

AUTOMATED FEATURE EXTRACTION AND CONTENT-BASED
RETRIEVAL OF
PATHOLOGY MICROSCOPIC IMAGES USING K-MEANS
CLUSTERING AND CODE RUN-LENGTH PROBABILITY DISTRIBUTION

by

Lei Zheng

BS, Genetics and Genetic Engineering, Fudan University, China, 1989

MS, Information Science, University of Pittsburgh, 1998

MS, Pathology, University of Pittsburgh, 2002

Submitted to the Graduate Faculty of
School of Information Sciences in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH
SCHOOL OF INFORMATION SCIENCES

This dissertation was presented
by

Lei Zheng

It was defended on

October 31, 2005

and approved by

David J. Foran, Ph.D., Department of Pathology, UMDNJ

John R. Gilbertson, MD, Department of Pathology

Hassan A. Karimi, PhD, Department of Information Science and Telecommunications

Dissertation Advisor: Michael J. Becich, PhD, MD, Department of Information Science and

Telecommunications

Dissertation Advisor, Dissertation Committee Chair: Paul W. Munro, PhD, Department of

Information Science and Telecommunications

Copyright © by Lei Zheng

2005

**AUTOMATED FEATURE EXTRACTION AND CONTENT-BASED RETRIEVAL OF
PATHOLOGY MICROSCOPIC IMAGES USING K-MEANS CLUSTERING AND
CODE RUN-LENGTH PROBABILITY DISTRIBUTION**

Lei Zheng, PhD

University of Pittsburgh, 2005

The dissertation starts with an extensive literature survey on the current issues in content-based image retrieval (CBIR) research, the state-of-the-art theories, methodologies, and implementations, covering topics such as general information retrieval theories, imaging, image feature identification and extraction, feature indexing and multimedia database search, user-system interaction, relevance feedback, and performance evaluation. A general CBIR framework has been proposed with three layers: image document space, feature space, and concept space. The framework emphasizes that while the projection from the image document space to the feature space is algorithmic and unrestricted, the connection between the feature space and the concept space is based on statistics instead of semantics. The scheme favors image features that do not rely on excessive assumptions about image content

As an attempt to design a new CBIR methodology following the above framework, k-means clustering color quantization is applied to pathology microscopic images, followed by code run-length probability distribution feature extraction. Kulback-Liebler divergence is used as distance measure for feature comparison. For content-based retrieval, the distance between two images is defined as a function of all individual features. The process is highly automated and the system is capable of working effectively across different tissues without human interference. Possible improvements and future directions have been discussed.

TABLE OF CONTENTS

PREFACE	X
1.0 IMAGE AND DIGITAL IMAGING	1
1.1 A TAXONOMY OF IMAGES	1
1.2 IMAGE REPRESENTATION	4
1.3 IMAGE FEATURE	7
2.0 IMAGE INFORMATION RETRIEVAL: A LITTLE THEORY	10
2.1 A BRIEF HISTORY	10
2.2 DOCUMENT AND QUERY	11
2.3 FEATURE AND RETRIEVAL	13
2.4 RETRIEVAL METHODOLOGY	17
2.5 CONTENT-BASED IMAGE RETRIEVAL MODEL	19
3.0 IMAGE INFORMATION RETRIEVAL: VARIOUS APPROACHES	24
3.1 IMAGE MANUAL INDEXING	24
3.2 AUTOMATED IMAGE FEATURE EXTRACTION	27
3.3 IMPORTANT IMAGE FEATURES	30
3.3.1 Color Feature Matching	30
3.3.2 Texture Feature Matching	33
3.3.3 Shape and Sketch Features	37
3.3.4 Image Feature Matching in Compressed Domain	41
3.3.5 Composite Image Features and Region Based Matching	54
3.3.6 Conclusions	57
3.4 FEATURE WEIGHTING, FEATURE SELECTION, AND RELEVANCE JUDGMENT	58
4.0 SYSTEM IMPLEMENTATION AND PERFORMANCE EVALUATION	61

4.1	FEATURE INDEXING.....	61
4.2	USER-SYSTEM INTERACTION: RELEVANCE FEEDBACK.....	64
4.3	PERFORMANCE EVALUATION.....	66
5.0	METHODOLOGY: COLOR QUANTIZATION OF PATHOLOGY MICROSCOPIC IMAGE,FEATURE EXTRACTION, AND CONTENT INDEXING	69
5.1	BACKGROUND	69
5.2	SCIENTIFIC CONTRIBUTIONS	77
5.2.1	Major Scientific contributions	77
5.2.2	How many colors do we need?	80
5.2.3	Discrete domain processing.....	82
5.3	HYPOTHESIS AND EVALUATION.....	85
5.4	MATERIAL AND METHODS	89
5.4.1	Materials	89
5.4.2	Methods.....	92
5.5	EVALUATION	102
5.5.1	Information Packing Capability	102
5.5.2	Content-Based Image Retrieval Performance.....	104
6.0	RESULTS AND INTERPRETATIONS	107
6.1	PRINCIPAL COMPONENT ANALYSIS	107
6.2	K-MEANS CLUSTERING OF COLOR PIXELS	109
6.3	K-CODING	111
6.4	K-CODE RUN LENGTH PROBABILITY DISTRIBUTION	114
6.5	RUN-LENGTH PROBABILITY DISTRIBUTION FEATURE FOR CONTENT CLASSIFICATION.....	116
6.6	CONTENT-BASED IMAGE RETRIEVAL PERFORMANCE.....	118
6.7	PERFORMANCE ANALYSIS AND INTERPRETATION	121
6.7.1	Interpretation of Performance Metric	121
6.7.2	Pathology Classification and Specificity of Image Feature.....	125
6.7.3	Within-Class Divergence	129
6.8	SUMMARY	131
7.0	CONCLUSION AND FUTURE WORK	134

7.1	PATHOLOGY MICROSCOPIC IMAGES AND RUN-LENGTH FEATURE.....	134
7.2	FEATURE EXTRACTION AND SIMILARITY MEASURE.....	135
7.3	RUN-LENGTH FEATURE FOR TISSUE CLASSIFICATION AND CONTENT-BASED IMAGE RETRIEVAL.....	136
7.4	FUTURE WORK.....	137
7.4.1	Image quality control.....	137
7.4.2	Rotation invariant and two-dimensional features	138
7.4.3	Feature extraction and modeling.....	140
	BIBLIOGRAPHY.....	142

LIST OF TABLES

1	Table 5.1 Images and tissue types.....	91
2	Table 6.1 Comparison of distortion	111
3	Table 6.2. Content-based Image Retrieval Performance	119
4	Table 6.3. CBIR performance confusion matrix.....	120

LIST OF FIGURES

1	Figure 2.1 Content-Based Image Retrieval.....	13
2	Figure 2.2 Three-Space-Two-Mapping Model.....	21
3	Figure 3.1. Information scale of various retrieval tasks.....	25
4	Figure 3.2. Color feature matching.....	33
5	Figure 3.3. Retrieval based on texture feature.....	36
6	Figure 3.4. Sketch matching.....	40
7	Figure 3.5. Wavelet popularity.....	47
8	Figure 3.6. Wavelet functions.....	48
9	Figure 3.7. Diagram of Haar's wavelet transformation.....	52
10	Figure 5.1. A typical run-length probability distribution for a random code.....	97
11	Figure 5.2 Flowchart.....	101
12	Figure 6.2 K-means clustering.....	109
13	Figure 6.3. K-coding of pathology microscopic image.....	113
14	Figure 6.4 K-coding error.....	113
15	Figure 6.5. K-code run-length probability distribution of one image.....	115
16	Figure 6.6. K-code run-length probability distribution.....	117
17	Figure 6.7. Content-based Image Retrieval Performance.....	121
18	Figure 6.8. Query session: brain.....	123
19	Figure 6.9. Query session: heart muscle.....	124
20	Figure 6.10. Query session: autolysis.....	127
21	Figure 6.11. Query session: kidney tubules.....	128
22	Figure 6.12. Query session: prostate cancer Gleason's grade 3.....	129

PREFACE

I thank Dr. Michael Becich for his generous support, guidance, and patience through the full course of this research, without which every bit of this thesis is simply not possible. I also want to thank Dr. Paul Munro for his mentorship, insightful discussion about the problems and solutions in the very details that usually take hours of his precious time. I owe debt to Dr. John Gilbertson, who helped me with data collection, image classification, provided advices, and for his many other roles.

1.0 IMAGE AND DIGITAL IMAGING

It is said that one image is worth a thousand words. Visual information accounts for about 90% of the total information content that a person acquires from the environment through his sensory systems. This reflects the fact that human being relies heavily on his highly developed visual system compared with other sensory pathways. The external optical signal is perceived by eyes, and then converted into neural signal; the corresponding neural subsystem specialized for visual system is specially organized to detect subtle image features and perform high-level processing, which is further processed to generate object entities and concepts.

The anatomy of the visual system explains from the structure aspect why visual information is so important to human cognition. The cognitive functions that such a system must support include the capability to distinguish among objects, their positions in space, motion, sizes, shapes, and surface texture [stillings95, p463]. Some of these primitives can be used as descriptors of image content in machine vision research.

1.1 A TAXONOMY OF IMAGES

There are two formats, in which visual information can be recorded and presented – static image, and motion picture, or video. Image is the major focus of research interest in digital image processing and image understanding. Although a relatively recent development, computerized

digital image processing has attracted much attention and shed lights to a broad range of existing and potential applications. This is directly caused by rapid accumulation of image data, a consequence of exponential increases of digital storage capacity and computer processing power. There are several major types of digital images depending on the elemental constituents that convey the image content. Images can take the form of:

1. Printed text and manuscript. Some examples of the kind are micro-films of old text documents, photograph of handwriting.
2. Line sketch, including diagrams, simple line graphs
3. Halftones. Images are represented by a grid of dots of variable sizes and shapes.
4. Continuous tone. Photographic images that use smooth and subtle tones.
5. Mixture of above.

Among all above, continuous tone or photographic images are most common in the digital imaging practice and of major concern of content-based image retrieval. They used to be generated by converting from other media using a scanner. The process is labor intensive and costly. It was estimated that the image capturing and the subsequent manual indexing may account for 90 percent of the total cost of building an image database [besser95] [gettyedu]. This was the situation a few years ago. Now, digital cameras are becoming very popular and images are also being converted from analog electronic format (such as analog videos) to digital format.

At the conceptual level, an image is a representation of its target object(s). According to the Webster's 3rd New International Dictionary [webster93], an image is “the optical counterpart of an object ... a mental picture, a mental conception ...” The definition reflects the fact that an image has its content, which captures the optical or mental properties of an object; with its format varying across different kinds of media. It is determined by how the optical properties are quantized, or the

degree of mental abstractions that is required. Here are some examples of image according to the above broader definition:

1. An image can be an array of pixel values stored in uncompressed bitmap digital image format. In this format, each value represents the color intensity at discrete points or pixels. A well-known example of this is Microsoft's BMP format. Although BMP format allows for pixel packing and run-length encoding to achieve certain level of compression, its uncompressed version is more popular.

2. Popular Internet standard image formats see more extensive image transformation and compression, such as GIF and JPEG. The GIF standard defines a color degeneration process, which maps the colors in an image into no more than 256 new colors. Further compression using LZW algorithm (after the names of the inventors, Abraham Lempel, Jacob Ziv, and Terry Welsh) is then applied. The GIF image file stores the color palette along with the compressed pixel values. On the other hand, JPEG applies transformation rather than mapping to the pixel intensity. Discrete cosine transformation (DCT) is used to transform the pixel value representation in the spatial domain into a frequency domain representation. Instead of pixel values, cosine transformation coefficients are used to represent the wave signals. The two-dimensional image is represented as the combination of the wave signals of different frequencies. DCT is an orthogonal transformation that is computationally reversible. The information packing capability of DCT is harvested by the subsequent quantization step, which takes advantage of the characteristics of human vision by filtering out the part of the signals that is not significant to human visual system. Entropy coding or arithmetic coding is applied to achieve higher compression ratio before the coefficients are stored [miano99].

3. Specially defined signature file stores specifically extract images features in numeric or Boolean format, indicating the presence (or non-presence) and strength of the features. This is a very compact representation of image that is targeted for fast retrieval instead of display, archival, etc. Only the essential image features are conserved in the signature file and they are algorithm dependent. This allows for easy indexing and fast search for matching features of the query. An example of this can be found in image coding method using vector quantization (VQ), in which image blocks are coded according to a carefully chosen codebook. If the image blocks are similar to each other or the images in a set bear significant similarity, higher compression ratio can usually be achieved than the general- purpose compression algorithms such as GIF, JPEG.

4. Textual annotation can also be thought of as an instantiation of mental image, and sometimes, the descriptors can be coded by a predefined convention, or a thesaurus. The fact that two visually different images can convey the same concept and different concepts may present in images that share many similar optimal properties brings about a gap between image retrieval by content, and retrieval by concept [rasmussen97, Faloutsos].

1.2 IMAGE REPRESENTATION

The design of various image file formats usually has its root in the physical hardware implementation that supports the imaging process, the purpose of the image content, and how the image is to be processed. For example, the raster representation of image is closely related to the digital image sampling and display apparatus. Wavelet coefficient representation is used in some systems that require continuous transmission through computer network with zooming capability.

Simple image formats such as Microsoft's uncompressed Windows BMP format consists of 2 very simple parts, a simple header part that contains the metadata information about images and a second part that stores the intensity of each pixel sequentially. The header part has information such as: the size of the header part, the image width, image length, number of bits per pixel, preferred resolution (in pixel per meter), number of significant colors, and some other reserved fields. In uncompressed BMP file with 24-bit per pixel, right after the image header, the pixel values are stored in the order of scan line starting from the bottom to the top with the last line coming first. In each scan line, the pixel values are arranged from left to right, with 3 bytes for each pixel defining the value for Red, Green, and Blue, so that each color channel takes 8 bits which maps to a range from 0 to 255. For memory access efficiency, the number of bytes per scan line is rounded up to a multiple of four. The rest of the bytes are padded with arbitrary values. For compressed BMP files, or when the number of bits per pixel does not equal to 24 (number of bits per pixel could be 1, 4, 8, or 24), extra steps may be required to generate a BMP file from the raster representation.

Since image files require tremendous storage space, and demand substantial network bandwidth in raw pixel value format, compression is much desired in most occasions. Usually, image compression algorithms can bring about 10 fold of reduction in the file size without significantly affecting the visual quality. On the other hand, the compression is at the price of extensive processing and CPU power. The most commonly used compression routines usually include some kind of transformation, such as discrete cosine transformation (DCT) and discrete wavelet transformation (DWT), followed by general compression algorithms, such as run length encoding (RLE) and entropy encoding (i.e. Huffman encoding). There is an optional color space transformation/quantization at the beginning of the compression routine as can be seen in GIF format and JPEG standard.

The pixel values should undergo sophisticated processing before being stored in JFIF format, which is better known as JPEG file. There are four major steps involved, namely: color space transformation, discrete cosine transformation (DCT), quantization, and entropy compression. First, the pixel values should be converted from RGB color model into YCbCr color model, which has the advantage of concentrate the information in Y channel so as to allow better compression for the Cb and Cr channels. Then, the image is divided into 8 x 8 blocks, on which two-dimensional DCT is performed for each of the three channels. DCT is a kind of orthogonal transformation that can be implemented as matrix multiplication with a standard 2-dimensional DCT matrix. The resulting coefficients are quantized according to their impacts to the visual quality, with more information conserved from the low frequency components, while less from the high frequency ones according to empirical experience. This is achieved through dividing the coefficient matrix by two quantization tables, for the Y components, and Cb-Cr components. Huffman encoding is used to make the final compression. The JPEG standard allows for many different implementations and the compression ratio can be adjusted. The JPEG standard even accommodates a lossless format.

A more hardware oriented image format is TIFF (which stands for Tag Image File Format) format. TIFF serves as a wrapper that holds one or more image frames in one file, each with metadata information stored with tags. The image in each frame can be in different format, either compressed or uncompressed. TIFF handles the image frames individually. Usually, TIFF file is handled through a library of functions that provide a consistent API for manipulating the image at different level of granularity, such as pixel, scan-line, stripe, and block. One such library is the libtiff, which can be found at [<http://www.libtiff.org/>].

The image format is an important factor in the image retrieval, as it determines the primary image constituents that convey the visual information content. They could be pixels, frequency coefficients, color histogram, concept terms, as in above 4 kinds of images. Traditionally, all image processing researches deal with the first kind of image format, as other 3 can be derived from that. However, to manipulate compressed images is regarded of great importance in some real world applications, such as high fidelity video transmission, replay and manipulation; signature file is a common practice in many retrieval systems; and the most widely used and reliable image retrieval systems still employ text annotation as a mediator for retrieval, where expert manual labeling of the images is essential. This also forms a ladder of image abstraction from the physical optical entity. The kinds of features we are interested in, and the corresponding algorithms or methodology to extract the most descriptive features determine the details of the abstraction.

1.3 IMAGE FEATURE

Some researchers have reported the kinds of characteristics that are useful for image content description. They are usually well defined by a set of supporting image-processing routines. On the other hand, they can also be easily mapped to a certain aspect of human visual perception. The human visual perception process and the construction of mental image are outlined by the theory of primal sketch by Marr [marr76] [marr82]. The human vision system can differentiate objects, identify their location in space, their motions, sizes, shapes, and their surface textures [stilling95, p463-464]. Accordingly, machine vision systems also devise algorithms for object recognition, motion detection, edge detection and contour defining, shape descriptive, and texture pattern identification. Among these, object recognition, edge and contour, shape, and texture are relevant to

static image retrieval, and the color feature is very often used to help in every aspect of these characteristics, or used alone as a global feature of the whole image.

Now, we can safely define the concept of image feature based on the above analysis.

An image feature can be defined as the result of a computation or an expert evaluation criterion, performed to a target image. Practically, from the perspective of machine vision and computerized content-based image retrieval, we can define the content of an image as the set of all possible features, or combination of basic features, of that target image. The nature of image content we need to deal with, the user's information need, and the way human users interpret the content of the image, all three of these vary from task to task. As an exploratory approach to understand image content using machine intelligence, it is not uncommon to resort to image features less intuitive to human perception, or even develop new algorithms to define novel image features, in order to address particular image content and user information need. These facts welcome an open definition of image feature as given above that is extensible to accommodate particular cases of application. It is also worth noticing that an image feature is tied to the algorithm(s) that defines it.

According to different interests, image content can be perceived in three levels with increasing complexity: machine vision is good at discerning primitive features including, but not limited to, color, texture, and shape [rui99]; human image information consumers are more concerned with the composition of the primitive features to form objects and the semantic relationship among multiple objects in the same image context. At the highest level, image content can be symbolized by the concepts that it conveys [eakins96] [eakins98]. Accordingly, there are queries that the users can generate focusing on different levels of content. Some of them can be satisfied by focusing on only one or two basic image features, while others require understanding of the semantics of the images.

In chapter 2, I will introduce the concept of document feature, of which image feature is a special case, in the context of general information retrieval framework. The utilization of image features in content-based image retrieval will be further explained in chapter 3.

2.0 IMAGE INFORMATION RETRIEVAL: A LITTLE THEORY.

2.1 A BRIEF HISTORY

Although it was not until the early 1970s when the problem of retrieving digital images from archives started to draw people's attention, the practice of information retrieval has been a part of the librarians' daily responsibility long before electronic computer was popular enough to enter this arena. Some researchers would date the practice of information retrieval back to the earliest library systems that can be found in the record, which is as early as the third century B.C., in the Library of Alexandria with catalogs and classifications for its 500,000 stored volumes. Since then, theories and methodologies have been developed for library information retrieval that target at handling printed text documents. Most of them are still applicable for general information retrieval tasks. With a little bit of generalization and modification, they are able to accommodate the multimedia information retrieval scenarios, and provide us an insight into the multimedia information retrieval problem from a general information content perspective [squire00]. It was not until the late 1990s that the research on image, video, and audio retrieval took off.

2.2 DOCUMENT AND QUERY

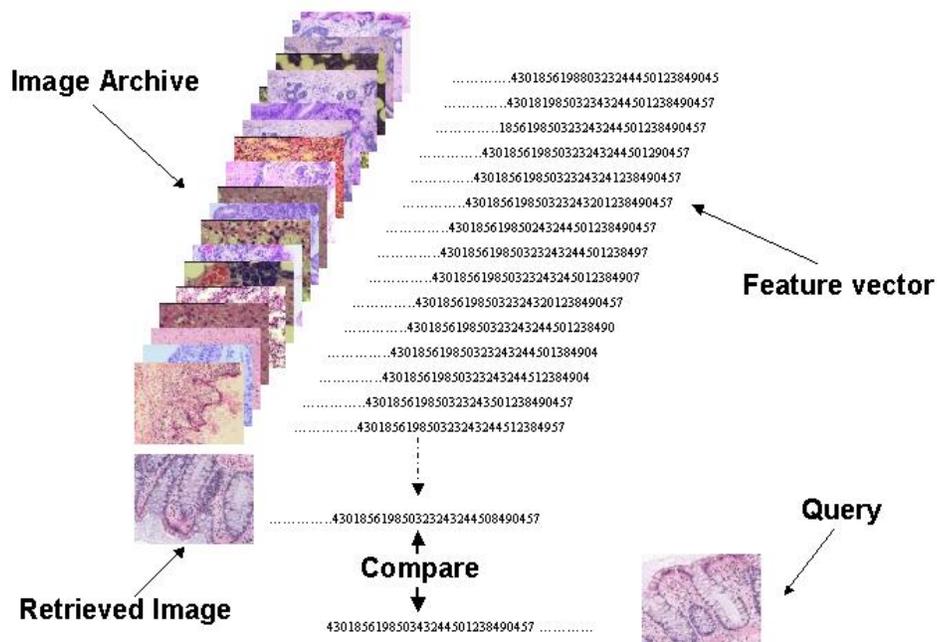
An information system consists of two major types of entities: document and query. Documents can be of any format, with any kind of internal structure, in organizing the information it carries. An information retrieval system treats all documents in collection as unstructured files. Documents can be represented with document surrogates. A document surrogate is usually a concise representation of the original document being processed for particular purposes, and contains only a small portion of the information that the complete document carries. Some examples of a document surrogates are article abstracts and image icons. They are more suitable for efficient storage and fast access and processing. A query is an indication of user's information need. It can take many different forms, ranging from Boolean query that is very concise, to a very detailed specification of the kind of documents the user needs that spans longer than the actual documents themselves, to one or more sample documents (such as images queries) that are set as examples for the kind of desired documents.

