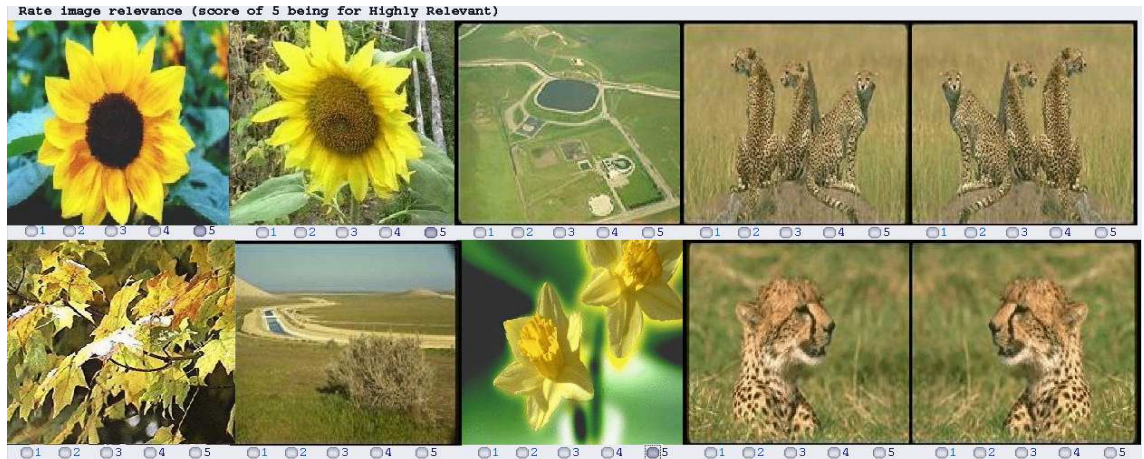


Figure 7.11: Retrieved Images without refinements

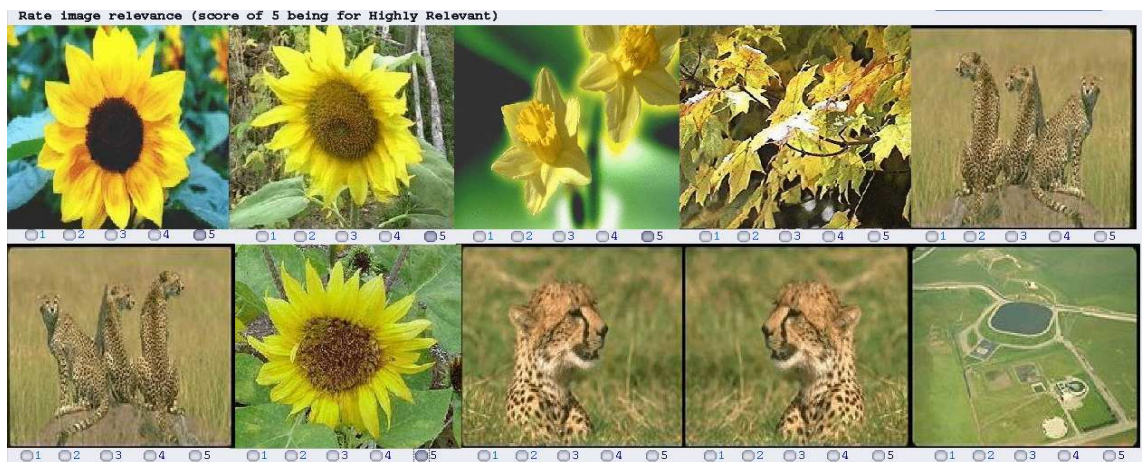


Figure 7.12: Retrieved Images after 1 refinement

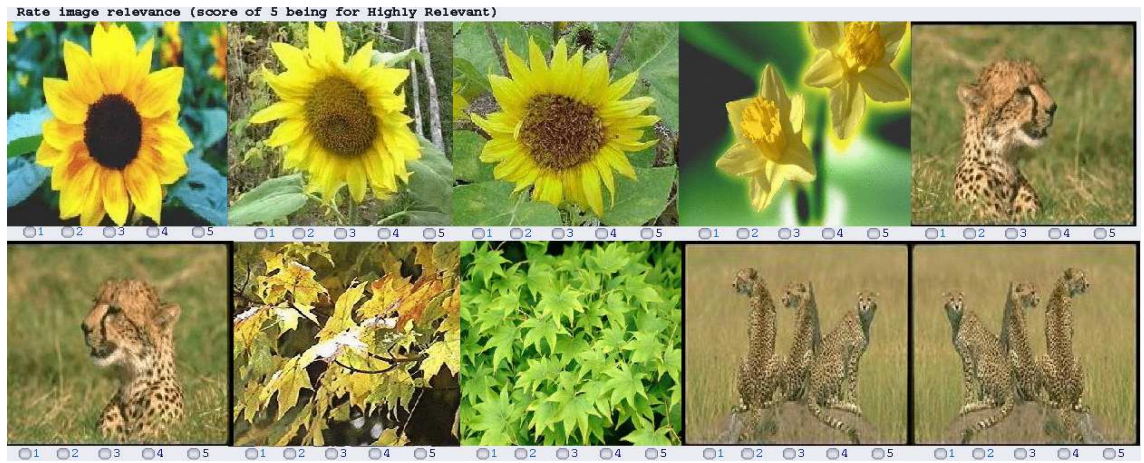


Figure 7.13: Retrieved Images after 2 refinements

Chapter 8

Conclusion

In this report, the change in accuracy in CBIR retrieval was investigated. A CBIR system was designed and developed to enable experiments to be carried out. It can also be put to actual use alongside the existing library of CBIR systems with minor changes.

The effects of dimensionality, feature weighting and relevance feedback in the retrieval process were shown and proved successful. More importantly, the effects of scope and number of irrelevant semantic classes were demonstrated, to underline those essential, yet usually overlooked factors. It was observed that the accuracy increased with increasing scope. Furthermore, this improvement had a tendency to saturate at a certain value. This cut-off point can be used in a retrieval process to limit the number of irrelevant items. Additionally, it was shown that the accuracy decreased with an increasing number of irrelevant semantic classes. This degradation in performance was most prominent in the presence of fewer irrelevant semantic classes and at lower scopes. The change in precision was converging towards a constant rate although further investigations are necessary before formulating this function. Nevertheless, scope and number of irrelevant semantic classes do play a significant role in the retrieval process and should not be ignored for a classifier's design.

For our investigation, it was necessary to survey performance evaluation techniques. It was noted that those sample size factors are often omitted in CBIR performance measures. Their proven impact emphasize the need to illustrate them for a proper representation of a system. This can be overcome by more in-depth analysis of results beyond the standard measures used as well as normalization procedures with respect to scope and number of irrelevant semantic classes.

Chapter 9

Future Work

CBIR is a vast research area and has many open questions and challenges. Designing a CBIR system involves choosing particular feature representation techniques, optimal dimensionality and reliable similarity functions in order to achieve best results. The ultimate aim is to reduce the gap between semantic information in the image and the extracted low-level features.

The developed CBIR system can be extended to include stronger features and additional learning capabilities. This will provide higher accuracy values thus facilitating the investigation of results. A larger database can also be used to increase confidence in the results obtained. Furthermore, experiments can be run on a different data set for more rigorous proof of concept. A function, determining precision from a specified embedding ratio, is yet to be formulated. This is desirable as it will enable the performance of a scaled-up version to be predicted. Investigations of the experimental results are further required for additional insights into sample size issues.

Moreover, this report is a starting point for determining a possible set of metrics for appropriate and consistent performance evaluation. This is crucial to enable comparison between CBIR systems on similar grounds.

Appendix A

Screen Shot of User Interface

The code was written in C++ with the FLTK and FLUID toolkit used for the GUI component. The user interface allows the database initialisation, individual queries, weights adjustments, search refinements, scope specification and other experimental options as required. The developed system is shown in Figure A.1