Researchers tend to think of queries and document surrogates as special kinds of documents that bear many of the same characteristics as regular documents. The retrieval process is to find documents that match the query document based on the specified criteria, and processing procedure that is applied to both query and documents in the archive [korfhage97, chapter 2]. This vision helps to bring the image retrieval problem and many derived techniques into the big picture of information retrieval.

With image retrieval, all the images in the archive are usually processed by machine or indexed by human experts to generate the signature file and the textual annotation for feature abstraction and, later, faster retrieval. Document surrogates are a constituent part of the image database that remains relatively static. On the query side, query-by-example has gained wide

acceptance as a major image database querying scheme. In this approach, the query is not used for direct comparison. Instead, the query image is processed first, following a similar procedure to generate the "query surrogate", which is then compared with the document surrogates from the database.

The document surrogate (and also the query surrogate) should retain as many useful features as possible while filter out those noises. Some researchers would define document feature as, "Information extracted from an object and used during query processing" [yate99]. Since, as stated previously, all objects in an information storage and retrieval system can be unified as documents, and all documents are usually processed to generate document surrogates, before a query is even defined. The document surrogates, essentially a compact representation of the set of all document features of interest, are relatively static and stored as a persistent part of the information system, with a longer lifetime than the ephemeral queries. Figure 1 below shows the typical architecture of a content-based image retrieval system with image documents, document surrogates (feature vectors), and query illustrated. In this work, a document feature refers to an information element that is extracted from the original document for the purpose of indexing and retrieval.



1 Figure 2.1 Content-Based Image Retrieval

- 1) Each image is broken down into a feature vector of numerical features;
- 2) image comparison is based on vector arithmetic;
- 3) database search is to find the feature vector(s) that have the minimal difference to the query image.

2.3 FEATURE AND RETRIEVAL

Document feature is an important concept in information retrieval. It defines the way that documents are compared with each other. The definition of document feature also implies several fundamental methodology of information retrieval. Essentially, information retrieval is the study of the statistical characteristics of unstructured data [callen00]. In the study of text retrieval, a document feature is an indexing term, which could take the form of a word, a phrase, or an N-gram.

Information retrieval tries to find statistical patterns of the document features in the data without understanding the structure of document or the implications of the features.

Firstly, all documents, by definition, are unstructured, which means, either the structure of the data is unknown, or the semantics of the data component are unknown. In information retrieval, a document is defined as an information object with unknown structure (Callan00). In contrast to the structural representations that are popular in other research areas such as relational database management system (RDBMS), natural language processing (NLP), and extended markup language (XML) techniques, most free text documents, audio, and pixel representation of images are unstructured data.

Secondly, information retrieval doesn't make any effort in interpreting the meaning of the indexing terms. Nor does it make any prior assumption about those meanings. The only thing that is used to make the retrieval decision is the statistical patterns of the document features. In the content-based image retrieval, object recognition and image understanding are also based on the identification of statistical patterns that are characteristic to the content.

Thirdly, all features are treated as independent to each other at the time of processing. This is a further assumption from the second one, and could be dangerous in theory. The redundant, ambiguous nature of the language we use determines that there are correlations among the features. Also, different interpretations can be applied to the same word in different contexts, which means the interpretation of a feature is not self-contained, independent, and context free. However, in practice, all text retrieval models work fine under the assumption, and it helps to clear many barriers and bring about the major success we have seen in the last decade of last century. The rising of many successful web search engines is a good proof for the success of the theory. The above three assumptions are usually referred to as the "bag of words" assumption in text

information retrieval, where every non-stop word is used as indexing term without regarding to their actual meaning and the internal document structure.

These assumptions actually make more sense with multimedia information retrieval, where multimedia objects/documents are records of their physical properties: 1) there is no internal semantic structure for those records; 2) human perception of those physical properties by sensory system is not necessarily connected to the interpretation of the content; and 3) either the recorded physical properties are independent to each other, or we can extract only independent properties from the object and leave out any possible redundant information. These properties can be used as indexing features of the multimedia documents. It is plausible that in the special case of image retrieval, we can adopt the same set of assumptions but under a different name as "bag of pixels". This scheme might exclude such techniques as object recognition from the big map in a pure sense, just as text information retrieval would with natural language processing. However, just as having been proved with text retrieval, the combination of the techniques is expected to generate much more power than each of them used alone.

The feature extraction step is usually designed to meet the users' specific information need with the domain context of the information in mind. Special attention should be paid to the types of the queries that system is expected to answer, and the nature of the document collection. Although it is one of the most important steps towards retrieval, there are no general rules that are applicable to a variety of problems. A document could either be manually indexed (such as keywords identified by the authors of a publication), or automatically indexed by computer. The features can be carefully selected to target at a particular type of content, e.g. indexing text documents with a controlled vocabulary such as the terms from a thesaurus, or the features can be massively extracted automatically from the documents, and then selected according to their relevance to the retrieval.

Initial studies in 1960s showed that either manual indexing or automatic indexing usually gave equally good performance (canfield experiments), and the conclusion has been supported by other researchers since then. Primitive studies in multimedia retrieval have also shown very promising results without resorting to manual indexing by human experts. However, the best results are often achieved by combining both approaches due to the fact that it is not likely for both approaches to make the same mistakes in retrieving non-relevant documents, but rather to agree upon retrieving those truly relevant documents. This is due to the benefit of the combination of evidence.

The features of the documents from the database (in information retrieval, the term database is used interchangeably with document collection, rather than its meaning in database management system research) are usually indexed and stored as document surrogates, which become a static component of the database. Fewer real systems require each document in the collection to be processed for every query session (one exception being the Fast Data Finder system). The features can be stored in several different kinds of formats for the kind of query and the kind of users' information need the system will serve. Those that have been implemented in real systems include bitmap, signature file, inverted list, and permuterm. They could be either compressed to improve performance, or uncompressed for better manageability. All of these have been found working well in various text retrieval systems, while some of them are more general-purpose and have found better usage in multimedia information retrieval. The detail of the storage and index format of the feature values will be further discussed in the following chapters.

2.4 RETRIEVAL METHODOLOGY

Having defined the concept of document features, the retrieval process can be viewed as a mapping process, in which documents in the document space are projected into a feature space, followed by a matching processing according to a retrieval model that group the documents with similar concepts in proximity in the concept space. In the document space, feature space, and in the concept space, there are patterns in the distribution of objects. Exploiting these patterns would help to improve the retrieval performance by selecting better features, assigning weights to these features, and designing better transformation metric. The topic of application of data mining techniques, such as cluster analysis, multi-dimensional scaling, in addressing this problem will be revisited where appropriate in later chapters.

According to the assumption made about the data distribution of the feature values, several retrieval models have been used to make retrieval decisions. All features do not contribute equally to the distinguishing of the document content. The distributions of the values also affect their quality as distinguishing features. Different retrieval models employ different mechanisms in determining the contribution of each of the features to the retrieval decision.

Several information retrieval models have seen great success in the past, including Boolean model, vector space model (one of the variations is latent semantic indexing (LSI)), probabilistic model (including basic probabilistic model, Bayesian inference networks, and language models) and fuzzy matching model, citation analysis models (including hubs & authorities, and page ranking mechanisms). Besides, artificial neural networks, logic-based, and natural language processing (NPL) techniques have also been used. These models take very different views, and make different assumptions about the distribution of the features. Accordingly, the definition of document similarity varies greatly across these different models. For example, in some cases, there

is a measurable document distance that is inversely proportional to the similarity measure; while in other systems, document similarity can only be inferred indirectly from the probability, of which two documents are similar to each other. The latter situation includes retrieval criteria determined by co-occurrence data and/or user feedback that are not characteristics of the documents themselves but rather the connections imposed to them through users' information seeking behavior. Thus, relevance of the document may be a better term than document similarity in describing the retrieval criterion. It is important to view the retrieval process as a process that matches the documents and the query, which is an instantiation of the user's information need. Because of the massive success of the vector space model, the document similarity measure is mistakenly regarded as the sole criterion for retrieval.

Not surprisingly, most models in practical use give more or less the same level of performance. Although this leaves the necessity of research on various retrieval models doubtful, they have been constantly improved by combining the strength of different models, and bringing in new techniques, particularly, data mining techniques, to the existing models.

Boolean-based matching is the most obvious approach to text document retrieval. Boolean retrieval systems appeared along with the technology such as punched cards and edge-notched cards in 1930s. In such a system, a query is an aggregation of specification of binary features in the format of a first-order logic (FOL) statement, and deductive inference is used as retrieval algorithm. A feature can take one of the three values, present, non-present, or unspecified. The documents that meet the requirement are considered relevant to the query. In spite of its simplicity and high search efficiency, a pure Boolean model has little strength in dealing with real world problems. However, Boolean matching can be found as an essential flavor in many retrieval systems. For example,

proximity matching is a very common add-on to the Boolean model and makes it very powerful in text document retrieval.

Vector space model combined with feature weighting is arguably the most important and most widely used retrieval model. It is a framework that allows various customizations. Definition of the distance metric is core to the vector space model. Euclidean distance, Bayesian distance, *Mahalanobis* distance (simplified *Mahalanobis* distance) or variations of them are among the popular choices. Latent semantic indexing (LSI) takes another step from the scheme of vectorized features of the vector space model. Probabilistic model and fuzzy matching are similar in implementation but make different assumptions about the data distribution in theory. Bayesian inference network is an application of probabilistic model. For introduction of various retrieval models, a recent textbook is [yate99].

2.5 CONTENT-BASED IMAGE RETRIEVAL MODEL

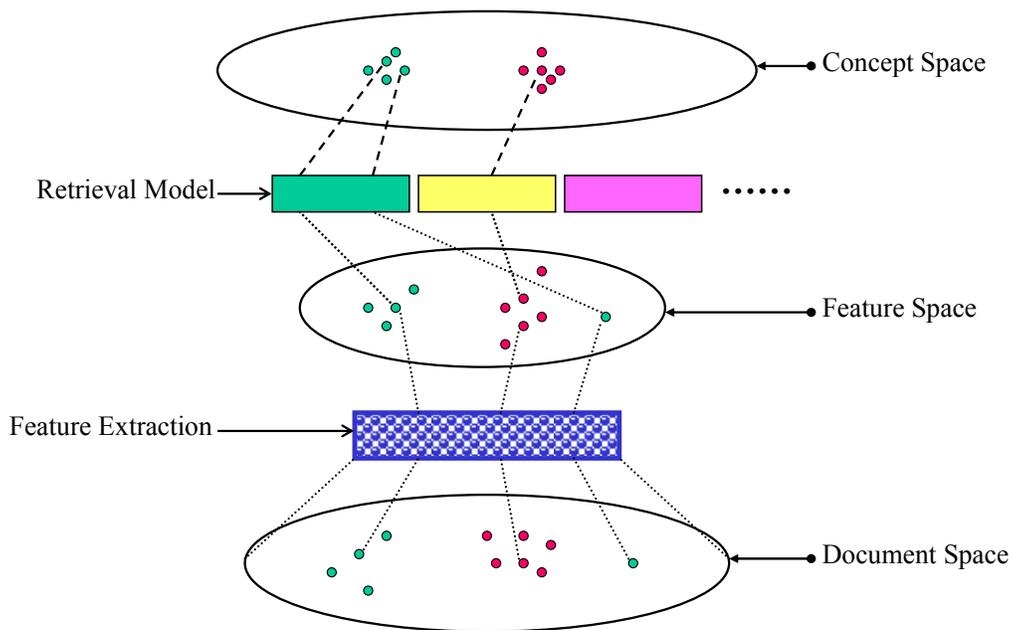
To summarize, information retrieval is a process of finding relevant documents in a document collection based on query. For both information processing and practical system performance reasons, documents are not compared directly. Instead, they are processed following a similar routine of essential feature extraction to generate document surrogates. The feature values are then used as input to a particular retrieval model. A retrieval model makes certain assumptions about the distribution of the feature values in the feature space, and determines the relevance of the documents by manipulating the document surrogates in the feature space. The adopted retrieval model would project those document surrogates bearing similar concepts to proximity in the concept space. There are 3 spaces in this simplified model, the document space, the feature space,

and the concept space. Two transformations are required to project the objects in the lower level spaces to the more abstract spaces. The first transformation is the feature extraction step, in which documents are processed to generate feature values. The goal of the second transformation is to overcome the “semantic gap [smeulder00]” between the image features and the concepts in the same domain in which the user’s information need is specified. A three-level model was introduced by Eakins [eakinsURL]. In that model, all three levels are used as handles for retrieval. Another four-layer was introduced along with the VIPER (now GNU GIFT, www.gnu.org/software/gift/) system.

The advantage of the model proposed here is that it clearly defines the basic image representation, the derived features, and the concepts that the image delivers, and the user's information need in their own right. The two projections are core to the retrieval process where most uncertainties rise and most researches take place. There is no guarantee that the projections are always valid and provide a viable solution for retrieval based on only the bag-of-pixels assumption, that is, without really "understanding" the image intelligently. This is because that the definition of image content varies from user to user. Even with the same query, different users may have different information need. The information content of the image can be consumed in different ways by even the same user. Besides, the whole notion of artificial intelligence is not to replicate human intelligence, but rather achieve the same level of intelligence with the approaches that are most suitable for machine. At last, while an indexing term in text retrieval may be used as a concept by itself, there is no deterministic or causal relationship that connects an optical property of an object to the concept it bears.

Thus, human semantic framework about image content is not guaranteed to be replicated faithfully in a computerized information retrieval system. However, within a well-defined domain context,

and when the image content in general is under careful control, it is safe to make the assumption that correct mappings between these three spaces exist and we can develop such algorithms based on the statistical patterns. Under certain circumstances, primitive image features, such as the color histogram, alone are sufficient to devise a working retrieval algorithm. Computationally "cheap" features might be just as important as those high-level features that require more processing and complex algorithms. From the perspective of the goal of retrieving relevant documents, it is not appropriate to divide features into different levels based on the complexity of feature extraction process.



2 Figure 2.2 Three-Space-Two-Mapping Model

The document space is organized according to the physical properties of the image, the image format, and the optical properties being recorded (color intensity, or frequencies of changes, or other coefficients; it has been shown that

all features, low level or high level, are treated indiscriminately by machine, and can play roles of similar importance in the retrieval process; according to users' different information needs and domain context, the concept space is very much fragmented. The mapping from the image document space to the feature space is based on image processing algorithms, which are usually quite subjective and universal to all content description processes regardless of data and users' information needs. The extracted features are reorganized, weighted, with noise removed according to specific task definition, training data, and retrieval model of choice. Different retrieval models can be used for different tasks or for same task, so that the same set of features can be mapped to very different fragments of the concept space. This is the part that contains the most intelligent content of the system.

There are refinements added to how these two transformations are performed. An important manipulation that accompanies both transformations is the feature weighting. In the feature extraction step, features can be weighted according to the nature of raw data and the algorithms that are used to process the data. Those features that help to distinguish the objects in the document space in general often receive heavier weights. Those don't, receive lower weight or even be eliminated. An example of this is the quantization step following the discrete cosine transformation (DCT) in the JPEG compression. The DCT coefficients and the three color channels are weighted differently according to the two quantization matrices, which is designed to reflect the significance of those coefficients to the visual quality of the image.

Many retrieval models often imply some kind of feature weighting scheme according to their assumptions about the distribution of the feature values. This fact also brings about the variations even in the same retrieval model. A typical example of this in text document retrieval is that how term frequencies are normalized. They can be either: 1) not normalized, 2) normalized by document length, 3) normalized by their frequencies in both relevant documents and irrelevant

documents from a ground truth collection, or 4) normalized by any kind of mathematically reasonable manipulations, such as taking log of them.

As a senior researcher stated it, information retrieval in general is an area that sees a little theory, and a lot of practice. There are many different approaches. In the next chapter, the current state of the art of research of content-based image retrieval will be reviewed with an emphasis on image feature manipulation.

3.0 IMAGE INFORMATION RETRIEVAL: VARIOUS APPROACHES

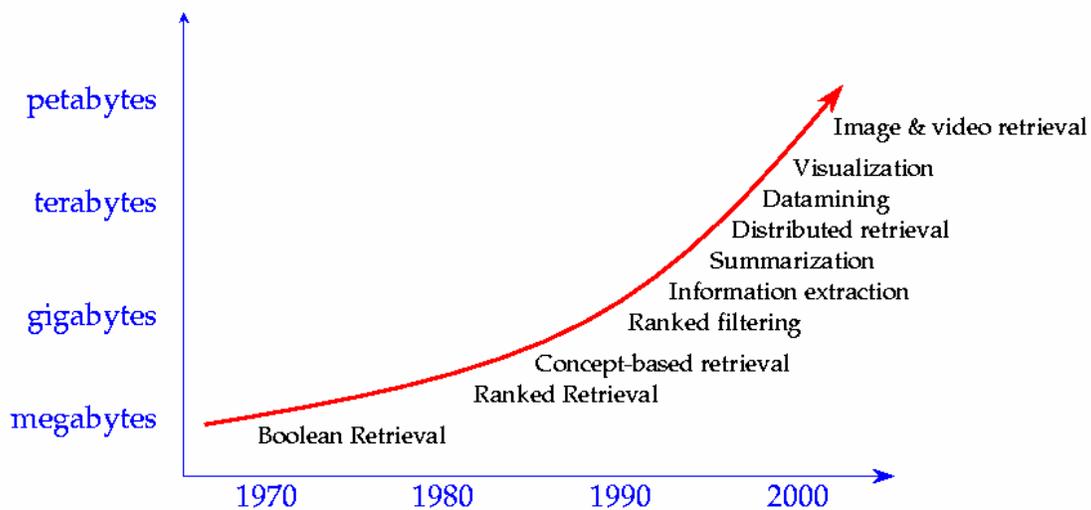
This chapter will examine the image features that have been used for image content description and content-based image retrieval research and practice. Not all surveyed research works on the image feature description necessary lead to implementation of a working retrieval system. Some use synthetic data or image collection for test purpose, such as those in texture features testing. Vector space model and distance measure is the most common feature matching scheme, in which a feature gives a vector of values, and the matching is based on the pair-wise distances of the feature vectors. Individual work may derive a customized distance equation that works best in a specific setting.

3.1 IMAGE MANUAL INDEXING

Early content-based image retrieval started with the emergence of large digital image databases in the early 1990s. It was not until 1997 that it began to draw wide attention, which was 2200 years after the founding of Library of Alexandria that started the earliest practice of information storage and retrieval; 30 years after the first production automated search service in 1967, marked by Lockheed's DIALOG system serving NASA projects, and Data Corporation's OBAR (Ohio Bar Automated Research) full-text retrieval system. From a hardware perspective, CBIR only came to

realization and matured as the powerful computer hardware had brought about the challenge of managing large scale image collection and had also provided computation power for the capable image processing and feature extraction algorithms.

With the availability of fast, cheap storage and digital imaging technology in the last decade, the advantage of capture, processing, storage, and transmission of image in digital format became apparent. Large image archives, such as those of satellite images and medical images, have been built up so rapidly that a methodology for automatic indexing and retrieval is in demand in order to better meet the consumers' information need. In many of those areas, digital imaging has been automated to great extent. Indexing those images manually would consume a huge amount of human effort. The Internet provides another driving force and incubator for the fast proceeding of the research in the area.



3 Figure 3.1. Information scale of various retrieval tasks.

(from lecture notes of Prof. Jamie Callan, Language Technology Institute, Carnegie Mellon University.)

Nevertheless, well before the term content-based image retrieval was forged [kato92], images were interpreted by human experts manually as an important step of indexing. In this process, text descriptors, either open set free format text or keywords from a standard thesaurus, are associated to the individual image. Text retrieval methodology can then be used to identify those descriptors that convey the same concept of the query. Those keywords and text terms are treated as image features. In other words, the projection from document space to the feature space is carried out manually by human expert. This approach is time consuming, arbitrary, susceptible to environmental changes, and inconsistent among different human experts, requiring substantial human labor in creating an indexing system, setting up conventions and building/customizing a thesaurus, and indexing every image in the database. The indexing scheme is pre-defined to serve a particular purpose and the anticipated users of the image archive, and thus content dependent. After the image archive has been indexed, it can only be used for limited purposes. Any change in the usage of the database, in the indexing scheme, or in the selected image features will cost another round of human labor to re-index all the images.

The tradeoff is that it requires very minimal computation power devoted to image processing and feature extraction step. This made it a viable solution when machine time was more expensive than human labor as in the history. At that time, some image processing techniques that are computationally complex either took too long or were regarded as too expensive and thus prohibitive for wide use, although the mathematical foundation of the algorithms might have been in existence for a long time and seen many successful applications in other area. Only very simple image primitives, such as the color histogram, were feasible to be used as indexing features, and were not capable of defining complex image content. The CBIR approach had to wait for its time

until storage and computation power are cheap enough to handle big image archives, and the world-wide-web provides incentive for providing digital content online.

Some of the old techniques have still proven to be useful. For example, the color histogram (see page 28 for a definition of color histogram) is still one of the most widely used indexing features. In retrieving web images, the text on the same page of the image, and the linkage text pointing to the image or the page provide a indexing setting somewhat similar to the text-based image retrieval scenario.

3.2 AUTOMATED IMAGE FEATURE EXTRACTION

With the maturation of computing technology and the CPU power of the computer, image features can be identified for the purpose of indexing. We are approaching to the point that we can use these features to describe the content of an image so that the images in a large database can be defined reasonably uniquely. In another sense, image documents can be projected from the document space to the feature space uniquely, so that fine granularity classification on the basis of the feature space manipulation becomes much possible. This allows us to define the image content from a machine intelligence perspective, which is not expected to duplicate exactly the process of human cognition. Practically, we can define the content of an image as the set of all possible features, or any combination of basic features, of that image; while an image feature can be defined as the numeric value generated by a pre-defined image-processing algorithm applied to the image of interest. Usually a feature value can be either a vector or a scalar. Thus, the feature extraction step allows us to reduce the two dimensional image to one dimension feature values. The feature values will be

ultimately reduced to relevance estimate, which is a scalar of zero dimension. This is within the framework of various retrieval models, which have been introduced in previous chapter.

Due to the fundamental differences between image processing, numerical feature extraction, and the understanding of image semantics, visual language, the validity of retrieval based on common primitive features is, at large, questionable [liuy]. The task of building an all-purpose CBIR system is equivalent to building an image understanding system that duplicates human visual perception, reasoning, and specific domain knowledge. Although limited success has been achieved in some CBIR systems with retrieval algorithms focusing on one or two kinds of carefully-selected primitive image features [das99] [smith95] [brandt99] [squire95] [deng99] [brunelli99] [srikanth99] [graham98], and progress has been made in applying artificial neural network classifiers directly to the digitized image [rowley98] [squire95] [ikeda00], the problem in general is still unsolved.

However, under careful image quality control, and by limiting the retrieval problem to a well-defined domain, it is possible to map primitive features to the content semantics. The existence of a mapping between machine generated image features and high level concepts of human cognition is a basic assumption that we rely on in order to justify all research efforts on applying artificial intelligence to machine image understanding. Restricting the scope of a CBIR system to a particular domain also helps to formalize the evaluation metric of system performance by making the relevance judgment meaningful.

Image content can be perceived in three levels with increasing complexity: machine vision is good at discerning primitive features including, but not limited to, color, texture, and shape [rui99]; human users are more interested in the composition of the primitive features to form objects and the semantic relationship among multiple objects in the same image; at the highest

level, image content can be symbolized by the concepts that it conveys [eakins96] [eakins98]. Accordingly, there are queries that the users can generate at all these levels. Some of them can be satisfied by focusing on only one or two basic image features, while for others understanding of the semantics of the images is essential.

Machines excel in more than one image processing schemes in deriving various features. Some are intuitive, some are not. Spatial domain processing is very intuitive, and widely used for the objectives such as image enhancement, object recognition. The manipulation focuses on pixels or aggregation of pixels within a defined region. Frequency domain processing is less intuitive, exploiting various transformation coefficients as basic elements, and widely used to generate compressed image/video formats. Recently, with the rapid rising of wavelet transformation, extracting image feature in frequency domain has attracted substantial attention in the wavelet community and content-based image retrieval community alike. Some higher level image processing techniques can use a mixture of algorithms developed in both domains. Early CBIR systems tended to rely on simple and computationally cheap features. It is a recent trend that multi-step, complex processing algorithms be used in identifying more comprehensive features with stronger connections to the targeted concepts and retrieval goals. However, the approaches here are mostly ad hoc. The difficulty in establishing a standard collection of test images like the those can be found in text retrieval, such as Cranfield, CACM, ISI, and most recently, TREC (A video TREC is available though. see <http://www-nlpir.nist.gov/projects/trecvid/>).

3.3 IMPORTANT IMAGE FEATURES

3.3.1 Color Feature Matching

The use of color feature can be found in many image retrieval systems because it is very easy to implement, costs little computation time, and is intuitive to human visual perception. Quite often, a homogeneous color block in an image corresponds roughly to a separate object. Machines can distinguish more subtle color differences than the human eye. It has been reported that in some particular image database, color feature alone can support fairly good retrieval accuracy.

In spatial domain representation, an image is a two-dimensional array of color pixels. To represent a color, it is necessary to define a color space model. There are several color space models, with the simplest and the most commonly used one as RGB model. It has its root in computer display hardware. In RGB model, there are three primary colors -- red, green, and blue, and they are additive. Any color is combination of these three primary colors with different contribution from each [buford94] [foley92]. Visualized in a three dimensional hyper space, RGB color model can be represented by a unit cube with eight corners as pure black, pure white, the three primary colors, and three secondary colors as cyan, magenta, and yellow). Many image formats use this color model or a palette derived from it, such as the popular BMP, GIF, and PNG.

To derive a color histogram of an image, a quantization step is usually added to distribute the pixels into a finite number of bins before the color histogram being derived, so that fewer colors need to be dealt with than actually in the image. The color intensity can be either uniformly quantized, or in a non-uniform way. Besides, vector quantization, including tree-structured vector quantization, and product quantization are also used [heckbert82] [jacobs95] [wan96]. The number

of pixels in each bin is then summed up. The process of dividing pixels into color bins can be refined to take into account the local features [pass99].

The generated histogram is a vector of N-dimension, where N is the number of colors after quantization. The color difference between two images is measured by the distance between the two color histograms. A variety of distance functions can be used, including Euclidean distance (aka L2 norm) [niblack94] [tseng95], quadratic distance [hafner95]. Other measures have also been proposed, such as L1 norm (color histogram intersection) [swain91] [srihari95]. Distance contributed by different colors can also be weighted so that the differences in the perceptually similar colors receive lower weights while those in the perceptually very different colors receive higher weights. A weight matrix has been proposed as a part of the MPEG-7 standard [IBM99]. In the measures mentioned above, L1 norm and L2 norm have been used by many researchers despite that L1 measure leads to low recall in missing matching images, and L2 tends to give low accuracy retrieval as the matching requirement is more relaxed.

A global color histogram can be used to represent the color feature of one image. In many cases, a region-based matching between two images produces more accurate relevance judgment. The images to be compared with each other are divided into sub-regions, either according to a hierarchical sequence such as a quad-tree, or through image segmentation process. This falls into the category of region-based matching, and will be covered later in the chapter.

The color feature of two images can be compared by a distance or similarity measure rather than by counting the number of pixels in each color bin. This measure is then combined with the presence of that color in the image to form the color feature descriptor. The descriptor can be of lower dimensionality than the color histogram and efficiently indexed to facilitate fast retrieval performance. The method was introduced as dominant color descriptor in [deng01]. The color

distance measure the authors adopted has been shown to be equivalent to the quadratic color histogram distance measure. The color descriptors were arranged in a D3 lattice structure for fast range query search.

In IBM's QBIC project, following color representation schemes are used:

1. Mean color.
2. Mean + standard deviation.
3. Multi-bin histogram. The distance of two color histogram is defined as following:

$$d_{hist}(I, Q) = (h(I) - h(Q))^T A (h(I) - h(Q))$$

where $h(I)$ is the color histogram of the image in the database, $h(Q)$ is that of the query,

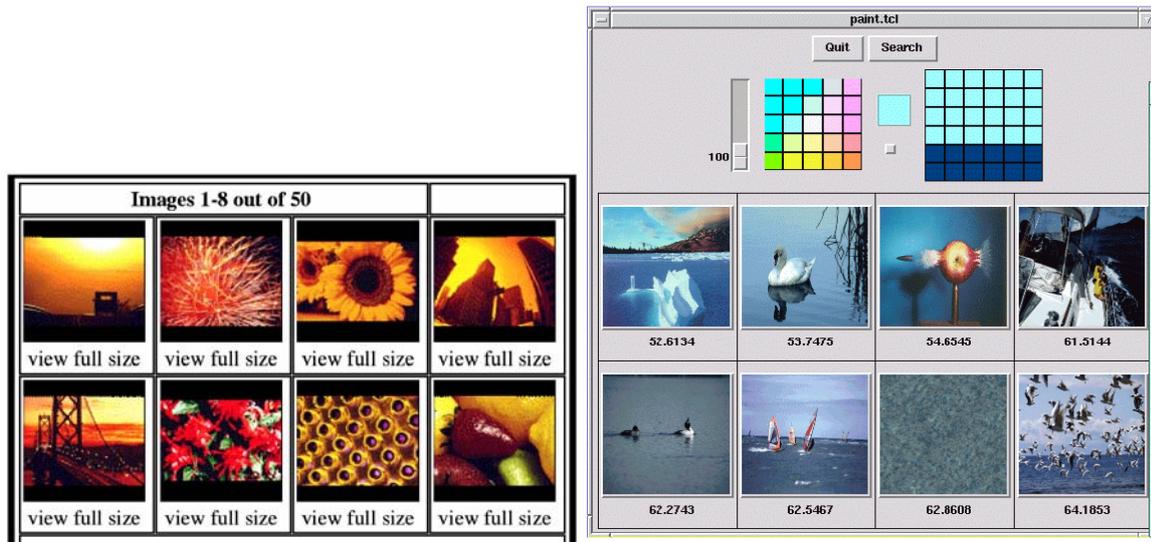
A is $k \times k$ similarity matrix denoting the similarity among different color pairs.

4. Color grid. The distance is defined as:

$$d_{gridded_color}(I, Q) = \sum_g \hat{d}_{color}(C^I(g), C^Q(g))$$

where C is the color in grid square g .

The following two examples show the color histogram matching (left) and color grid matching (right) respectively.



4 Figure 3.2. Color feature matching

Left: matching is based on global color histogram. Right: matching is based on grid color matching. Query is specified in the format of color grid layout.

3.3.2 Texture Feature Matching

Texture is the characteristics of the spatial distribution of gray level among neighboring pixels. The texture features capture the repeating patterns of local variations in image intensity which is too fine to be distinguished as separate objects at the observed resolution. Texture is very important in human perception of the optical characteristics of discrete objects and provides important clue in reconstructing 3-D structure from 2-D image. Thus, it is a topic under extensive investigation for a variety of purposes, including image segmentation, computer vision, and content-based image retrieval.

There are two major approaches in studying the texture property of image: statistical approach, and model-based approach. The statistical approach exploits the statistical properties of image or image regions in a bottom up fashion, starting from the

pixel values in the neighborhood. The co-occurrence matrix is in wide use in representing the dependence in the distributions of gray-level [haralick73]. The co-occurrence matrix is a function of: 1) the image region, 2) a displacement vector $d = (dx, dy)$, and 3) the number of gray-levels after quantization. The matrix contains frequencies of co-occurrence of two gray-levels. After normalization, it becomes a probability matrix with all the elements summing up to 1. Viewing these elements from a signal processing perspective, some feature values can be defined as:

$$Entropy = -\sum_i \sum_j P_{ij} \log P_{ij}$$

$$Energy = \sum_i \sum_j P_{ij}^2$$

$$Contrast = \sum_i \sum_j (i - j)^2 P_{ij}$$

$$Homogeneity = \sum_i \sum_j \frac{P_{ij}}{1 + |i - j|}$$

where P_{ij} is the probability of grey level i co-occurring with grey level j --- after a quantization step.

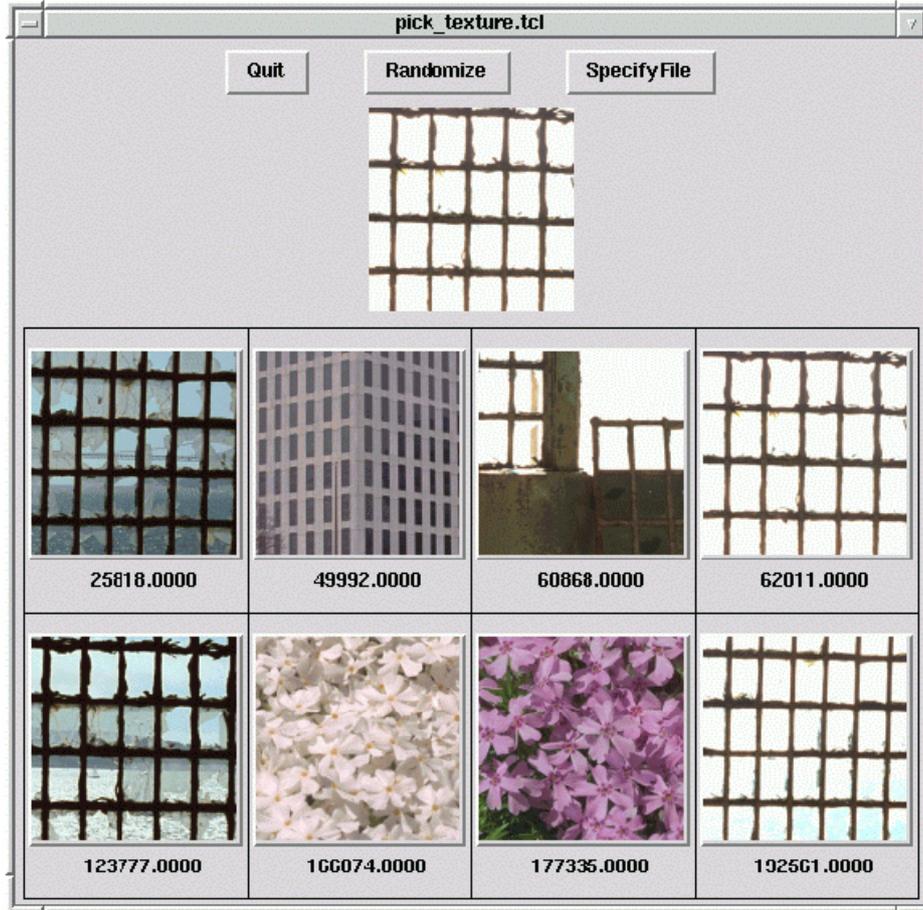
Experiments have suggested that *contrast*, *inverse deference moment*, and *entropy* have the biggest discriminatory power [gotlieb90]. Other features based on the subjective measure of human visual perception have also been defined. Tamura et al. [tamura78] defined six basic textural features as coarseness, contrast, directionality, line-likeness, regularity, and roughness. This system has been used in QBIC system [niblack93] and the MARS system [huang96] [ortega97]. Some researchers used only a subset of the above features including only contract, coarseness, and directionality [liu94] [niblack93]

[niblack97]. Another study identified repetitiveness, directionality, and granularity as the three most important orthogonal dimensions in human texture perception [rao93]. Sometimes, when the image primitives are big enough, it is more appropriate to separate the image primitives first and describe them individually, before texture analysis.

Another way to describe the texture features is through modeling, where a model is constructed with parameters determined so that the model can be used to generate the kind of texture to be described. Markov random field model (MRF) and fractal analysis have been used for this purpose [cross89] [pentland84] [kashyap81]. Taxonomy for texture description and identification can be found in [rao90].

Like color features, texture features can also be used for image segmentation. The generated regions can support different texture features. Recent progresses in image compression techniques have provided new possibilities in image feature extraction. Many of the features that result from the transformation algorithms are related to the image texture. The application of region-based matching and image transformation in CBIR will be discussed later in this chapter.

The figure below shows retrieval based on texture feature [shapiro01].



5 Figure 3.3. Retrieval based on texture feature

A query is specified in the format of a texture pattern. Texture features are extracted from the query and compared with those extracted from the images, while the color differences are ignored.

The similarity measure with regards to the representation is:

$$d_{pick_and_click}(I, Q) = \min_{i \in I} \|T(i) - T(Q)\|^2$$

$$d_{gridded_texture}(I, Q) = \sum_g d_{texture}(T^I(g), T^Q(g))$$

where $T(i)$ is the texture description vector at pixel i of image i , and $T(Q)$ is that of the query.

3.3.3 Shape and Sketch Features

Shape description is also an active research area [li94] [mehtre97]. There are three ways to describe shape feature, boundary-based, region-based, or a mixture of two. Unlike color and texture, which provides information in image segmentation step, correct image segmentation is instead a prerequisite for generating meaningful shape descriptor. The shape descriptors are usually required to be invariant for translation, scaling, and rotation. Fourier descriptors and moment invariants have proved to be successful in the two approaches respectively.

Three mathematical formulations are useful for boundary feature representation, including chain code, Fourier descriptor [zahn72] [persoon77] [gonzalez92], and UNL descriptor [lee97] [jain86]. They describe only the outer boundary of a region while ignoring the area inside. In the former method, the signal that defines the boundary is Fourier transformed and the extracted features are compared in the frequency domain. Following are four commonly used shape feature descriptors:

- a. Unit vector:

$$v_k = \frac{V_{k+1} - V_k}{|V_{k+1} - V_k|}$$

- b. Cumulative differences:

$$l_k = \sum_{i=1}^k |V_i - V_{i-1}|, k > 0, V_0 = 0$$

- c. Fourier descriptors:

$$a_n = \frac{1}{L \left(\frac{n2\pi}{L}\right)^2} \sum_{k=1}^m (v_{k-1} - v_k) e^{-jn(2\pi/L)l_k}$$

d. Fourier distance measure:

$$d_{\text{Fourier}}(I, Q) = \left[\sum_{n=-M}^M |a_n^I - a_n^Q| \right]^{\frac{1}{2}}$$

In all above formulas, V_i is a vector representing a minimal segment of the boundary of the region; a_n^I and a_n^Q in formula d is defined in formula c.

A more relaxed matching than shape feature-based matching is boundary matching, including polygon representation-based boundary description, and elastic boundary matching. In the polygon representation, matching is based on the lengths of the sides of the bounding polygon of an object and the angles between them.

Region-based representations focus on the changes of the area inside the boundary. The descriptors developed along this approach include the early moment invariants [hu62], Zernike moments [jain86], the morphological descriptor [prasad97], and pseudo-Zernike moments [mehtre97]. New methods in shape feature description include finite element method (FEM) [pentland96], turning function [arkin91], and wavelet descriptor [chuang96].

Sketch is another useful feature that usually focuses on the contour as well as skeleton structure of the objects in image [leung02] [leung02-2]. In a query session, the outline of an object can also be specified by the user as a sketch as proposed in [hirata93], and implemented in QBIC [niblack93] [flickner95]. The feature can be derived by edge detection, line thinning techniques, and compared based on the correlation between the

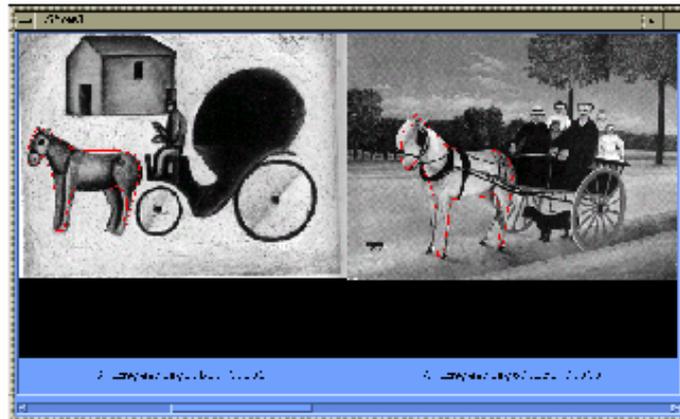
user-defined sketch, and the feature extracted from the image. Following is a formula for sketch-based feature matching and a query example using the matching criteria.

$$d_{sketch}(I, Q) = \frac{1}{\sum_g \max_n [\hat{d}_{correlation}(shift_n(A^I(g)), L^Q(g))]}$$

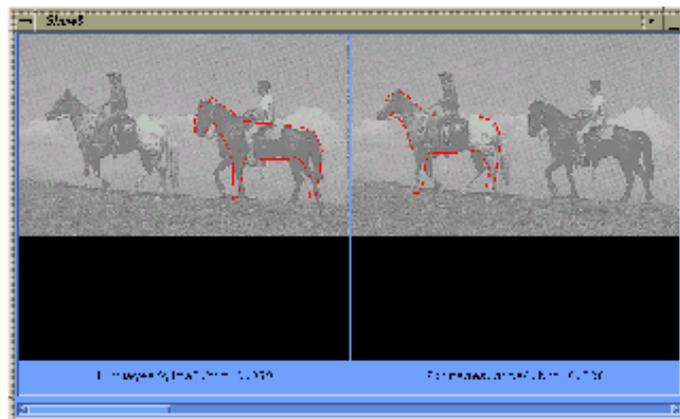
where $L^Q(g)$ and $A^I(g)$ are sketch feature specified in the query and that derived from the image from the database respectively. One of them needs to be shifted and the maximal similarity score is used to determine the distance measure.



a) The user's query shape



b) Two of the retrieved images.



6 Figure 3.4. Sketch matching

The user sketches the outline of a horse as a query. The system needs to identify all the images with at least one object with the same shape. In this implementation, the sketch feature is not orientation invariant. Please notice that the algorithm can tell the rider on the horse apart from the horse and identify the correct contour of the horse. Color and texture features are ignored in this case.

3.3.4 Image Feature Matching in Compressed Domain

Signal transformation algorithms are primarily used for image compression purpose. For example, the JPEG encoding and decoding standard employs the discrete cosine transformation (DCT) algorithm to transform the raster representation in the spatial domain into DCT coefficients in the frequency domain. In this process, the color information is concentrated in one channel thus high compression rate is achieved in other two channels. Besides, high frequency signals are further quantized so as to take less storage space [wallace91]. Wavelet transformation also concentrates signal in a subset of wavelet coefficients thus achieve its compression goal. These advantages of compressed image format have become the incentive for compressed domain manipulation to avoid dealing with the much larger raster format image file. These algorithms provide new types of feature in addition to the three discussed above, although, there are efforts in deriving primitive features, such as texture, from wavelet transformed images [boggess01] [kuo93] [unser95] [wouwer]. Despite of the advantages, the algorithms working in the compressed domain could be more difficult to develop. It is hard to relate the less intuitive coefficients in the transformed domain to the higher level visual perception in the spatial domain. The characteristics of indexing images in the compressed domain have made it a distinct topic, abbreviated as CDI (compressed domain indexing). It usually covers transform domain techniques such as discrete Fourier transformation (DFT), discrete cosine transformation (DCT), Karhunen-Loeve transformation (KLT), sub-band and wavelet transformations, and image vector quantization. Good review papers that cover CDI include [mandal97] [mandal99].

a. Discrete cosine transformation

DCT is closely related to the discrete Fourier transformation (DFT) but with certain advantages in handling signals with finite elements and computational simplicity. With an information packing capability close to the optimal KLT, and with advantages such as all the coefficients guaranteed to be real numbers [gonzalez92], DCT is also the basic compression algorithm underneath the MPEG 1, 2 H.261/H.263 standards. Indexing JPEG format images is desirable for storage, computation efficiency, and availability of JPEG images on the web. Similar researches in image/video manipulation and feature extraction have made progress, and the algorithms developed have been used in real-time video processing, motion estimate, and the signal processing circuit in color TV set [koc95] [liupatent98].

One way to handle DCT compressed images is to reconstruct the image from the DCT coefficients (IDCT: inverse discrete cosine transformation) either completely or partially using only a small subset of the coefficients (partial IDCT) so as to save processing time [armstrong01]. Better yet, image features can be derived directly from the DCT coefficients [feng02] [jiang02] [jiang02-2]. A very simple statistical modeling of the DCT coefficients by computing the means and variants of the AC components followed by Fisher discriminant analysis was used in [smith94] [reeves97].

Another technique that exploits the DCT coefficients is to use the within-image block-wise correlation of the coefficients in the query image or the target image as a key. The overall image similarity is the similarity of keys from both images [shneier96].

However, in that paper, no semantic significance could be established corresponding to the similarity assessment. DCT coefficients have also been used in face recognition. In [pan99], the information capacity of only a few (as little as 35 from a human face image) DCT coefficients from a facial image is demonstrated, and the effect of the block size on the sample images analyzed. Artificial neural network is then used as a supervised learner to generate a classifier, with the selected DCT coefficients as input, and the object ID as output.