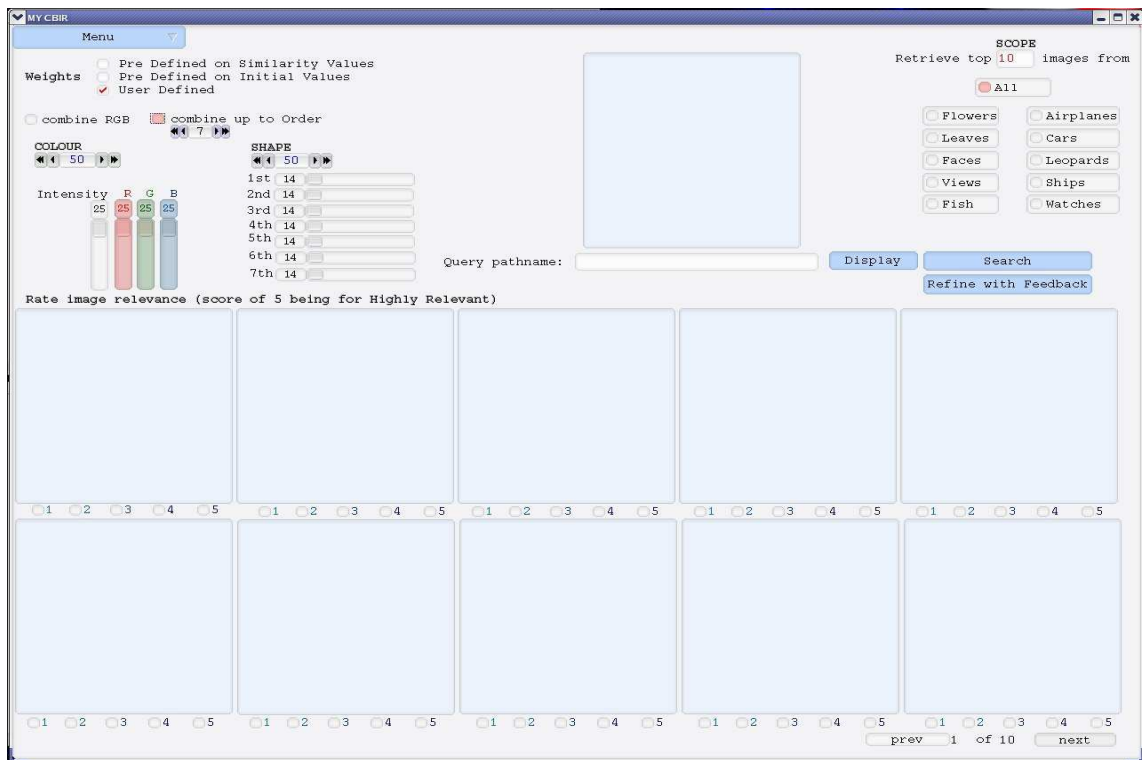


Figure A.1: Screen Shot of Developed CBIR System

Appendix B

Experimental Results with Weights

For the Query Image C from Figure B.1, the results obtained without weights are illustrated in Figure B.2 and the improved ones with weights are shown in Figure B.3.

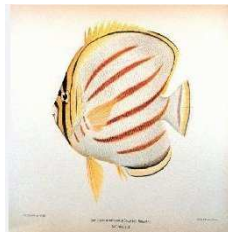


Figure B.1: Query Image C



Figure B.2: Results without Weights



Figure B.3: Results with Weights

Appendix C

Experimental Results with Feature Combination

For the Query Image D from Figure C.1, the results obtained with

- 3D-Colour Feature are illustrated in Figure C.2,
- 7D-Shape Feature are illustrated in Figure C.3 and
- with both combined using the SRI algorithm are illustrated in Figure C.4.



Figure C.1: Query Image D



Figure C.2: Results with 3D-Colour Feature



Figure C.3: Results with 7D-Shape Feature



Figure C.4: Results with combined 3D-Colour and 7D-Shape using SRI algorithm

Bibliography

- [1] Carson C., Belongie S., Greenspan H. and Malik J. Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, Aug. 2002.
- [2] Clements S. *Web Document Analysis*, Unpublished BSE Honours Thesis, SCSSE, Monash University, Clayton, Victoria, Australia, 2003.
- [3] Cox I., Miller M., Omohundro S. and Yianilos P. Pichunter: Bayesian Relevance Feedback for Image Retrieval, *International Conference on Pattern Recognition*, vol. 3, pp. 362-369, 1996.
- [4] Cox I., Miller M., Omohundro S. and Yianilos P. Target Testing and the PicHunter Bayesian Multimedia Retrieval System, *Proceedings of ADL '96*, NEC Research Institute, Princeton, 1996.
- [5] Cox I., Miller M., Minka T., Papathornas T. and Yianilos P. The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments, *IEEE Transactions on Image Processing*, vol.9, no. 1, pp. 20-37, 2000.
- [6] Das G. and Ray S. A Compact Feature Representation and Image Indexing in Content-Based Image Retrieval, *Proceedings of Image and Vision Computing*, New Zealand, Caine, Dunedin, Nov. 28-29, 2005.
- [7] Flickner M., Sawhney H., Niblack W., et al. Query By Image and Video Content: The QBIC System, *IEEE Computer*, pp. 23-31, Sept. 1995.
- [8] Foley D. Considerations of Sample and Feature Size, *IEEE Transactions of Information Theory*, vol. IT-18, no.5, Sept. 1972.
- [9] Fukunaga K. and Hayes R. Effects of Sample Size in Classifier Design, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 8, Aug. 1989.
- [10] Hore E. and Ray S. A Sum-Result Indexing Algorithm for Feature Combining in Content-Based Image Retrieval, *Proceedings of the Fourth*