Edge detection can also be simulated in the DCT transformed domain. Different subsets of DCT coefficients measure the gradient in the change of the intensity in vertical, horizontal, and diagonal directions, respectively. A straight line of slope m in spatial domain is represented in DCT transformed space by a straight line with a slope of approximately $1/m$ in the DCT domain [ng92]. The approximation of more features with DCT coefficients was described in [shen96], covering edges, edge orientation, edge offset from center, and edge strength. Due to the fact that DCT is also the underlying coding scheme for video, there is a strong interest in it in the content-based video retrieval community. The topic will not be covered any further in this writing.

b. Singular value decomposition

Singular value decomposition (SVD) or under another names as Karhunen-Loeve transformation (KLT) or principal component analysis (PCA) or latent semantic indexing (LSI), has been applied successfully in face image classification. The feature extracted by KLT/PCA/SVD/LSI method is called eigen face [pentland94]. It is a method for

projected to the corresponding eigenvectors, while UA give coordinates of points in that axis. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix. When performing dimensionality reduction, the smallest eigenvalue correspond to the least significant features, which can be removed while conserving the most discrimination power.

SVD has special implication in content-based image retrieval. If we use each row in vector A to represent an image feature vector, with every column representing one feature, then the following can be inferred:

1. The value K represents the number of concepts in these images;
2. The elements of each row in U represent the membership strength of the corresponding image to the respective concepts; U is an “image-to-concept” similarity matrix.
3. The diagonal elements in W are the representation of the concepts in these M images; W is a “concept strength” diagonal matrix.
4. The elements in V are the weights of the features that contribute to the classification decision on an image bearing an individual concept; V is a “feature-to-concept” similarity matrix; the columns of V are called eigenvectors, with the most prominent ones appearing on the left most columns.

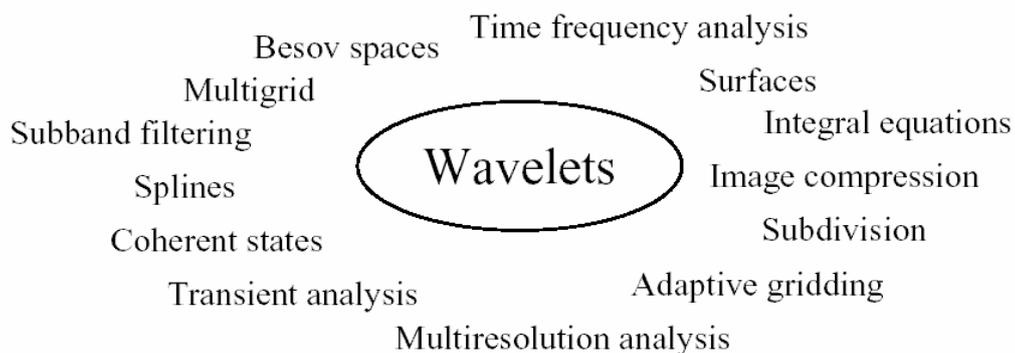
The complexity of SVD is $O(N*N*M)$, but can be reduced if only eigenvalues are needed, or if only first a few eigenvectors are needed, or if the matrix A is sparse. The algorithm is available in most linear algebra software packages such as LINPACK (and its modern version, LAPACK, <http://www.netlib.org/lapack/>), MatLab, S-plus, and Mathematica. The application of SVD can be found in multi-lingual IR, compression, PCA (“ratio rules”), Karhunen-Lowe transform, query feedbacks, and the recent google/Kleinberg algorithms [fukunaga90] [press92, chap2] [<http://www.cs.utk.edu/~lsi/>].

c. Discrete wavelet transformation

Discrete wavelet transformation (DWT) has gained wide popularity due to its higher compression ratio, its hierarchical representation structure that allows progressively transmission of images of different magnification, the localization information it carries and its adoption as a part of the JPEG 2000 standard. Data sets without obviously periodic components cannot be processed well using Fourier techniques. For example, the United States FBI compresses their fingerprint data base using wavelets (<http://www.c3.lanl.gov/~brislawn/FBI/FBI.html>).

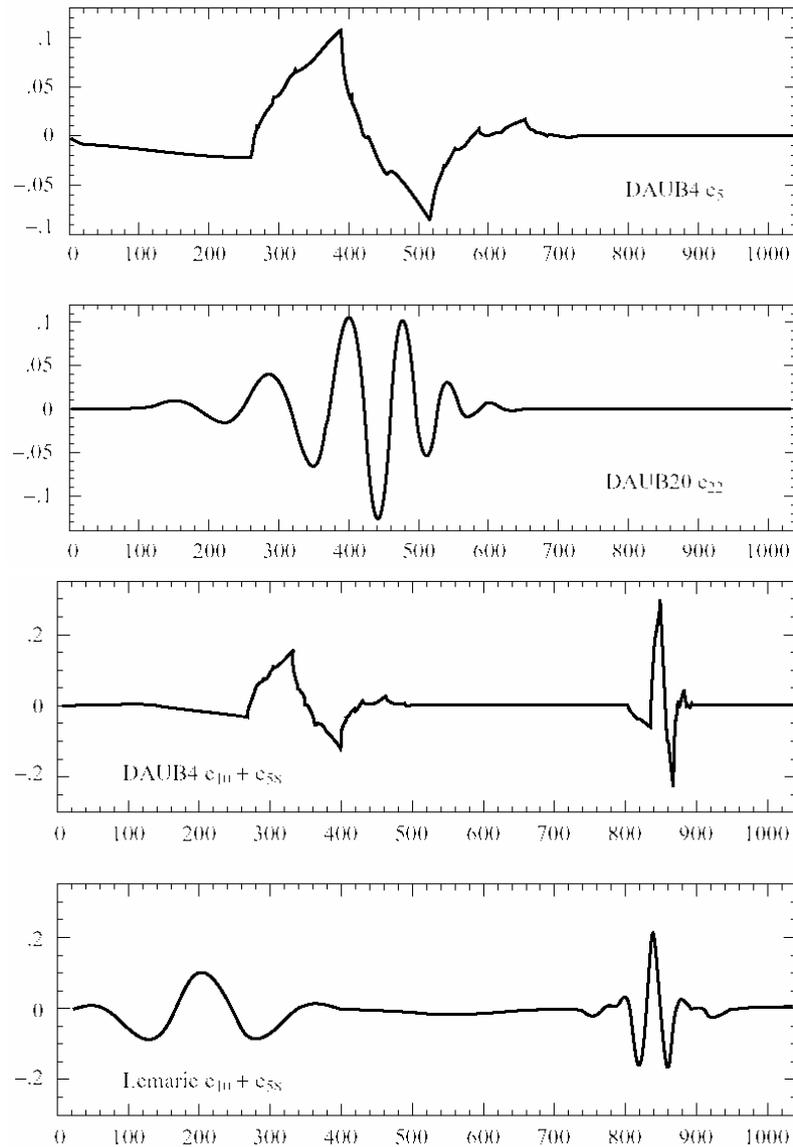
Like Fourier transformation and other signal transformations, wavelet transformation also involves convolution of target signal and base functions that satisfy certain restraining conditions. Unlike DCT, which needs to divide images into small blocks, DWT transformation processes over the whole image, thus avoids the blocking artifact. At least, dividing image into smaller blocks is not practically required for any reason. The coding bit rate is also better adapted to the change of the image signal over the

complete image, rather than assigned evenly as with DCT blocks. With DWT, image is passed through a filter that is defined by a base function, and decomposed into a low-pass component and a high-pass component. The low-pass component is then passed through the same filter again, and the process goes on recursively [mallat89] [antonini92]. It has also been shown that DWT is better adapted to the signal by choosing different base functions. There are an infinite number of base functions to choose from, while making the tradeoff between how compact they can be localized in the space and how smooth the functions are. A good base function should resolve more wavelet coefficients in the low frequency sub-bands. For this reason, Daubechies' wavelets are preferred over the earliest Haar's wavelet in processing 2-D natural images, for its continuous derivatives provide better modeling of continuous functions of image [wang98]. But Haar's wavelet still has some usage.



7 Figure 3.5. Wavelet popularity

Many areas of science, engineering, and mathematics have contributed to the development of wavelets (Wim Sweldens and Peter Schröder, “Building your own wavelets at home.”)



8 Figure 3.6. Wavelet functions

(Press92, Numerical Recipes, Chap13)

- Left: single basis functions from the wavelet families: upper: DAUB4: inverse DWT of a unit vector in the 5th component of a vector of length 1024; lower: DAUB20: inverse of the 22nd component.
- Right: sum of two unit vectors, $e_{10} + e_{58}$, which are in different hierarchical levels of scale, and at different spatial positions: upper: DAUB4 wavelets defined by a filter in coordinate space; lower: Lemarie wavelets defined by a filter written in Fourier space.

The (one dimensional) wavelet transformation is defined in its continuous form by the following formula:

$$CWT_x^\psi(\tau, s) = \Psi_x^\psi(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(t) \psi^*\left(\frac{t-\tau}{s}\right) dt$$

where:

τ is the translation factor;

s is the scaling (or dilation) factor;

$\Psi(t)$ is called the mother wavelet;

ψ denotes the wavelet base function;

x denotes the test signal.

According to the above definition of the inner product, the CWT can be thought of as the inner product of the test signal with the basis functions, and takes the form as following:

$$CWT_x^\psi(\tau, s) = \Psi_x^\psi(\tau, s) = \int x(t) \psi_{\tau, s}^*(t) dt$$

where:

$$\psi_{\tau, s} = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right)$$

(Robi Polikar, The Wavelet Tutorial, the engineer's ultimate guide to wavelet analysis, <http://engineering.rowan.edu/~polikar/WAVELETS/WTtutorial.html>). In a space

defined by orthonormal vectors, the coefficients can be calculated using the following formula:

$$\mu_k = \langle f, \phi_k \rangle = \int f(t) \cdot \phi_k^*(t) dt$$

where:

$\langle f, \phi_k \rangle$ denotes inner product of the two functions: signal function and the base function;

ϕ_k are orthonormal vectors of that space, where $k = 1, 2, \dots, N$, with N as the dimension of that space; the orthonormality holds when the following equation is satisfied:

$$\int_a^b \phi_k(t) \phi_l^*(t) dt = 0 \quad k \neq l \quad (\text{orthogonality condition})$$

and

$$\int_a^b \{|\phi_k(t)|\}^2 dx = 1$$

or equivalently:

$$\int_a^b \phi_k(t) \phi_l^*(t) dt = \delta_{kl}$$

where the Kronecker delta function is defined as:

$$\delta_{kl} = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases}$$

In the discrete version, the signal is decomposed into different frequency bands by successive highpass and lowpass filtering of the time domain signal. The low frequency component can be sub-sampled by 2, according to Nyquist's rule:

$$y[n] = \sum_{k=-\infty}^{\infty} h[k] \cdot x[2n - k]$$

The process is called lifting scheme and generates pyramidal representation. The two filters satisfy the following equation:

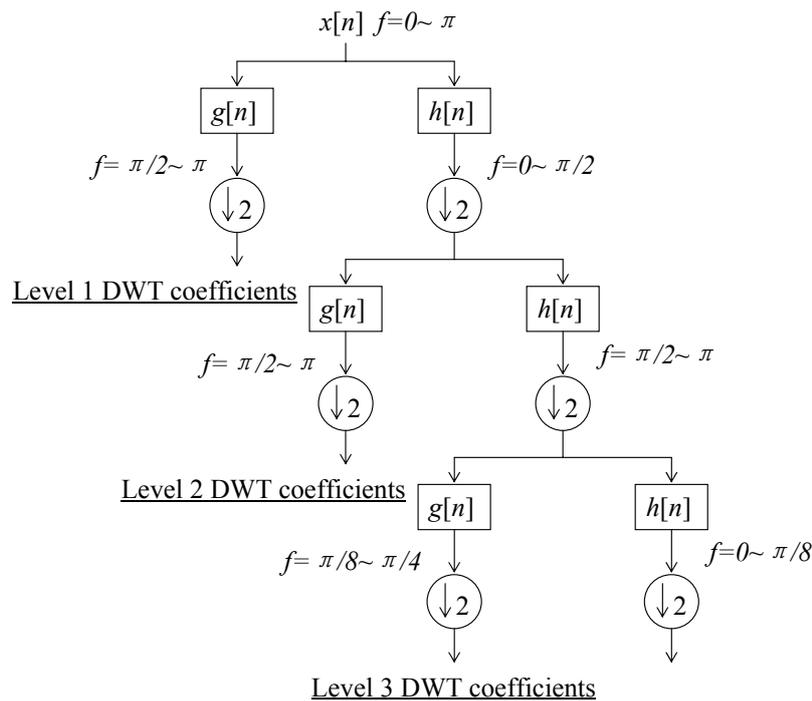
$$g[L - 1 - n] = (-1)^n \cdot h[n]$$

where $g[n]$ is the highpass, $h[n]$ is the lowpass filter, and L is the filter length. This kind of orthogonal filters is known as the Quadrature Mirror Filters (QMF). The filtering and sub-sampling operations can thus be expressed by:

$$y_{high}[k] = \sum_n x[n] \cdot g[-n + 2k]$$

$$y_{low}[k] = \sum_n x[n] \cdot h[-n + 2k]$$

The lifting scheme is shown in the following diagram:



9 Figure 3.7. Diagram of Haar's wavelet transformation

Haar's wavelet is the first wavelet, and the simplest wavelet base discovered by Haar, in which the highpass filter is a differencing function and the low-pass filter is an averaging function. Daubechies' wavelets are a group of wavelets with base functions of similar properties. DAUB4 is one of them with relatively compact support.

The wavelet transformation's spatial localization property and the multi-resolution construction of the pyramidal model make it a very popular scheme for texture feature classification. Much work has been done towards this direction [chang93] [smith94] [chen94]. DWT has shown superior performance over other compressed domain or pixel level processing for texture classification [smith94]. Hidden Markov model (HMM) has

been applied to exploit the co-occurrence patterns among wavelet sub-bands. The texture features extracted from different wavelet sub-bands were analyzed and modeled separately. Image comparison and retrieval decision is based on the comparison of parameters of the models [chen94]. A more common approach to classify texture features based on wavelet sub-band coefficients is through histogram analysis [mandal96] [mandal99]. Gabor wavelets were used in multiple directions to achieve better directional discrimination [manjunath96]. Correlation of the spatial domain features and the wavelet coefficients was induced in [wang96]. After careful down-scaling of the images, the wavelet coefficients extracted from the smaller, normalized version of the images were used for direct comparison [jacob95] [wang97]. Other efforts in achieving rotation, translation, and scaling invariance have seen various success [qi94] [rashkovskiy94].

d. Other image transformation techniques

Other image transformation/compression/decomposition methods are used for image content description and feature extraction. Discrete Fourier transform is closely related to DCT and sees limited use in image feature extraction:

$$H(\omega) = \sum_j c^j e^{ij\omega}$$

Vector quantization (VQ) can usually achieve very good compression ratio based on the fact that it encodes vectors (or data structure of higher dimensions) rather than scalars. It has been used to code both grey level and color image content for content-based retrieval

purpose [vellaikal95] [idris96] [lu99]. At the beginning of the coding process, a codebook, which consists of the most representative pixel block-wise patterns (or patterns in the processed features) in the image, is generated either manually or through unsupervised learning [kohonen95]. Then, image blocks or regions are coded with indices to the most similar blocks. The code set is used as the vocabulary of the images. Techniques such as histogram, vector space model-based distance measure, and co-occurrence measure can be combined with the image vocabulary analysis. The whole scheme bears much similarity with the text retrieval paradigm using indices as counterparts of indexing terms in text retrieval, with one exception that two dimensionality property is important for image in forming image layout, texture, etc, which relates to the one dimensional term proximity in text. The idea of vector quantization plus the concepts and retrieval models (plus co-occurrence analysis) from text retrieval was formalized in [zhu01] [zhu02]. The authors named the approach "keyblock-based image retrieval" in a series of their publications. Besides being applied directly to the image pixels, these transformations are also used on generated image features for the purposes such as dimensionality reduction, data structures for fast search and retrieval.

3.3.5 Composite Image Features and Region Based Matching

As the computation power becomes more accessible and cheaper than ever before, a CBIR system tends to use more than one feature and more than one technique, and the retrieval decision is based on the summation of the contributions from all the features so as to

achieve better accuracy and accommodate more types of image content [antonini92] [li97] [idris95] [swanson96] [podilchuk96].

A further step beyond image feature extraction and indexing is image understanding that bridges the semantic gap between machine vision and human visual perception. ("Semantic features aim at encoding interpretations of the image which may be relevant to the application.") Usually, understanding an image requires clear identification of the objects in the image, and the spatial relationships among them. This further requires accurate image segmentation to separate the objects, which is, despite of the successes in the restrained domains [podilchuk96], in a broad sense, an unsolved problem. On the other hand, it is usually not necessary to carry out an accurate segmentation that captures individual objects, if we can get enough information about the object from other, more general image features that can be related to the objects [smeulders00].

Along the track towards object feature extraction and image understanding, but with a fuzzy, non-deterministic approach, a region is a rough match of an object. Without ignoring the structural grouping of the pixels, a region is usually formed by homogeneous properties such as color, texture [gevers00] [wang98]. Different from the approaches of constructing histograms and exploiting statistical parameters such as mean, variants, levels of moments, and treating images as "bags of pixels", the pixels in an image are first grouped to form regions according to their spatial proximity and homogeneity in intensity. Then the layout of multiple regions in the image along with the regions' individual properties is combined to form a measure of the similarity between images. One of the possible representations of the spatial relationship is attributed relation graphs (ARGs) as

described in a textbook [chang89] and a recent paper [petrakis02imagemap]. Due to the distortions in the shapes, shift in the color, variations in the relative positions among objects, etc, of the object/region in the image, fuzzy logic is often used in combining multiple features rather than deterministic measures. One of the earlier region-based retrieval systems used color and texture as region defining features. EM was then used to cluster pixels and/or image blocks into regions [carson97]. WALRUS (WAVElet-based Retrieval of User-specified Scenes) was wavelet/region based, in which wavelet features from small sliding windows were used to cluster the windows to form regions. Regions were indexed and the index adopted an R*-tree structure to increase retrieval speed. Similar regions from two images were compared in pairs [natsev99]. In the WindSurf (Wavelet-based INDEXing of imageS Using Region Fragmentation) system, wavelet coefficients were clustered with k-mean, regions with homogeneous features were identified, and spatial information extracted. Then a distance measure was derived, which was later coupled with an R*-tree like distance-based access method (DBAM) for indexing and fast retrieval [ardizzoni99] [bartolini00]. The wavelet features and *k*-mean clustering method were again used in a later research in [chen02]. Image segmentation was achieved by dividing an image into blocks and the blocks were grouped by *k*-mean based on their low frequency wavelet coefficients in LUV color space. Each image was coded with a *k*-dimensional vector representation by k-mean method, based on the kinds of blocks used to cover the complete image. The concepts of fuzzy feature and fuzzy model for fuzzy feature matching were defined. Further fuzzy feature vector matching was based on the fuzzy

model. Another recent paper that adopted a similar wavelet-based texture analysis approach is [suematsu02].

Bayesian framework is another approach adopted by some researchers [vasconcelos98] [vasconcelos00]. It has also been extended to process users' relevance feedback under the same framework [su01].

3.3.6 Conclusions

From the above brief survey of a variety of features used for CBIR, a few trends are obvious:

1. More features are under investigation than before.
2. New features are more complex than the simple color, texture schemes.
3. Combination of multiple features, and combination of processing techniques in processing one features are common in defining an image or a region in an image.
4. Compressed domain processing, such as wavelet coefficients and DCT coefficients, has drawn wide attention and gained popularity.
5. Machine learning/data mining techniques are used in extracting sensible image features
6. With the increase in the complexity of the constitution of the feature sets, and the realization of the uncertainties in the image feature processing, researchers start to exploit mathematical/statistical models to support better retrieval decision making, instead of resort to simple classification rules. The incorporation of other information retrieval techniques, such as relevance feedback, query expansion, will

take more advantage of those theories and techniques developed in text retrieval but applicable to general retrieval problems. This will be further discussed in the next chapter.

3.4 FEATURE WEIGHTING, FEATURE SELECTION, AND RELEVANCE JUDGMENT

With multiple features or single multi-dimension feature, an optional step that improves the features' distinguishing power is to provide a weighting scheme so that relevant, high quality features can have more contribution to the retrieval decision than not so relevant, noisy features. Traditionally, the measure of the similarity of two feature values can be adjusted by the statistical distribution. The Mahalanobis distance is such an example, and it has been used in CBIR as well as many classification tasks. The different color channels are weighted by representing the color in different color models, which have different properties and information concentration in three channels [wallace91]. However, data-dependent feature weighting is not commonly seen. Feature bagging and selection are usually based on empirical data. This is commonly seen with compressed domain feature processing, in which a small subset of all DCT and wavelet coefficients are used as features and the rest majority are thrown away. In other cases, some of these coefficients are grouped together so that only a few concentrated feature values need to be dealt with. Due to the general absence of gold standard image databases to facilitate supervised learning, most learning process targeted to assign weights for the features are either

unsupervised or empirical. However, one exception is seen in the user-system interaction process when the feature weights can be tailored to users' information needs, based on the users' relevance feedback on a per query session base. The topic of relevance feedback will be discussed in the next chapter.

Feature selection is related to the issue of feature weighting but with an objective to reduce the dimensionality of the feature set. A systematic approach is to eliminate the feature that has the least impact to the correct classification. The process is iterative until a threshold error rate is reached and the process can not be carried on any more, or the set has reached a satisfactory number of dimensions, and the process stops voluntarily. Another opposite approach is to select the most significant features first. This is usually preceded by a transformation of the feature space. Many of the image transformation methods surveyed above can be used here for this purpose. The most important method that has been reinvented in solving many similar problems is KLT/PCA/ SVD/LSI, which stands for Karhunen-Loeve Transformation/Singular Value Decomposition/ Latent Semantics Indexing. The method is highly efficient in identifying only a few feature values while retaining most of the distinguishing information.

Various information retrieval models take the features prepared from above procedures as input, and project the image features from the feature space to the concept space, so that the retrieval decision can be made based on the models' individual relevance judgment:

1. The presence or non-presence of the features satisfies a first order logic (FOL) that the query specifies. The Boolean retrieval model is not very useful in image retrieval as the image features are usually numerical and the Boolean representation of the features loses the information.
2. The document is in proximity of the query in the concept space, according to a distance formula with feature values as variables. The vector space model is most widely used in text retrieval, and is also favored by the image retrieval community for its adaptability to features of different natures.
3. There is a certain probability that the document may be relevant to the query based on the features it has: $P(\text{Rel} \mid \text{query}, \text{document}) > \text{threshold}$. The probabilistic model is theoretical sound, gives quite good retrieval accuracy in text retrieval, and provides a sensible way to unify features of all kinds of nature under the general scheme of probability of relevance. Yet, works need to be done to make it feasible to deal with image features.
4. Fuzzy model is very similar to the probabilistic model but without the constraints from the probability theory, such as all cases of an instance should sum up to 1. The implementation is quite similar to the probability model. It was recently used in [chen02].

There exist other retrieval models that shine in the domains that fit them. The kind of retrieval model in use plays an important role in the retrieval decision making process and the way how the image features should be represented.

4.0 SYSTEM IMPLEMENTATION AND PERFORMANCE EVALUATION

This chapter covers three major topics that are not directly related to the image classification but are important to the system performance and evaluation: 1) data structure for feature indexing, 2) relevance feedback, and 3) evaluation.

4.1 FEATURE INDEXING

There are two important aspects in implementing a good CBIR system: effectiveness and efficiency. Effectiveness emphasizes on the function of the system to satisfy users' information needs, which is mostly concerned by the retrieval accuracy. Efficiency is the performance-wise characteristics that allow the retrieval tasks to be accomplished in a predictable amount of time with reasonable resources, and the scalability of the design.

After proper features have been identified, an image document surrogate that is used for direct comparison is usually a feature vector. A Boolean example is an extreme case where each element of the vector takes a value of either 1, or 0, or undefined, indicating whether a feature is present, non-present, or undefined. A signature file can be generated for the database with K binary vector, where K is the number of features. Each vector is N bits long, with N as the number of total documents in the database. The bit i in

vector j indicates the value of i -th feature in j -th document. The retrieval process is as simple as Boolean AND operation: all binary vectors corresponding to the query terms are aggregated with Boolean AND operation, documents that have 1 in the resulting vector are qualified for retrieval.

However, binary features are not commonly used and provide harsh granularity in feature quantization. When feature values are scalars, an image is more likely to be represented as a point in a high dimensional feature space. The problem of retrieving similar images is then transformed into the problem of searching for points within proximity of the query in the high dimensional space. Unlike the case in Boolean kind of feature and query, the comparison between images gives a similarity (or dissimilarity) value. Users are usually interested in related images with similarity value above a certain threshold. This translates into two kinds of queries, nearest-neighbor query (k -NN) and range query. k -NN asks for the k images that are most similar to the query, while range query asks for the images with certain similarity score. The easiest solution is sequential search that compares every database image with the query. This approach doesn't require any special data structure and search strategy, and is used by small systems and test-bed with small collection of images.

With a large collection of images to search through, the performance of sequential search is not acceptable. Tree-like data structures are used to break down the search complexity. For a single feature, or features that can be mapped to one single index, many conventional search algorithms are available, such as B-tree. Insertion, deletion, searching, and concurrency issues have been well studied. For feature space of higher dimensionality,

multi-key indexing methods are often used. Quad-tree is especially useful in decomposing image into regions hierarchically [smith94] [vellaikal95] [chang95] [lin97] [natsev99]. K-D tree is designed to search in multi-dimensional space with multiple attributes [white96].

Spatial access methods (SAMs) are file structures for managing high dimensional points in large scales, either in the main memory or on the disk [gaede98]. One of the strengths of SAMs is that it provides a means to process range query. R-tree is such a SAM method that can be viewed as an extension of the one dimensional B-tree. In R-tree, multi-dimensional objects (such as a range, nearest neighbors of a query) are represented by their minimal bounding rectangle (MBR). Intermediate nodes contain the definitions of range and pointers to the corresponding child nodes, while leaf nodes have the object ID and the range defining information of the MBR of that object. The intermediate nodes of an R-tree are allowed to overlap with each other. In particular, range queries and nearest neighbor queries are easily represented as such objects in R-tree [guttman84] [brinkhoff93]. R-tree has been used with images with multiple objects/regions that have been successfully segmented and their spatial relationships clearly modeled with ARGs or editing distance [petrakis02fast] [petrakis02ImageMap]. Variants of R-tree include R+-tree [sellis87] and R*-tree [beckmann90].

As in text retrieval, image features could also be indexed to build inverted indices to facilitate locating images with particular features [voorhees86] [squire99] [squire00]. The indices are usually built as B-trees or hashing tables. Images can also be clustered together to facilitate retrieval and browsing.

4.2 USER-SYSTEM INTERACTION: RELEVANCE FEEDBACK

Another goal in CBIR implementation is to provide interaction between the system and the users, during which the users can express their information needs through an iterative refining process. Due the existence of "semantic gap", low level features alone are not sufficient in defining image content in a general scope. The input of the users' assessment to the retrieved images in a retrieval session helps to adjust the users' specification of the query and/or the behavior of the system, to provide a context for how the image features should be interpreted. One of the most important user-system interactions in information retrieval is relevance feedback, first developed for vector space model in text retrieval [rocchio71] [buckley95]. After a query session is complete and the system returns a list of relevant images, the user can review the retrieved images and give assessment of their relevance back to the system. With this information, the system carries out another session with added input in the hope that better retrieval accuracy can be achieved. The process goes on iteratively until the user gets what he/she wants.

There are three ways in which users' relevance feedback can be used to improve retrieval accuracy: through query refinement, or through feature weighting, or through a mixture of both. In the first approach, a new query is formed based on the old query by enhancing the feature components that are significant in the retrieved images that are classified as relevant, and at the same time, decreasing the ones from the non-relevant [rocchio71]. An important method of this approach in text document retrieval is query expansion.

The second approach is to adjust the feature weights in the favor of those in the relevant images, against those present in the non-relevant ones. In the IRIS (Interactive Retrieval of Images System), the variance of a feature is used as a measure for the importance for classification -- the feature receives a lower weight if it sees big variance across the images that are considered relevant according to users' feedback, or if it sees the same level of variance both in the relevant and non-relevant image groups [yang98]. In [rui98], the weight adjustment is based on the ranks of the relevant images and non-relevant images along each feature axis. MindReader adopts the Rocchio formula from the text document retrieval [ishikawa98] and refines query iteratively.

Heuristic methodology described above were mostly adaptations from the text document retrieval. Recent progresses have been made in the direction of viewing the user-system interaction as an optimization, learning, classification, or density estimation process, the goal of which is to improve the classification based on a small number of examples [rui00]. In PicHunter, a "stochastic comparison search" was proposed to search for the desired images with feedback as "relevance judgment" [cox98]. Boosting technique is used to learn the classification function from 45,000 features in [tieu00]. DCT coefficients are modeled in a Gaussian mixture model, and then Bayesian inference is applied for region based matching and learning in [vasconcelos00]. Self-Organizing Map (SOM) is used to group relevant and non-relevant images based on the user feedback in [laaksonen00]. Other important learning methods that have been used include support vector machine (SVM) [chen01]. Expectation-maximization (EM) is used in [yong01]. [huang01] has a good review on various approaches. Besides, [su00] introduces a Bayesian

framework to handle relevance feedback assuming different context for positive feedback and negative feedback. In [bang02], instead of adjusting the weights of the features or refining the query, the feature space is warped, i.e. the features of the database images are changed in order to reflect the relevance judgment in user's feedback. In each iteration, the positive images are brought closer to the query, while the negative images are moved away from the query. This effort is expected to correct the errors in the feature identification process.

4.3 PERFORMANCE EVALUATION

The evaluation of a CBIR system can focus on different components and aspects, such as ease of query formation, speed of retrieval, required resources, document presentation, and the ability to find relevant documents. The most important and common evolution measure is the retrieval effectiveness, which turns out to be a difficult task for image retrieval. First, a relevant image is one that is judged useful in the context of a query image. The judgment varies from person to person. The usefulness is also hard to define. Sometimes, an image database might be used to answer queries that are not expected to be common. The human judges are not consistent throughout the time, and from one judge to another. The retrieval model the system adopts also comes with its intrinsic definition of relevance, be it a FOL statement, a distance measure, or a probability estimate. In short, judgment depends on more than document and query. In image retrieval, we are yet to establish standard test sets of images for evaluation/comparison purpose. Current, researchers usually use one of

the test image sets that are easily accessible from the web, such as the one at <http://corel.digitalriver.com/commerce/photostudio/catalog.htm>, which includes 70.000 images in 500 categories. In real world cases, the complete set of relevant images from an image database is hard to examine individually.

Performance of CBIR is typically evaluated by empirical visual inspection as seen even in some publications, where query images and the respective retrieved images are laid out side by side. When necessary, two metrics, recall and precision, are used following the tradition of text retrieval. Precision is defined as the proportion of a retrieved set that is relevant; recall is defined as the proportion of all relevant documents in collection that also appear in the retrieved set. In ranked retrieval, precision and recall are displayed in the form of P/R curve [squire99] [squire00] [zhu02]. This is an easy way to compare the performances of different systems/algorithms over the same image database. However, evaluation centric works are scarce in the literature. A discuss of various CBIR evaluation issues in the light of text retrieval evolution methodology can be found in [müller01performance] [squire01design]. The important issues discussed include standard benchmark test image collections, relevance judgments, single-valued metrics, and graphical representations. The proposed methods were later used in [müller01automatic].

There is an effort to put together a CBIR benchmark portal on the web at (The benchathlon Network, home of CBIR benchmarking, <http://www.benchathlon.net/>) with a standard test image collection. The evaluation methodology is mostly borrowed from text document retrieval side. The metrics include ranks of the first match and average rank,

precisions at 20th, 50th, and of all the relevant images, recall with half precision scores 0.5, and finally, the P/R plot [müller01automatic].

A very recent work takes a very different perspective than above. Various application scenarios of image retrieval are discussed in the light of the retrieval task, users' very special information needs, and the interactions between system and users [jermyn02].

A new approach targeting at verifying the distance measure of CBIR engine within a domain context was used in [zheng03]. The pair-wise similarity scores among selected sample images from a medical image database were used for evaluation rather than the precision/recall scores to take advantage of the numerical output directly from the search engine. Clustering algorithm is applied to the similarity matrix to build a dendrogram, which was then compared against the medical taxonomy to ensure correct classification. This approach helps in difficult situations where massive case-by-case evaluation is prohibited due to insufficient relevant images in the database (one sample image per category might be all that is required for clustering algorithm), but domain knowledge can provide a gold standard. The complexity of the retrieval problem was then estimated using multi-dimensional scaling.

5.0 METHODOLOGY: COLOR QUANTIZATION OF PATHOLOGY MICROSCOPIC IMAGE,FEATURE EXTRACTION, AND CONTENT INDEXING

5.1 BACKGROUND

Automated image classification and content-based image retrieval (CBIR) have attracted wide attention from academic research as well as from commercial development. However, the most important driving force behind the strong trend is the rapid accumulation of digital images in a few particular domains, such as satellite imaging, medical imaging, and other special purpose picture archiving and communication systems (PACS). The digital images generated by modern imaging devices in these special sections have virtually run out of human experts' manual indexing capability. Understanding and solving problems in these particular domains requires special knowledge and careful research effort, and the result can be better adapted to that domain and lead to superior solutions to the real world problems. Domain context also provides a rich and stable resource for identifiable image features and a ground truth knowledge base for performance evaluation.

Pathology diagnosis, as a tradition, heavily relies on morphological analysis of microscopic images. Diagnosing an image in many ways resembles the process of classification, in which, each sample is labeled with one or more class names. The goal of

content-based microscopic pathology image retrieval is to mimic the intelligence of human pathologists as a classifier. A well-designed retrieval algorithm is expected to retrieve the images that bear the same class label as the query image. What is more, this classification-based retrieval algorithm is also expected to “understand” the taxonomy of the class labels and the histological/pathological phylogeny of the different tissues so that the tissues with similar origin and morphological structure will also be ranked higher than the other tissues [zheng03]. Hence, the relevance judgment of content pathology microscopic image retrieval is based on classification correctness. A positively retrieved image should bear the same class label as the query image. In this design, all images are collected from diagnosed cases, and each labeled with a short text descriptor. As a matter of fact, the images within the same category usually have very consistent morphological appearance. A good retrieval algorithm should focus on these common features and generate a relatively small within-group variation in classification. This indicates that a relatively small population of each category may be sufficient to achieve statistical significance in the evaluation process.

There are two major objectives and approaches for image feature extraction: 1) to identify strong features that are directly associated with a kind of object/concept of interest; 2) to pack/concentrate image information and to identify weak features, and at the same time, remove noises in order to allow further image data mining for discovery of the statistical correlation between the features and the concepts of interest. Modern content-based image retrieval techniques that deal with images of multiple different classes often resort to the latter approach, for it is inefficient or near to impossible to identify individual features for every potential classification in a large number of, and sometimes unexpected,

image collection following the first approach. General features allow for more entropy reduction than those features that are present only in a few particular cases, and thus are potentially more discriminative.

Principal component analysis (PCA) is one of the most important and powerful methods in feature extraction and information packing. Techniques based on the same mathematical formulation include KLT, SVD, and LSI under different names in different domains. It provides optimal information packing capacity that usually leads to significant reduction of problem complexity without significant loss of information. A data point can be represented by its projections on a few eigenvectors and still conserving 80-90% of the total information. However, the process is un-supervised and the technique is data dependent rather than classification driven. That means the derivation of the eigenvectors is affected by the statistical distribution of the sample data, rather than taking into account the classification of the data. The process is optimal in data representation, but may not be optimal for class representation/classification purposes. As a result, the dimensionality reduction may neglect features that are critical for classes with fewer representation of the total sample population, and in some cases, blind dimensionality reduction is harmful to classification.

Vector quantization has been used for image compression and the method usually gives a compression ratio much higher than other general-purpose compression methods. This is due to two important facts: 1) it is possible to encode a vector of features or a block of pixels (or coefficients if the image is in its transformed format already) at a time rather than a scalar value -- either an individual pixel or a single coefficient; 2) the codebook can

be learned from the images to be coded thus covers a reduced code space with improved accuracy. In some sense, the process bears some similarity with PCA in its first stage. The coded blocks can be treated as terms or keywords of an image document as in text information retrieval, and a rich set of proven feature extraction techniques, distance metric, and retrieval models can be employed to accomplish the retrieval task in the discrete domain. The fact that all the subsequent processing, feature extraction, content indexing are carried out in the discrete domain brings about reduction in the algorithm complexity and gain in processing performance, especially on modern computer hardware with longer word length, and more random access memory (RAM). Some success has been reported along this track [idris95] [lu99] [zhu00] [zhu02]. However, there are several drawbacks to this new image coding scheme: 1) the arbitrary dividing of the image into small, square blocks makes the transformation not robust against various affine transformations; 2) the effectiveness of vector quantization for the purpose of image content indexing is thus questionable; 3) vector quantization is very computation intensive in both codebook design and code vector searching.

But in implementation, VQ still suffers from significant computation overhead in generating the codebook, which is NP-hard in its non-heuristic version, and from the coding process, which involves searching in the codebook. The bigger the codebook is, the less information distortion can be achieved, and yet, the longer time and more memory space the codebook generation and coding processes take. To make a balance between the compression ratio and the information fidelity is critical in applying VQ to real problems. The decoding is fast and simply a table lookup process.

There are attempts to pack the image information through color space transformation and color quantization prior to feature extraction, the most significant of which include: 1) median cut, 2) octree algorithm, 3) Kohonen neural network quantization, and 4) k-means [Verevka95].

The prevalent color quantization method is based on median-cut algorithm. The algorithm is computationally simpler and easier to implement, the clustering boundaries are parallel, and its coding scheme is only roughly adaptive to a particular image and not targeted to minimize the coding error. When the size of the codebook is small, it generates serious color distortion.

In this dissertation, an EM-based (Expect-Maximization) color quantization approach is used to encode color information in H&E stained pathology microscopic images. The reasons for adopting this clustering-based color quantization over alternatives such as median-cut, octree, or self-organizing map (SOM) are based on several considerations.

Using a clustering algorithm has the advantage of the capability of classification of pixels based on their color intensity and the classification can be easily fine tuned by introducing extra parameters or by switching to a supervised version of the clustering algorithm that better reflects the nature of the classification problem and takes advantage of our knowledge about the image content. Many variations, modifications and improvements have been made to clustering algorithms like k-means to make them more suitable for particular problems and objectives. Special techniques that are potentially relevant to pathology image feature extraction. The number of clusters can be optimized

according to the data distribution. There are ways to change the unsupervised k-means clustering into supervised clustering with prior knowledge about the staining property and the class densities of the sample. The clustering is based on distance measure, which can also be modified to generate better clustering results. There are rooms for improvement in the E-M iterations. The result of k-means clustering is also easy to interpret in the light of our knowledge about image content, so that problems in the process can be easily identified to improve the efficiency of clustering and color quantization.

Alternative color quantization algorithms including median-cut, octree, and self-organizing map are either not an explicit classification algorithm, or difficult to incorporate prior knowledge about the data into the algorithm. Their quantization performances are also difficult to access. Octree and SOM are also less scalable to large image size, and not so flexible to allow improvement when compared with most clustering algorithms. The k-means clustering is also better modeled mathematically.

The k -means algorithm used here is a simple VQ algorithm optimized for minimizing squared error of Euclidean distance. K points are selected through an iterative, expect-maximization process so that they are optimal in representing all the pixels assigned to each of the k groups with minimal distortion. K -means algorithm for color encoding can be implemented as following:

1. Assign k initial centroids of k groups with the values of k random pixels from the image
2. E-step: assign each pixel to the centroid with the shortest Euclidean distance

3. M-step: recomputed each of the centroids as the average of the pixel values of the group
4. If any centroid changes location, loop to E-step; otherwise, stop.

Nevertheless, theoretically, the strategy adopted in this dissertation doesn't impose any restriction on what kind of color quantization algorithms to be used. The adoption of k-means clustering is mainly for practical reasons such as flexibility, efficiency, and clear interpretation of the result. This doesn't exclude the possibility of using other color quantization algorithms, or any other classification, coding algorithms to derive a symbolic representation of an original image, which is suitable for subsequent morphological feature extraction.

For the same reason, the proposed strategy doesn't necessarily rely on run-length probability distribution as morphological feature descriptors for image content. Many morphological feature extraction algorithms have been used with image objects, regions, or other entities that are rough matches of an object. These features include basic measurements based on area, boundary, radius, bounding rectangles, and any ratios, derivatives of them, to complicated concepts such as moment invariants, fractal features. All these may be appropriate if they find their use in solving a particular problem.

There are reasons that make run-length probability distribution an attractive idea to try out and a candidate for good solution here. First, code run-length, a basic concept in coding theory, has long been used for encoding and compression for a broad variety of images, and has been incorporated into several popular image file standards, such as BMP, GIF, PNG, TIFF, and TGA. Although this serves as a strong indication that it can be used

as a good image feature, it has not been reported of such usage for the purpose of content-based retrieval of color images. A texture feature description method was proposed by Galloway [galloway75] [tang98]. It has been mainly applied to black-and-white images with different shades, such as radiology images. It is mathematically simple, easy to derive, and the computation is efficient after the previous color quantization. Simple statistical models can be used to define distance metric without much modification, which is in accordance with the general guideline of the research design. The fact that run-length feature is only one-dimensional should not be a problem with most pathology microscopic images as the images are not considered directional. The run-length feature should be invariant regardless of the direction in which the image is captured and processed.

Another consideration is that other previously mentioned morphology feature extraction algorithms were developed as attributes of image entities that are better defined, such as objects identified by object recognition algorithms, image regions that are result of image segmentation. These tend to carry over various assumptions and prior knowledge that those preprocessing algorithms are built upon but may not hold true in the situation with k-means color quantization which doesn't carry out the promise of object recognition or image segmentation. The symbolic representation of the original image consists of color code blocks that do not necessarily resemble objects or regions. Specifically, the color codes may not aggregate to form discrete geometry shapes to allow morphometric measures. Instead, they form low-level distribution patterns.

Run-length feature is a kind of low-level feature that relies on minimal assumptions or prior knowledge about data set and thus remains valid across a broader range of data set.

The more specific a feature is, the more assumptions it makes about the data it handles, the higher probability it will fail with different data under different circumstances. So, those high level features may work well for objectives such as object recognition, image segmentation, with specific image content, but low level features are more suitable for the purpose of content-based image retrieval where it is expected to work well with many image types that it has no chance to be trained with. Such a search engine is usually expected to use only a relatively small number of general features and deliver sustainable performance over a broad range of content.

The adopted methodology treats all features as global to the whole image rather than as associated to particular objects, although it is possible to apply the same feature extraction algorithm after regions or objects have been identified. The combination of k-means color quantization and run-length probability distribution feature is a choice in favor of automated extraction of low-level, low-cost, weak features that do not rely on many assumptions about the nature of the two-dimensional image content and prior knowledge about interpreting and understanding the image content.

5.2 SCIENTIFIC CONTRIBUTIONS

5.2.1 Major Scientific contributions

Since early last decade when the term content-based image retrieval was coined, there have been various attempts in tackling the challenge using individual approaches with various levels of success. These efforts, however, are far less systematic when compared with text

information retrieval, which has matured in the same time period. The published surveys about the research area mostly focus on technical issues. Theoretical components of a general framework are not adequately defined. Few general guidelines can be found about how to select effective image features, and many times technologies are borrowed from other specialized image science and engineering areas that may not be suitable for building an image search engine. The problems of such technologies are that they make many invalid assumptions about the data and thus make the algorithms over-complicated, inefficient, and invalid.

First of all, this dissertation makes an effort to address some of the issues by defining key theoretical components, defining the relationships among them. Then, following these general guidelines, a new approach of defining morphological features for pathology microscopic images, feature comparison, and content-based retrieval is proposed. A brief summarization of these major findings is provided below. Other related issues are discussed in the following sections.