- IASTED International Conference*, SIP. Kaua'i Hawaii, USA. Aug. 12-14, 2002.
- [11] Hughes G. On the Mean Accuracy of Statistical Pattern Recognizers, *IEEE Transactions on Information Theory*, vol. IT-14, pp. 55-63, 1968.
 - [12] Hu M. Visual Pattern Recognition by Moment Invariants, *Computer Methods in Image Analysis, IRE Transactions on Information Theory* vol. 8, 1968.
 - [13] Huijsmans D. and Sebe N. Content-Based Indexing Performance: A Class Size Normalized Precision, Recall, Generality Evaluation, *Proceeding of the IEEE International Conference Image Processing, ICIP .03*, pp. 733-736, 2003.
 - [14] Huijsmans D. and Sebe N. How to Complete Performance Graphs in Content-Based Image Retrieval: Add Generality and Normalize Scope, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.27, pp.245-251, Feb. 2005.
 - [15] Kanal L. and Chandrasekaran B. On Dimensionality and Sample Size in Statistical Pattern Classification, *Pattern Recognition*, vol. 3, pp. 225-234, 1971.
 - [16] Li B. and Yuan S. A Novel Relevance Feedback Method in Content-Based Image Retrieval, *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC.04)*, Department of Computer Science, Jilin University, Changchun, China, 2004.
 - [17] Long F., Zhang H. and Feng D. Fundamentals of Content-Based Image Retrieval. In D. Feng, W. Sie and H. Z. (Eds.), editors, *Multimedia Information Retrieval and Management-Technological Fundamentals and Applications*. Springer, 2003.
 - [18] Ogle V. E. and Stonebraker M. Chabot: Retrieval from a Relational Database of Images, *IEEE Computer*, vol. 28, no. 9, pp. 40-48, Sept. 1995.
 - [19] Raudys S.J. and Jain A.K. Small Sample Size Effects in Statistical Pattern Recognition: Recommendation for Practitioners, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252-262, 1991.
 - [20] Rui Y., Huang T., Ortega M. and Mehrotra S. Content-Based Image Retrieval With Relevance Feedback in MARS *Proceedings of the IEEE International Conference on Image Processing*, 1997.

- [21] Rui Y., Huang T., Ortega M. and Mehrotra S. A Relevance Feedback Architecture in Content-Based Multimedia Information Retrieval Systems, *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries, in conjunction with IEEE CVPR .97*, 1997.
- [22] Muller H., Muller W., Squire D., Marchand-Maillet S and Pun T. Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals, *Pattern Recognition Letters*, 22(5), April 2001.
- [23] Muller H., Geissbuhler A. and Marchand-Maillet S. The Truth about Corel-Evaluation in Image Retrieval, *Proceedings of the International Conference on the Challenge of Image and Video Retrieval (CIVR 2002)*, London, England, July 2002.
- [24] Rui Y., Huang T., Ortega M. and Mehrotra S. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval, *IEEE Transactions on Circuits and Video Technology*, vol. 8, no. 5, pp. 644-655, Sept. 1998.
- [25] Shahshahani B. and Landgrebe D. The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, Sept. 1994.
- [26] Smeulders A. and Gupta A. Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, Dec. 2000.
- [27] Srinivasan B. and Ray S. Content Based Image Retrieval, *SIP 2000 Conference*, Las Vegas, USA. 2000.
- [28] Vasconcelos N. and Kunt M. Content-Based Retrieval from Image Databases: Current Solutions and Future Directions, *Proceedings of International Conference in Image Processing*, 2001.
- [29] Wang L., Chan K. and Tan P. Image Retrieval with SVM Active Learning Embedding Euclidean Search, *Proceedings of International Conference in Image Processing*, pp. 725-728, 2003.
- [30] Zhu X., Fan H., Luo H and Hacid M. Using Small Samples for Content-Based Image Retrieval System with Relevance Feedback, *ACM Multimedia Workshop on Multimedia Information Retrieval*, MIR 2002, Juan Les Pins, France, Dec. 2002.