Unlike with text information retrieval where word roots, phonemes, terms are inherently meaningful, the fundamental unit of image information is not obvious. A pragmatic definition of image feature is proposed as a numeric value generated by human experts or computer programs according to given criteria or following certain algorithms, which is also very different from the feature concept as used with image processing. Through defining the concept of feature in the context of image information retrieval, a three-layer framework -- including an image document space, a feature space, and a concept space -- for context-based image retrieval is proposed (figure 2.1). The framework

indicates that while there are usually algorithmic methods to extract image features from an original image, the projection from feature space to concept space is generally non-deterministic but rather based on statistics. The “semantic-gap” issue is not addressed with this model as it, like any other data mining approaches, relies on a data centric, bottom-up approach to explore the regularity in the distribution of data and their correlations with the distribution in the concept space. No restrictions have been imposed on the way how image features are derived. This allows the flexibility to ensure all classifying features can be included in the final feature set for image content description. It is not necessary for a feature derivation methodology to mimic any concepts in the process of human vision and image understanding. Weak features are especially favored by such an approach as they usually impose few assumptions regarding the distribution either in the feature space or in the concept space, and remain valid across different image types and image content. This is an approach that relies on feature extraction procedures to preserve all essential information for image content description, hopefully in a compact representation, and then statistical methods will be capable of finding the regularities among the features that can be used for the purpose of image content indexing, comparison, and content-based image retrieval. This can be achieved through modeling, data mining, or other methods.

The adopted k-means clustering based color quantization and run-length probability distribution algorithm exemplify the theoretical framework as summarized above. Both k-means clustering and run-length probability distribution are simple, popular methods in their respective area, and are mathematically well defined. Color code run-length is used for raster image compression and internally supported by several image file formats.

Although there is no previous report of using them as feature extraction and comparison method for the purpose content-based image retrieval, they are capable of modeling the tissue morphology effectively without resorting to morphometric measure, hence, the requirement of domain knowledge in the designing process is minimal. The method doesn't require any training, nor does it require human assistance when applied to new image types. The process is highly automated, with only one simple global parameter to decide on.

The results as shown later in the chapter indicate the effectiveness of the proposed k-means color quantization and run-length probability distribution as a feature extraction and comparison method for the purpose of content-based image retrieval, which demonstrates the feasibility of the forward-mentioned content-based image retrieval framework. Based on the lessons learned from this research, further improvements on the detail of the algorithms and future works are elaborated in detail in the last chapter.

5.2.2 How many colors do we need?

For most other imaging applications, the rapid advancing hardware technologies always push software development to handle more colors. More color depth also provides more subtle visual details for image content description and feature extraction. This is a major reason most researchers in building content-based image retrieval systems would prefer high quality images, some would even go beyond 24-bit color depth to adopt multi-spectrum imaging apparatus.

On the contrary, other evidences exist that show human visual perception relies more on the transformed domain image features and relatively tolerant to color changes, such as color shift, color distortion. The number of colors that human visual system can discern at one time is also far less than the state-of-the-art hardware can provide. Some reports of extreme tests also show that experts can work with digital images with extremely reduced number of colors and still give the same performance. In one report, pathologists worked with digital microscopic images of only 256 colors could make accurate diagnosis just as well as with more colors. The reduced color pallet didn't impair the accuracy of diagnosis at any significant level, and this was achieved using generic color quantization method that is not optimized for any particular purposes, and thus less efficient for the task.

Some image processing practice also contents with limited colors in many ways. Early image format standards, such as GIF, allow a maxim number of 256 colors at a time, yet, GIF has become one of the two earliest, most popular image formats of the Internet. Many image-processing routines involve the process of reducing the number of colors in the image.

In some of the major domains of imaging and image analysis, such as those satellite images and medical images, the color spectrum is inherently limited, where the colors are either artificially generated, stained, or digitally assigned. In many cases, human experts dealing with these images actually require only a limited number of colors compared with that the images are captured with. The domain specific context and the limited color spectrum also make it possible for algorithms to produce a color pallet that is much more

efficient than existing generic color quantization schemes can do, while maintaining the visual quality to human perception and minimizing the distortion during transformation, both criteria that are important in determining how far the color quantization algorithm should go.

This research will apply k -means clustering algorithm to H&E stained pathology microscopic images to derive a compact pallet with minimal number of colors under the constraint of maintaining reasonable image quality through visual inspection, as well as keeping minimal image distortion, which will be measured by numeric metric. According to the preliminary results, it is possible to achieve the goal with only 16 carefully selected colors, a color depth of only 4 bits, which is a significant reduction from the original 24-bit color depth. The expert visual inspection and close monitoring of quantitative distortion measure ensure that most of the visual information is well preserved just with these 16 colors.

5.2.3 Discrete domain processing

A successful color quantization step can open the door to some novel methods in the discrete domain, other than the traditional continuous domain processing. It is much anticipated that discrete domain processing is more efficient in patterns recognition, feature extraction, in content description, coding, classification, and searching. After the image is represented with only a limited number of codes, a diversity of data structures are available to facilitate further processing. Discrete domain processing is also less ambiguous and less error would be introduced in the subsequent processing. The advantage

of the new scheme with highly efficient color quantization followed by discrete domain processing is that the one-time price paid at the initial color quantization is under careful control and easily optimized to retain most visual information and adapted to the particular type of image. In the methodology proposed in this report, and as show in the preliminary results in the appendix, the k -means clustering generates codebook according to the color distribution of that particular image type and the transformed image can be both visually inspected by human experts and also monitored by statistical metrics. In the continuous domain, the error could be amplified in every involved step.

One important problem in image processing and image understanding is image segmentation. In most cases, a segmented region is a rough approximate of an object, or a part of an object. Many image segmentation approaches divide arbitrarily the image into small blocks, classify the blocks, and merge those that are similar to each other to form regions. A different approach uses VQ algorithm to code those blocks and use the distribution pattern, or frequency of the code as image content descriptor for content-based retrieval [zhu00] [zhu02]. The major weakness of both approaches lies in the arbitrarily dividing of image blocks, which is not tuned to adapt to the scale or location of the image objects. The k -means based color image coding naturally provides a simpler setting for image segmentation in the discrete domain through adaptive pixel clustering.

One most commonly used image feature in CBIR is color feature, which is used to produce color histograms. Many researchers have investigated the use of color histogram or its improved variants as image content descriptor. In solving some problems, color feature alone is sufficient to give satisfactory image classification and content-based

retrieval performance. However, the image color is susceptible to many environmental conditions and imaging hardware. It usually takes careful calibration at image capturing time or pre-processing routine to compensate the variance. This problem is addressed by adaptive color quantization as in k -means based color coding.

In this proposed project, the tissue samples are stained with chemical dyes that render the color difference according to the biochemical constitution of the tissue objects. After coding, the codes' run length correlates much to the scale and distribution of the tissue objects of various magnitudes. By investigating the run length probability distribution of the code, it's very likely to derive efficient content descriptors that are compact and adaptive to the type of images. For image classification and content-based image retrieval purpose, this could be more efficient than image segmentation and object recognition approach in the domain with continuous color value or transformation coefficients, and much more effective than the simple color histogram descriptors.

It can also be speculated that many other unconventional discrete domain methods can be applied directly to the coded images, without any need to intermediate processing. Such techniques include hashing based image block indexing, Hidden Markov Model (HMM). Although this approach could be limited in the kind of tasks it handles best, it provides a novel access point to solving many domain-specific image classification and content-based retrieval problems, and bring about a very rich and powerful set of discrete domain processing tools to solve the problems they can do best, and this approach to this problem has not been reported in literature before, and is author's major theoretical contribution. More detail will be explained in the rest of this chapter

5.3 HYPOTHESIS AND EVALUATION

In content based image retrieval, for the purpose of effective feature processing and fast searching through the signature file, image features are usually further quantized from their original version. It is expected that a data dependent, weighted quantization and classification-aware encoding should give better performance than the blind, uniform quantization. In the highly packed feature representation, less useful information would be filtered out, and high classification accuracy can be maintained.

A high quality codebook provides a set of essential image content descriptors for feature extraction and representation, and noise reduction, and thus, is crucial in improving retrieval accuracy and efficiency. Three important factors that determines the quality of the encoding process is the coding error, the information packing capability that is related to the size of the codebook, and the CPU time it requires. Using the same coding algorithm, a bigger codebook can usually minimize the coding error at the expense of a bigger coded file size and longer coding time, while a codebook with a limited vocabulary generates a compact coded representation, which is beneficial to efficient content indexing, in short time, but with bigger coding error. It desirable to obtain a relatively small codebook that contains the most important image features while leave out those less significant ones. This often translates into minimizing the coding error for a particular kind of images under the constriction of a fixed codebook size.

The error in the process of image VQ encoding comes from two sources: 1) shift in the color space caused by variation in the objects themselves or in the imaging process, such as uneven staining and light illumination; and 2) variations of image composition in

the spatial domain, including affine transformation and image warping. Under the constraint of codebook size, minimizing these errors can significantly improve the coding efficiency and accuracy.

This report proposes a color space information packing method using the k -means method that aims to adaptively reduce the complexity of color representation. The benefit of this includes better quality of color quantization and improved accuracy of image content classification based on the improved codebook, when compared with linear color quantization. This is especially true when the pixels from the image are not distributed evenly in the color spectrum. Such cases can be found in image collections with limited scope, or those images that are artificially colored, such as those images in GIS systems, and those medical images that are stained with chemical reagents, captured under unconventional illumination condition to induce photonic reaction only in certain bandwidth, or some of those images that are artificial colored during digital processing. This proposal focuses on pathology microscopic images that are H&E stained with blue and purple colors. More details will be provided in the next chapter.

Both the image distortion resulting from a transformation and the classification-based retrieval performance can be evaluated by methods and metrics that have been established in their respective domains. To measure image distortion, standard error is defined as the average squared difference between the pixels from the original image and those from the reconstructed version:

$$Error = \frac{1}{N} \sum_i d(\vec{p}_i, \vec{q}_i)$$

where N is the total number of pixels examined, $d()$ is the distance function, $d(\mathbf{p}_i, \mathbf{q}_i)$ is the distance between two representations of pixel i ; \mathbf{p}_i and \mathbf{q}_i are different vector representations of pixel i from the original image and the transformed image respectively. A better transformation algorithm with less distortion should give a smaller error value.

In statistical formulation, if we view the original image as a true data set with each pixel as a data point, the color quantization can be viewed as a sampling process. Assuming that the error follows normal distribution, and Euclidean distance is used to measure the distance between two pixels as following:

$$d^2(\bar{\mathbf{p}}_i, \bar{\mathbf{q}}_i) = \sum_j (p_{ij} - q_{ij})^2$$

where j is the index to the vector representing the color depth. Thus, the evaluation of the relative performance of two quantization algorithms is equivalent to the inference about two independent variances. This leads to the following hypothesis for the statistical testing: The null hypothesis represents the situation that there is no significant difference between the levels of fidelity of the two color quantization approaches, so that, their error, which is calculated with the same formula as variance here, won't show significant difference. Alternatively, the hypothesis to be proved is that the proposed new method generates smaller error (variance) than the old one. Thus we have following hypothesis for a one-tail test, assuming the observed error should not be larger for the proposed method than that for the control method:

$$H_0 : Error_{new} = Error_{old}$$

$$H_1 : Error_{new} < Error_{old}$$

The test statistic for testing H_0 against H_1 is the ratio, F , of the two sample variances:

$$F = \frac{s_{new}^2}{s_{old}^2}$$

When the images to be compared have the same number of pixels,

$$F = \frac{s_{new}^2}{s_{old}^2} = \frac{\sum_i \sum_j (p_{ij} - q_{ij})^2}{\sum_i \sum_j (p_{ij} - q_{ij})^2}$$

In this project, GIF color quantization scheme will be used as control. The $Error_{old}$ and s_{old} denote the error in the GIF color coding. Detailed explanation of the proposed quantization method and the hypothesis testing can be found in the second part of the next section [glass96].

The retrieval performance of both approaches can be compared graphically using precision-recall curve (P/R curve). Precision is defined as the portion of retrieved documents that are relevant. Recall is defined as the portion of all relevant documents that have been retrieved. The results from both methods can be plotted in the same graph.

Using the same query set and image collection, the better performing method will generate a curve with higher precision and better recall in comparison with the other. This is a standard practice in image retrieval performance evaluation as well as in more general information retrieval scenario. More detail will be discussed in the next chapter.

5.4 MATERIAL AND METHODS

5.4.1 Materials

The vast majority of pathology microscopic images are stained with a dye consisting of Hematoxylin (blue color) and Eosin (red color) (aka H&E staining). The two chemical reagents bind to protein molecules with different affinity according on their electronic charge. Hematoxylin binds to basic proteins with a positive charge, while Eosin binds to acidic proteins with a negative charge. This renders the originally thin, almost transparent tissue sample slice a purplish color with a spectrum spanning from purple red to dark blue.

Tissue micro-array is an emerging revolutionary technique that assembles a large array of samples from different sources on one single glass slides to facilitate uniform, simultaneous processing with various techniques, quality control, and better reproducibility. As a result, it is expected to improve productivity, quality, and save precious tissue samples and biochemical reagents. It also brings about new frontiers in biomedical research and diagnosis. Both hardware and software image processing techniques for automated tissue micro-array data analysis have started to gain momentum. Technically, one of the major reasons to use tissue micro-array is that all the images on the

same glass slide are stained through exactly the same process and using the same reagents, which means, the color variation is under control across different images. Besides, other variations due to the defects or difference in the optical property of the glass slides are also reduced to a minimum.

The proposed research will use digital images captured from the H&E stained tissue micro-array glass slides. A typical capturing station is an optical microscope, a digital camera hooked on top of it, and a computer that controls the microscope and also stores the captured images. So far, most image content description methods that have been carefully studied are sensitive to the image scale. For pathology microscopic image, 20x combined objective and ocular magnification captures enough detail of sub-cellular structure. An image size with 1 million pixels and a color depth of 24 bits (1000 x 1000 x 24 bits) should provide a field with enough scope to ensure statistical stability in feature extraction. At this magnification level, the object is captured at roughly 1 micron per pixel. These estimates are based on visual inspection of common pathology specimens, previous experience in CBIR system design, and the nature of the image data mining and feature extraction algorithms described below. It has been shown that human pathologists can still make diagnosis with a color depth of about 8 bits, when dealing with digital microscopic images. However, previous experiences also show that machine vision can distinguish and take advantage of more subtle color details, and some image processing/feature extraction algorithms are very sensitive to color variations caused by disturbance in the staining process and the variation in light sources of the capturing station. The image will be saved

in uncompressed RGB TIFF format, a format that is popular in various imaging applications.

Real medical cases will be used for both algorithm design and performance testing. Each case will include a digital image captured according above specification, a class tag, and an optional more elaborated description explaining any differentiation under the general class umbrella, both of which are assigned by the pathologists as a part of the diagnostic process.

1 Table 5.1 Images and tissue types

Tissue Type	Case
Brain	12
Fibrous tissue	13
Heart	73
Kidney	23
Prostate	26
Prostate Cancer	30
Brain*	23
Thyroid*	23
Total	225

More than 250 images have been captured from tissue micro-array slides, 217 of which are of good quality and have been diagnosed. Six most abundant types with a total of 179 images are selected. Besides, additional brain and thyroid images (brain* and thyroid* in the table) have been captured from normal glass slides. These add to a total of

225 images from 7 diagnosis categories, near 900 MB of image data. Following criteria are used for image quality control:

1. Images are inspected visually while capturing
2. Images captured from a micro-array spot that is not intact are stored but not diagnosed later
3. Images that are difficult to make an unbiased diagnosis for based on the content information within the captured scope are left out
4. The tissue types that have less than 10 quality cases each are not included

5.4.2 Methods

a. Color Coding and Image Transformation

There are various color models, each has special properties for different applications (see review of image color feature). One of the simplest and also one enjoys great popularity is the RGB color model. In RGB model, the color of a pixel is defined by values of three primary color channels, namely Red, Green, and Blue. Every unique color can be defined as a linear combination of three primary colors, or in a spatial analogy, a point in the RGB color cube. The color difference of two pixels can thus be defined as the Euclidean distance in the three dimensional RGB color space. This distance metric is suitable for k -means algorithm without the need for any modification when computing the centroids of the k -means classes.

After the distance metric has been determined, the single tunable parameter to decide upon is the value of k . K -means in general doesn't impose any particular restrictions

to the value of k . However, in this special case as for quantization purpose, it is more reasonable to assign k value with one from the series of 2^n , where n is the number of bits to be used to represent each code. For the kind of H&E stained pathology microscopic images, to find the optimal value for k that generates minimal image distortion while keeping the size of the codebook small, there are only a few reasonable candidates as will be used in this research. They are 4, 8, 16, 32, 64, and 128, as 2, 3, 4, 5, 6 and 7 bits are chosen as the bit number respectively. A k value of 2 degenerates the 24-bit color image into a binary image and loses essential details, because there are far more than 2 constituents in any kind of tissue. Previous study has shown that digitally captured microscopic images with a color depth of 8 bits (as the maximum that GIF format allows) are sufficient for human pathologists to perform diagnosis without significantly impairing accuracy. It is the goal of using k -means algorithm here to generate a more compact color representation than previous methods.

The method for determining the k value using the color distribution of sample images is summarized as following:

1. Read digital image file in TIFF format using `libtiff` (<http://www.libtiff.org/>) library function.
2. Draw a small portion of random pixels from the image.
3. Apply k -means clustering algorithm to the sample pixels with a selected values of k as 4, 8, 16, 32, and 64
4. Transform the image to represent each pixel with the centroid it belongs to in the k -means cluster

5. Compute the distortion (the formula for squared distortion is given) of the

transformed image using formula: $error^2 = \frac{1}{N} \sum_i |p_i - q_i|^2$

6. Plot the distortions, e , against n , the number of encoding bits, which is $\log_2(k)$.

Pick the bit number corresponding to the point with the biggest derivative. Thus the corresponding optimal k value is determined.

This is a crude solution for picking a characteristic value from a very sparse series of a limited number of choices. With the chosen value of k and the determined codebook through k -means clustering, the image is coded by representing the RGB value of each pixel with the one of the k codes from the codebook.

b. Feature Extraction and Distance Measure

After the image has been coded as described previously, feature extraction is achieved by finding statistical pattern in the code distribution in the coded representation. The coloration of tissue is mostly according to its affinity to the dye based on its biochemical composition. The transformed H&E stained microscopic image will have limited number of kinds of pixels rather than a wide range of almost continuous spectrum of color values. A much more anticipated consequence than the reduction of codebook size is that the transformed image is segmented into small patches of same code that are roughly corresponding to the functional tissue, cellular, and sub-cellular components, which are important for medical diagnosis. This results in code distribution patterns of very high co-occurrence, with a much higher probability for a code to be used at the location with neighboring pixels represented by the same code. In a simplified one- dimensional

representation, it reveals very regular and characteristic code run lengths distributions rather than one that is totally random (see figure below). Different codes also bear different characteristics what can be attributed to the morphological properties of the various functional components the codes represent. The distribution pattern of the code in the image is a function of the k value, the magnification level, and is also characteristic to that image and, hopefully, other images of the same class.

In this research, a simple pattern descriptor based on code run length is proposed for image feature extraction. For each code in the codebook of size k , one distribution (actually a histogram without being normalized to sum up to 1, due to its one-dimensional nature) of code run length, L , can be derived; for each image, a total of k discrete distributions can be derived. Assuming l elements are actually used from a series of L , and all k distributions are used for image comparison, an image can be represented by a feature vector of $l*k$ elements. The distance between two images, D , can be defined as linear combination of all pair-wise distances between the corresponding code distributions:

$$D = \sum_{i=1}^k w_i d_i$$

where k is the size of the code book and a parameter of k-means algorithm, w_i is the weight assigned to code i according to its relative contribution to the image content representation, d_i is the distance between two run length probability distributions, and is a function of the two series $L1_i$ and $L2_i$. It is expected that d_i , when $i = 1, 2, 3, \dots, k$, are not independent to each other due to the nature of k-means as a clustering algorithm. However,

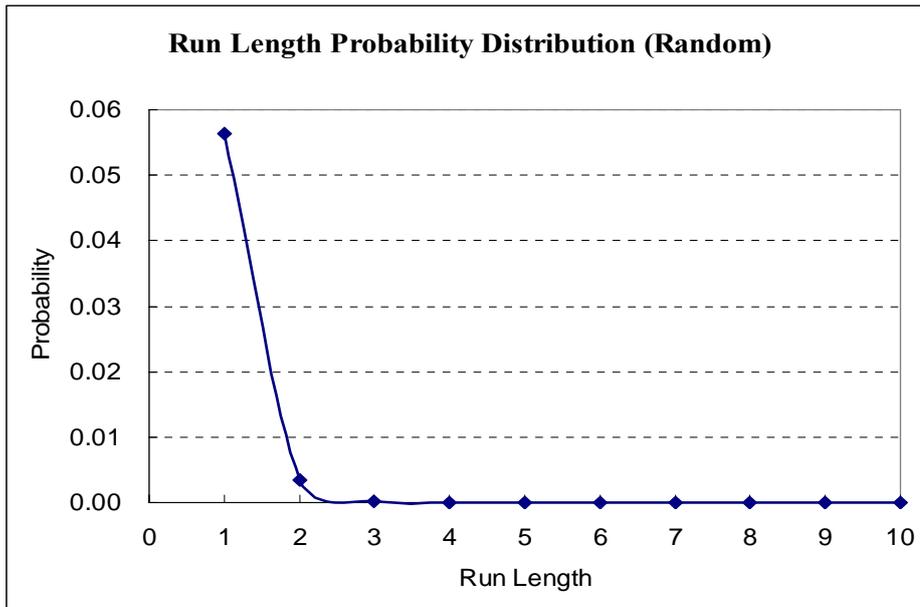
when the codebook size is minimized to approach the number of essential pathological components in the image, the inter-dependency is expected to be kept to minimal.

There are a few different methods in computing the distance between two discrete probability distributions. The two most popular ones are L_1 -norm, which sums up all the differences between the corresponding probabilities, and Kullback-Liebler distance, including its variances and metrics that based on the theory of information distance [weeds02]. Some previous researches also tested kurtosis as a measure for the skewness of a distribution. The difference between the two kurtosis values was then used as the distance of the two distributions. However, Kullback-Liebler distance is widely used to measure mutual information, and is the method used for this research.

For a given code density p_c , the estimated probability, $prob(n)$, for run length n is computed using following formula:

$$prob(n) = p_c^n (1 - p_c)$$

Using this formula, a plot of probability distribution of different run lengths of random pixels (with density $p_c = 0.06$) is shown below.



10 Figure 5.1. A typical run-length probability distribution for a random code.

Assuming k value equals 16 here, the average probability density is roughly 0.06. Only the first 10 run lengths are shown.

The code run length probability distributions from actual images are very different from the above curve and show the particular characteristics of different functional components in the image. Also, different codes have very different probability density values in the image.

The original form of Kullback-Liebler distance is defined as:

$$d(p, q) = \sum_{x \in X} p(x) \cdot \log \frac{p(x)}{q(x)}$$

The unmodified version of Kullback-Liebler distance formula has problem when $p(x) = 0$, and this is not uncommon with sparse data and using Maximum-Likelihood Estimate (MLE) without preprocessing such as smoothing. There exist many improved

information distance measures that fix this problem. In this research, however, the run length data is not really sparse, especially with shorter run lengths. With the run length probability distribution of random pixels, the first 2 run lengths add up to more than 0.99 when $p=0.06$. With the average run length probabilities of a real image, the first 5 run lengths could add up to 0.9 (see the results shown in appendix). By using only the first several run length probability values, the sparse data problem with Kullback-Liebler distance can be avoided. In this research, the probability values will be computed from all the pixels of the image using MLE.

Image retrieval is based on this distance measure D . The images from the database with the shortest distance to the query image are returned to the user as the answer to the query.

The feature extraction step is summarized as following:

1. Encode the image using k -means algorithm with the k value determined in the previous step
2. For each of the k codes, derived the run length probability distribution; the code and its probability distribution is defined as a feature; k such distributions are used as the signature of the image.
3. The similarity of two images is determined based on their signatures and according to the distance functions given above. Image retrieval is achieved based on the distance measure.

c. Algorithm Tuning

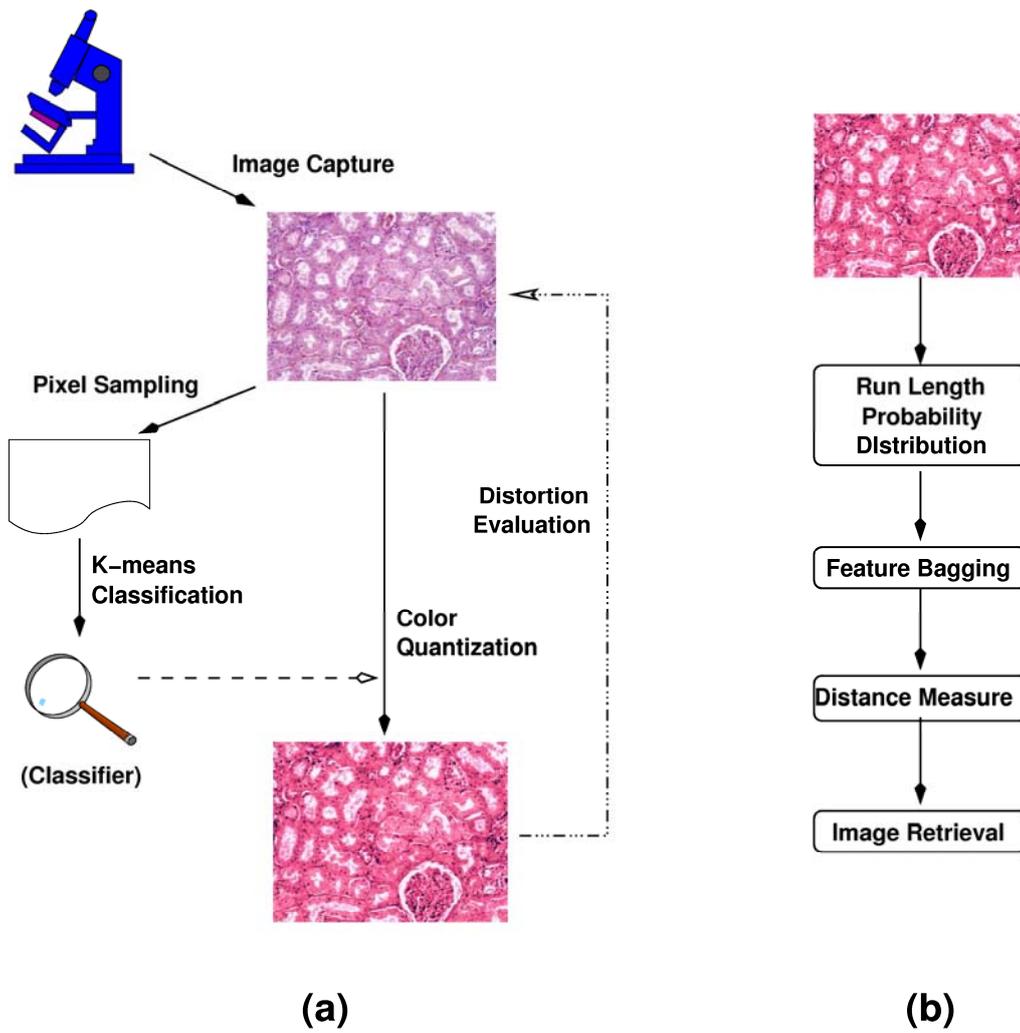
There are three kinds of parameters to be determined:

1. The value of k as described in the previous section.
2. The weight of each code in the distance formula in computing D . This is the same process also called feature selection/weighting. In many simplified cases, all features are assigned the same weight so that each of them contributes equally to the distance measure. The preliminary results have shown that this solution may work well with the data collected. A more reasonable solution for complicated situations with a less complicated approach is to group those features based on the similarity of their probability distribution and assign the same weight to each group. This is called “*bagging*” in information retrieval, and is also a technique related to dimensionality reduction in the feature space, where several features with similar probability distributions are combined to form a complex feature.
3. The probability coefficients s_j in the formula for d_i . It is a function of run-length, j , and the probability density of code i , which can be estimated from the query image.

For efficiency consideration, since the goal is to estimate the distance between two run-length probability distributions, not all probability values are necessary to be taken into account during signature generation. Those run length probabilities that have the most representation in the pixel population and most characteristic to the image content are more important. The first factor emphasizes on those corresponding to the shorter run lengths, which is shown as peak in the above estimated run length probability distribution plot; the

second usually refers to those run lengths that are of the scale of the functional tissue, cellular, and sub-cellular components, which is represented by the part of the real probability distribution summarized from an image that is inconsistent to that of random pixels (see the preliminary results). This could vary from tissue to tissue and should be determined according to the particular cases. However, since the size of most human cells remains relatively constant, the sub-cellular components that are significant for diagnostic purpose are usually under 20 pixels in diameter under the specified magnification of 1 micron per pixel. The run length probability values that are in that scale should be most significant for content description.

The following flowchart shows the steps involved in: a) color quantization, and b) feature extraction, processing, and retrieval algorithm.



11 Figure 5.2 Flowchart

From glass slides to digital image retrieval, digital imaging, image data mining, classification, numeric feature extraction, dimensionality reduction, and distance measure.

5.5 EVALUATION

5.5.1 Information Packing Capability

An important measure of the performance of feature extraction algorithm is the amount of visual information that has been conserved during transformation, or in other words, the distortion of the transformed image. To evaluate the performance of k -means based color quantization, distortion formula given in the previous section is used, and the result will be compared against that of the popular, general purpose median cut algorithm (such as the one used in GIF compression) using the same codebook size. Each test image will be coded using both color quantization methods. The encoded version should have much smaller uncompressed file size due to reduced color number:

$$\text{File Size} = N \cdot \log_2 K$$

where N is the number of pixels in the image; K is the size of the codebook. The result is in the number of bits. A 1280x1024x24 bits color image of almost 4 MB can be compressed into 650 KB using a codebook with 16 words, and yet, the original image can be reconstructed, given the codebook. The distortion is computed by the error function:

$$\text{error}^2 = \frac{1}{N} \sum_i d^2(p_i, q_i)$$

where N is the total number of pixels in the image; p_i and q_i are pixels from the original image and the reconstructed images respectively; $d(p,q)$ is the distance function between the two pixels. In the proposed research, Euclidean distance in the RGB color space will be used to compute the difference between the original image and the reconstructed version, so that the error formula takes the same format as variance in statistics:

$$error^2 = \frac{1}{N} \sum_i |p_i - q_i|^2$$

and in our particular case:

$$error^2 = \frac{1}{N} \sum [(p_R - q_R)^2 + (p_G - q_G)^2 + (p_B - q_B)^2]$$

Based on this formulation, the hypothesis to be tested is that the new method doesn't reduce the error in vector quantization:

$$H_0 : Error_{new}^2 = Error_{old}^2$$

$$H_1 : Error_{new}^2 < Error_{old}^2$$

Assuming that the new method does actually give smaller error than the old one does, one tail F test will be used here:

$$F = \frac{error_{new}^2}{error_{old}^2}$$

the hypothesis is rejected at 0.95 confidence level if $F < .05F_{N-1,N-1}$, where N is the number of pixels in the image. In this research, the median cut color quantization scheme used in GIF image format will be used for comparison. The same image as used in k -means coding is transformed into GIF format using a color pallet of size k , which is the same as in k -means coding. The GIF image will then be transformed back to bitmap representation and compared with the original image to compute the coding error. Only color mapping is carried out in the process, and the error is brought in by the color quantization algorithm only. The $Error_{old}$ in the above formula donates that error. Images from different classes will be tested using the above statistic to assess the performance of the new method across different kinds of images. An image of a dimension of 1000 x 1000 contains 1 million pixels. This sample size should be sufficient for statistical significance.

5.5.2 Content-Based Image Retrieval Performance

A microscopic image set consisting of 20 different pathological classes, each of which consists of at least 10 unique images, will be built. The images will be captured from H&E stained, classified tissue micro-array slides in the tissue sample repository at the department of pathology, University of Pittsburgh Medical Center. The capturing condition will be carefully controlled to maintain uniformity with following specification: 1) magnification: 20x objective lens; 2) image dimension: 1280 x 1024. Each image in the

collection will be assigned a unique ID. Then, the image is coded using the 256-code codebook. By comparing the histograms of the code usage in each of the images, a distance measure between two images can be derived. For example, in a simple situation, a Euclidean distance can be computed between two histograms. Given a sample image as query, retrieval of similar images is a process of searching for the images in the database with the smallest distances to the query.

The relevance judgment of content pathology microscopic image retrieval is based on correct classification. In this design, all images will be used have already be diagnosed, and each labeled with a short text descriptor. A correctly retrieved image should bear the same class label as the query image. Otherwise, the retrieval is false. In some cases, a normal tissue image can lead to retrieval of cancerous images of the same kind of tissue. This is especially common when the sample is of lower cancer grade [zheng_thesis]. This phenomenon could be a very interesting research topic by itself. But in this proposed research project, these cases will be counted as false retrieval.

The performance of the k -means based color quantization in image retrieval will be evaluated by Precision and Recall based on the following formulae:

$$precision = \frac{|relevant \cap retrieved|}{|retrieved|}$$

$$recall = \frac{|relevant \cap retrieved|}{|relevant|}$$

The retrieved images are ranked according to their distance to the query. The results will be plotted graphically. The precision at a fixed recall point, say, 20% recall, will also be used as a criterion.

The majority of the published CBIR research works lack a quantitative evaluation part. The performance is usually demonstrated by empirical visual inspection of the query images and the retrieved images from sample query sessions. In a few more carefully worked out projects, evaluation process follows the Precision/Recall, and P/R curve approaches, which have been used as a standard in text information retrieval for more than three decades, but was just proposed as a standard for image retrieval in 2001 [muler01] [muler01_2] [zhu02]. In pathology, CBIR projects usually lack a formal evaluation part [korn98] [wang], or the classification accuracy is used in place of the performance measure [foran98]. This design will follow the evaluation methodology used in [zheng_thesis] with standard Precision and Recall measures.

6.0 RESULTS AND INTERPRETATIONS

6.1 PRINCIPAL COMPONENT ANALYSIS

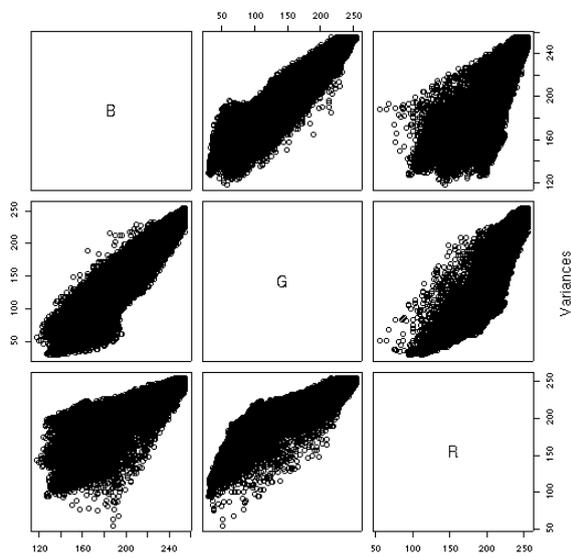


Figure 6.1a

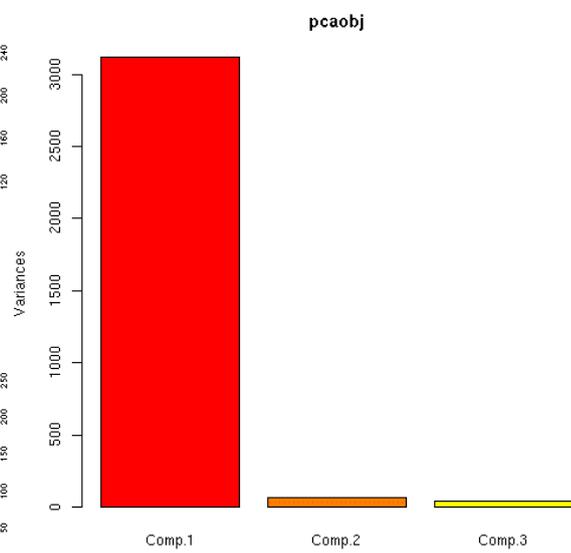


Figure 6.1b

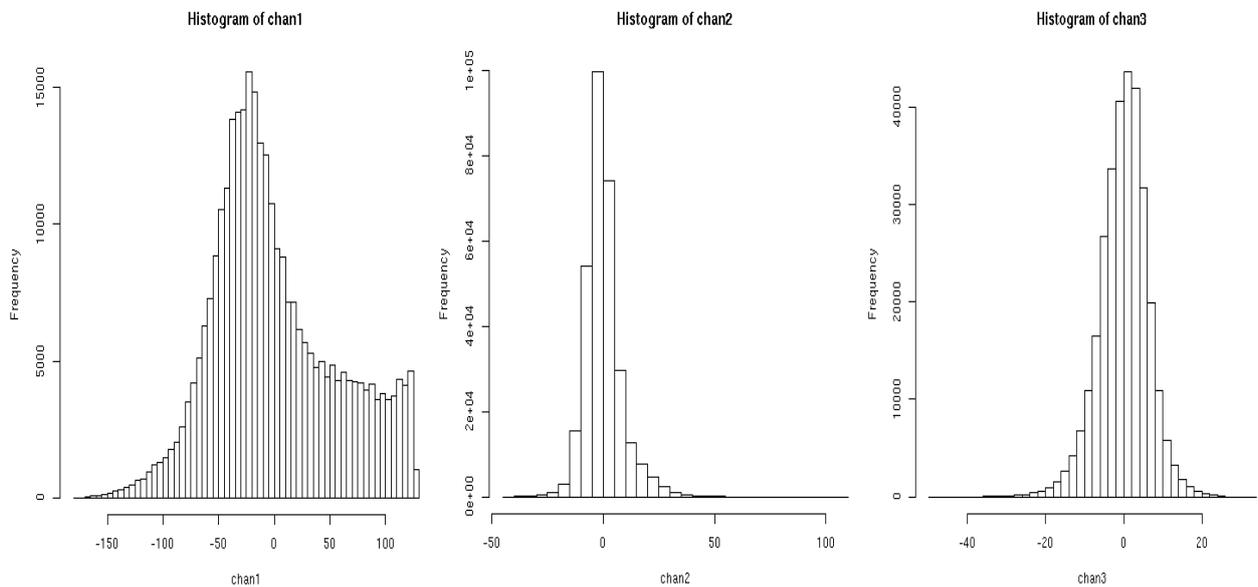


Figure 6.1c

Random pixels were selected from all images in the collection, 100 pixels from each. The results mainly reflect the color space distribution of the pixels. Figure 6.1a displays scatter plots, each along two out of the three color (RGB) dimensions a time. Figure 6.1b shows variances along three directions after transformation. Figure 6.1c have histogram plots for three components (R package (<http://www.r-project.org/>) was used for all processing and plotting.).

The results show that the distribution of the pixels in the RGB color space is extremely uneven, and this is a common characteristic of H&E stained microscopic image, regardless of the tissue type and the diagnosis.

6.2 K-MEANS CLUSTERING OF COLOR PIXELS

12 Figure 6.2 K-means clustering

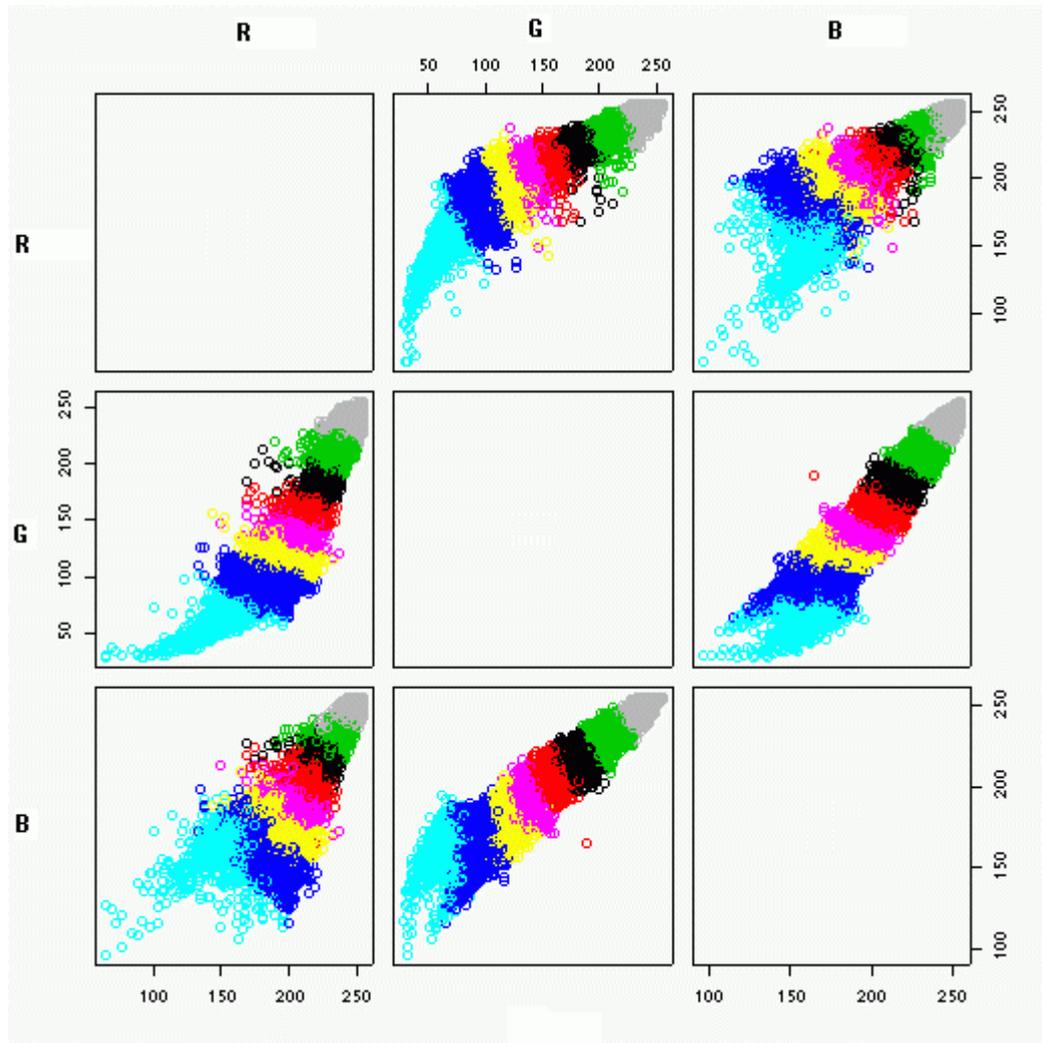


Figure 6.2a K-means clustering of image pixels with $k = 8$

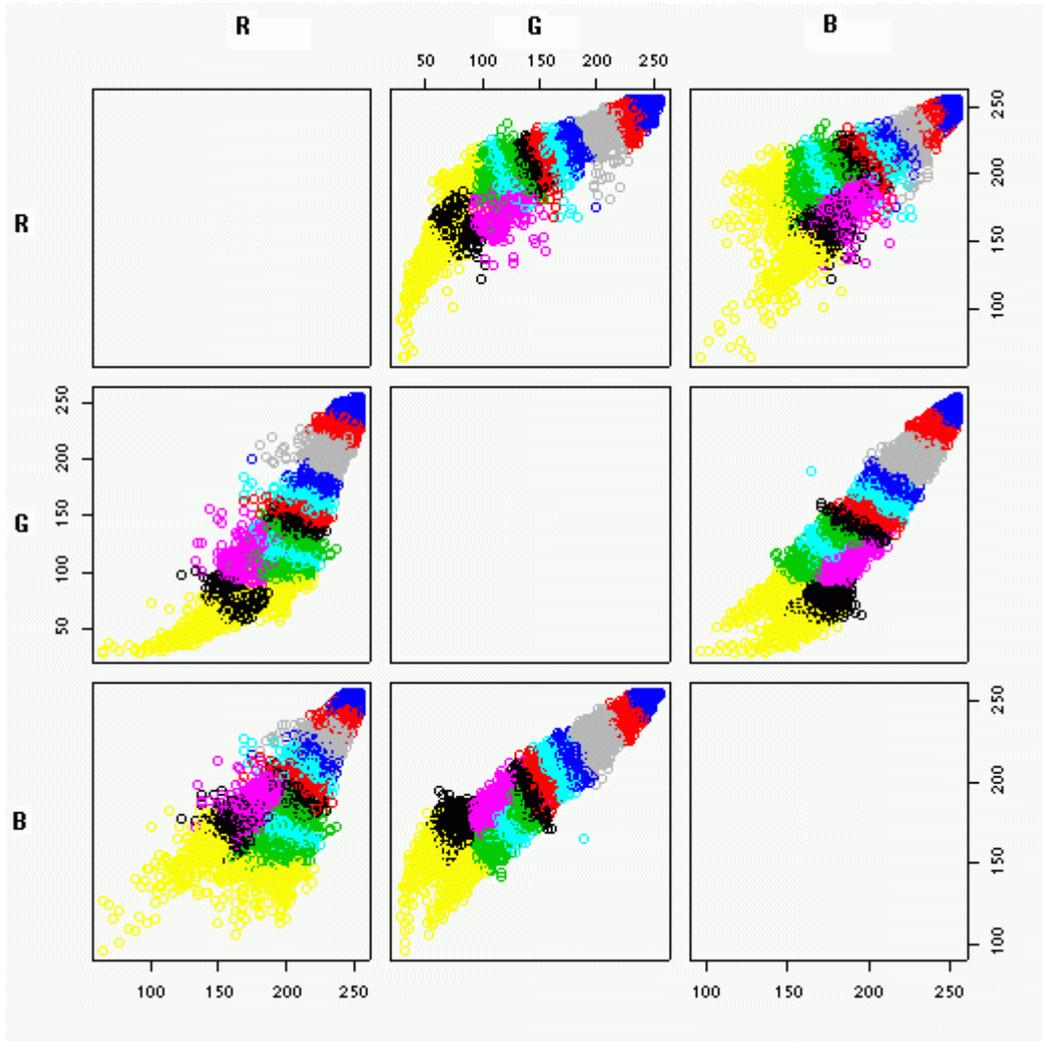


Figure 6.2b *K*-means clustering of pixels with $k = 16$

Coding efficiency is achieved at the price of image distortion. The coding efficiency is measured by the size of the codebook, while the distortion at each pixel is defined as error in the coding process, which is the Euclidian distance of RGB vector between the pixel in the original image and that in the transformed version. In the black-white subtraction images (figures 5a, 5b, 5c), the error is displayed after multiplied by 10, thus, a darker image indicates smaller error, while a brighter image indicates bigger error. K values are chosen as $k = 2^n$, where $n = 2, 3, 4, \dots$. The result shows that, while a k value

of 16 gives a very faithful depiction of the characteristics of the tissue, a k value of 4, although segmenting the whole image into four different kinds of regions quite effectively, causes loss of important detail. It also shows that more error occurs in the regions under represented in the pixel population, such as nuclei. Other sub-cellular compartments see lower error levels.

6.3 K-CODING

Table 6.1 Comparison of distortion

IMAGE ID	K-CODE ERROR	MEDIAN-CUT ERROR	F	TISSUE	DIAGNOSIS
ha01_01	11.58	23.33	0.2473	Kidney	Autolysis
ha03_02	9.61	16.78	0.3518	Heart	Heart Muscle
ha06_02	10.97	22.53	0.2437	Kidney	Kidney Tubules
ha08_03	7.83	14.73	0.3264	Brain	Brain
he01_02	9.80	23.06	0.2049	Prostate	Prostate Cancer Gleason 3
he01_09	9.55	16.93	0.3314	Prostate	Fibrous Tissue
he03_05	11.02	22.51	0.2602	Prostate	Benign Prostate
he04_07	9.02	18.11	0.2569	Prostate	Fibrous Tissue
hu03_08	9.38	19.16	0.2650	Fibrous Tissue	Fibrous Tissue
hu03_09	8.74	15.17	0.3320	Fibrous Tissue	Fibrous Tissue

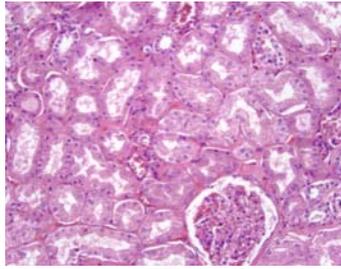
All original images are of dimension $1315 \times 1033 = 1358395$ pixels, and 24-bit color depth. Transformation error is defined as the average Euclidean distance between the original pixel value and the pixel value after transformation in the RGB color space, as defined in section 5.4.2. “Optimized median-cut” color reduction was performed using the “Decrease Color Depth” function in Paint Shop Pro 7.05, along with “nearest color”. K-coding was performed using the author’s modified k-means clustering code. All images are processed to

reduce the size of color pallet to sixteen (16) different colors, with both `tiffmedian` and `k-coding`. F statistic is computed as

$$F = \frac{S_{tiffmedian}^2}{S_{k-coding}^2} = \frac{\sum_i (p_i^R - q_i^R)^2 + (p_i^G - q_i^G)^2 + (p_i^B - q_i^B)^2}{\sum_i (p_i^R - q_i^R)^2 + (p_i^G - q_i^G)^2 + (p_i^B - q_i^B)^2}$$

For one-tail F test, $_{0.01}F_{1000,1000} = 1.112$, $_{0.01}F_{\infty,\infty} = 1.001$. The above F values in the table are all bigger than 1.112, which indicates that `k-coding` performs better than `tiffmedian` for all the tested images at 99% confidence level. The error per pixel is only a half with `k-coding` as with `tiffmedian`.

One sample image is transformed using `k-code` as described above, with `k=4, 8, 16` respectively. The result is shown in figure 5.5. The coding error is computed as the difference between the original image and the `k-coded` version. In figure 5.6, the corresponding coding error is magnified by 10 times and displayed as pixel intensity/brightness. The original version of `k-means` clustering assumes equal weight for all the sample points, thus clusters with a smaller population sees greater within-class variation than those with a higher sample density. As with pathology microscopic images, the nucleic regions that are darkly stained by Hematoxylin suffer from bigger coding error than cytosolic regions, as highlighted in figure 5.6, especially when the codebook size is limited.

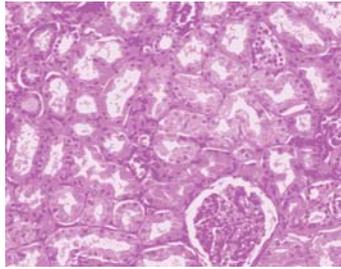


(6.3a)

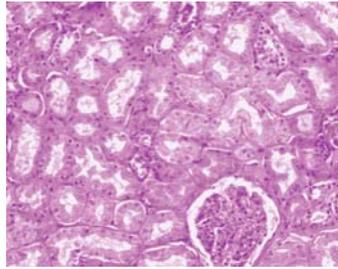
Figure 6.3 and **Figure 6.4**. *K*-coding is performed exactly as described above with *k* values pre-assigned as $k=2^n$, $n = 2, 3, 4$. The coding error is calculated according to equation (3).

5.5a) Original image; 5.5b) $k = 4$; 5.5c) $k = 8$; 5.5d) $k = 16$.

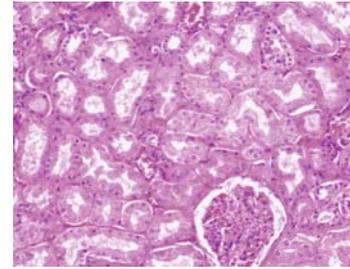
5.6a) $k = 4$; 5.6b) $k = 8$; 5.6c) $k = 16$.



(6.3b)

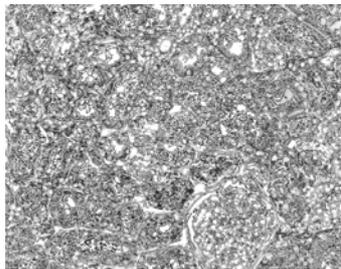


(6.3c)

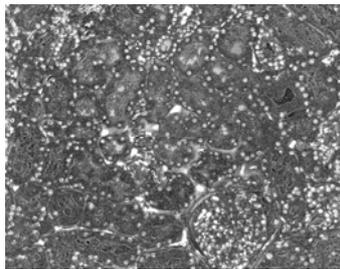


(6.3d)

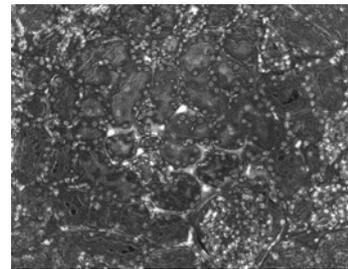
13 Figure 6.3. K-coding of pathology microscopic image



(6.4a)



(6.4b)



(6.4c)

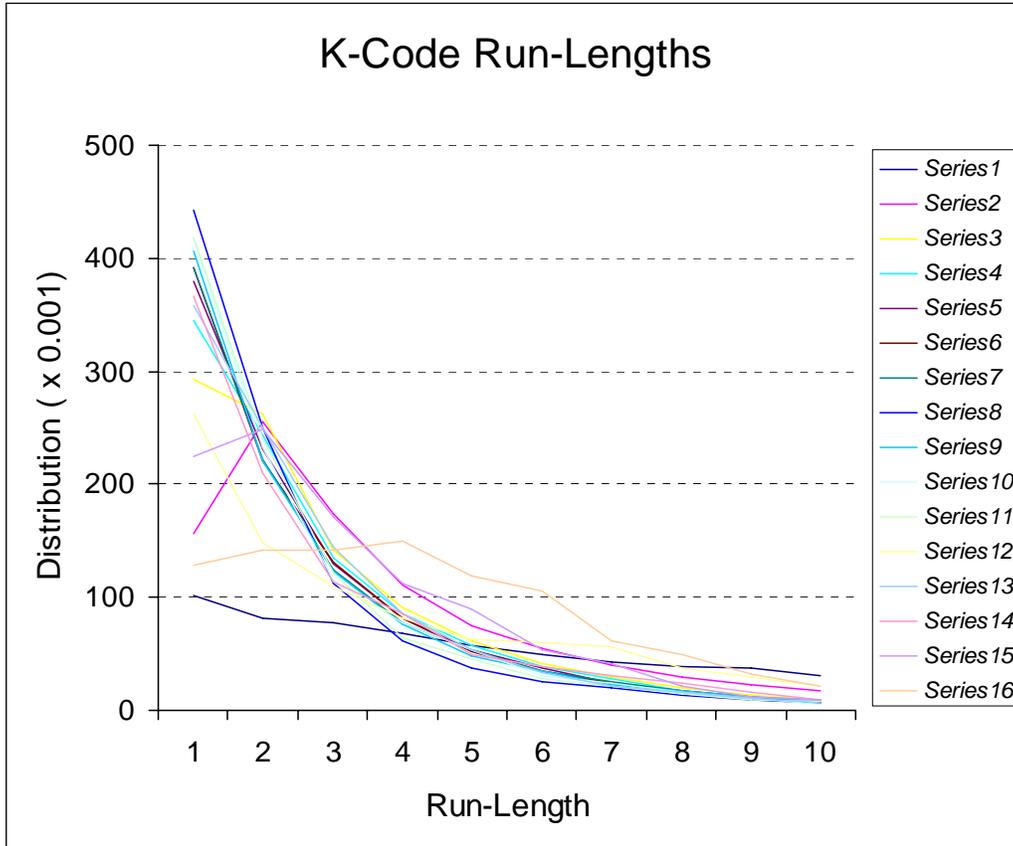
14 Figure 6.4 K-coding error

showing the difference between the original image and the *k*-coded version

6.4 K-CODE RUN LENGTH PROBABILITY DISTRIBUTION

The run length probability distributions of k -codes from a real image are very different from those from an array of random pixels. Different tissue types, through different k -means clustering process, may generate slightly different codebooks. To maintain pair-wise comparison consistency across multiple tissue types, k -codes are ordered according to the centroid pixel color intensities they represent. Since each pixel in the RGB color model has three color intensity values, we define that the relative order between two arbitrary pixels is determined by any two of the three values with the same ordering. A pixel takes precedence if two out of its three color channels take precedence in relative to another pixel, i.e., one of the two pixels is considered “bigger” if two of its three R, G, B intensity values are bigger. This ordering is guaranteed to be unique with the RGB color model.

Since the images used in this study are not considered directional, a run-length of any k -code is defined as a continuous, horizontal run of pixels of the same code. Vertical patterns should carry the same information, and are therefore redundant and not taken into account. This assumption will not always hold true with other image types under different conditions. Code run length probability distribution is a function of code density and collocation pattern. Under a particular condition, i.e. at a given image magnification, with the same number of pixel clusters, code run length probability distribution is a property of the image content itself.



15 Figure 6.5. K-code run-length probability distribution of one image

Plot of run length probability distributions of the k -coded image with $k = 16$. All run length probability values in this plot have been multiplied by the corresponding run length l , so that they represent the relative probabilities of pixels to be a part of run length l . The purpose of doing this is to show the small probability values clearly. The raw values of all the run-length probabilities of a code should sum up to 1.

6.5 RUN-LENGTH PROBABILITY DISTRIBUTION FEATURE FOR CONTENT CLASSIFICATION

To test the usefulness of run-length probability distribution feature for the purpose of tissue classification, two tissue types, brain and thyroid, were chosen and images were captured from routine glass slides. On one single slide, multiple, non-overlapping fields could be captured. The images were processed with k -coding using a k value of 16, and run length probability distributions have been computed. The two figures below show both average values (dots) and confidence intervals (bars), which indicate the within-group distances. Only the first 8 run-length probability values from each k -code distribution are shown. All distributions appear one-by-one in a row. The probability values are multiplied by 1000 before plotting.

The result as shown in Figure 6a and 6b indicates that the differences between the averages of the corresponding probability values are significant enough when compared with the within-group differences, so that the two tissue types can be readily distinguished from each other using the run-length feature alone.

16 Figure 6.6. K-code run-length probability distribution

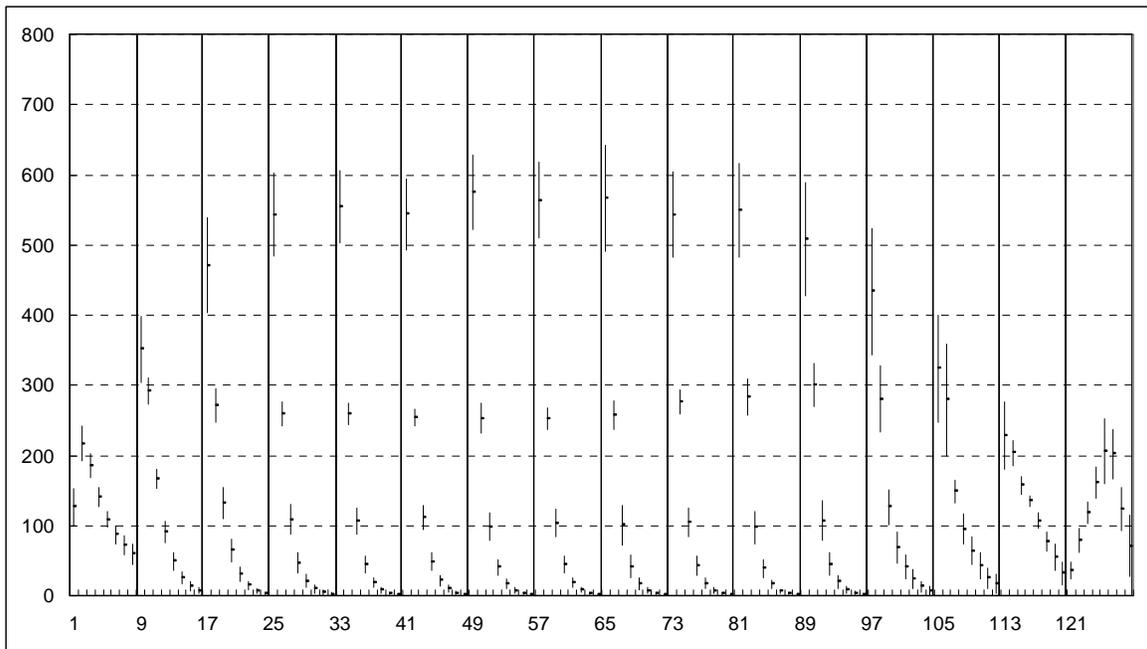


Figure 6.6a. *K*-code run-length probability distribution of images from one tissue: brain

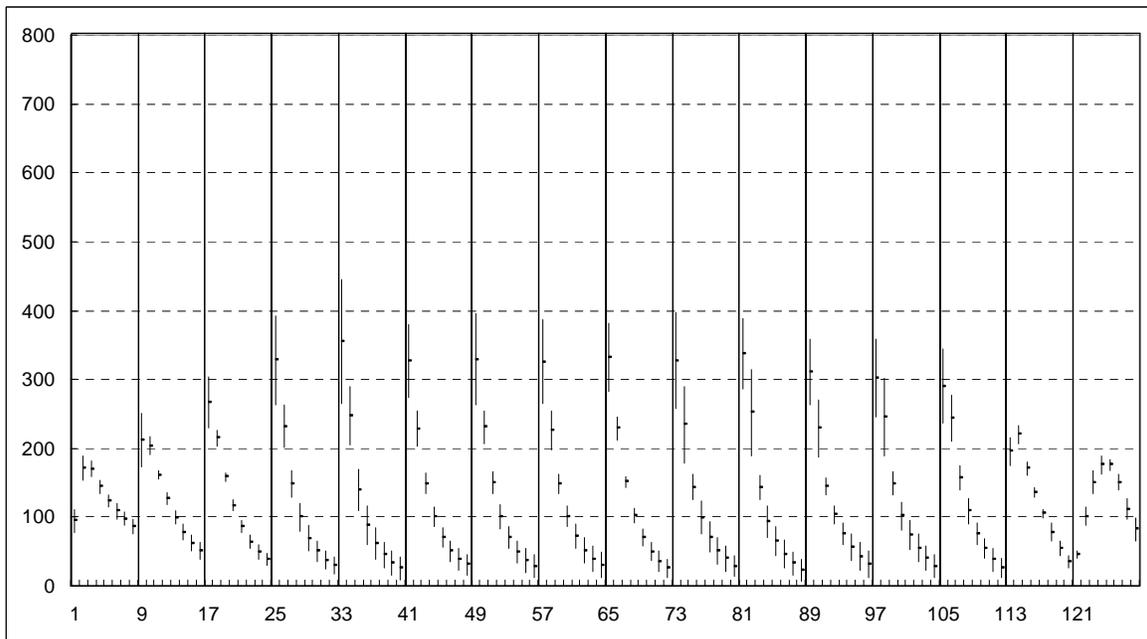


Figure 6.6b. *K*-code run-length probability distribution of images from one tissue: thyroid

6.6 CONTENT-BASED IMAGE RETRIEVAL PERFORMANCE

To assess the overall performance of k-code run-length probability distribution as image content descriptor across multiple tissue types and to investigate the possibility of using the feature for the purpose of content-based retrieval, the feature from different images were compared and the dissimilarity was measured quantitatively according to a distance metric. Images were captured from tissue micro-array slides with one image from each of the spots, and processed as above in section 6.5. Pixel run-length was counted up to 40, and maximum likelihood probability distribution was computed for each pixel code of the image. C code was developed to batch-process all the images and the generated the probability distribution data was stored in files.

Each image in the collection was used as query to compare against all the rest of the images and the distances were computed according to the original Kulback-Liebler distance formula. The top ten (10) images that showed the lowest scores in distance metric were recorded and used to compute the precision and recall scores. Class-wise precision and recall were computed as the average of the images of that tissue type. A Python script was used to take the probability distribution data from the previous stage, compute the pair-wise Kulback-Liebler distances, and then, using diagnoses as a gold standard to quantify the precision/recall retrieval performance metric.

Meanwhile, to allow examination of individual cases interactively, a simple graphical user interface has also been developed to show individual query sessions, using Python and wxWidgets (<http://www.wxwidgets.org/>), an “open source, cross-platform native user interface framework”, a suite known as wxPython (<http://www.wxpython.org/>). The interface allows user to select a query image from the archive by choosing from a list of displayed icons, then, execute the search command. The query, image as well as the ten (10) best match images with the

smallest distance measures, are displayed for each query session.

The results from batch mode processing are summarized in the Table 6.2 and Figure 6.7. Individual cases are further analyzed using the graphical interface to interpret the results in the light of pathology domain knowledge, and the findings are discussed in detail in the following section.

Table 6.2. Content-based Image Retrieval Performance

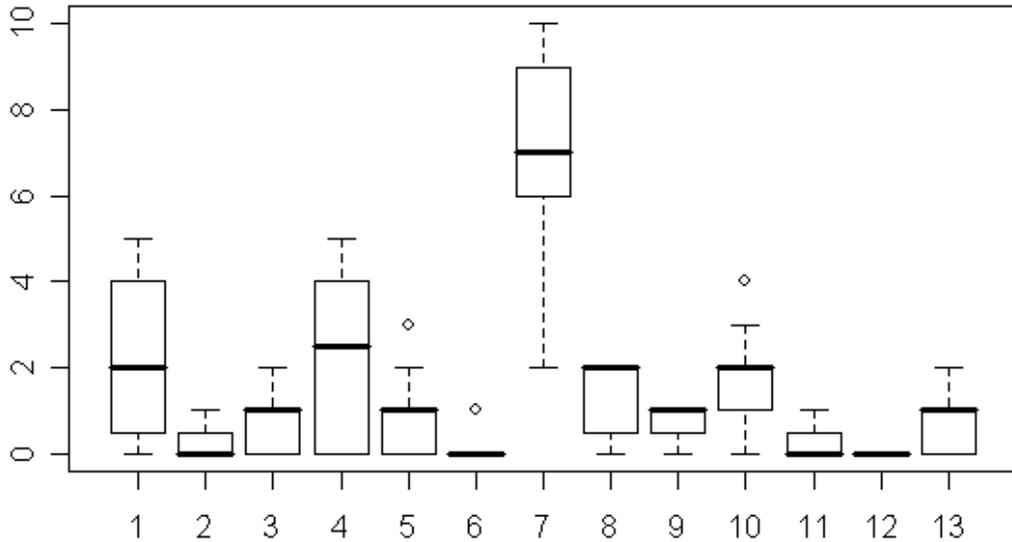
Tissue Type	R₁₀ / Max	P₁₀	Images
<u>Autolysis</u>	0.32 / 1.0	0.22	8
Autolyzed Renal Tubules	0.08 / 1.0	0.02	4
<u>Brain</u>	0.18 / 0.91	0.21	12
Fibrous Tissue	0.06 / 0.83	0.07	13
<u>Heart Muscle</u>	0.11 / 0.15	0.73	68
Heart Muscle and Fat	0.05 / 1.0	0.02	5
Kidney	0.33 / 1.0	0.06	3
Kidney Tubules	0.19 / 1.0	0.13	8
Benign Prostate	0.05 / 0.66	0.08	16
Prostate Cancer Gleason 3	0.09 / 0.5	0.19	21
Prostate Cancer Gleason 4	0.04 / 1.0	0.02	7
Prostate Cancer Gleason 5	0 / 1.0	0	2
Prostate Fibrous Tissue	0.10 / 1.0	0.09	10

Every image in the collection was used as query to retrieval similar images and the results were averaged for each of the different tissue types. The precision and recall values, referred to as P₁₀ and R₁₀, are the class-wise average of first ten (10) retrieved images. Max is the maximum possible recall value at the top 10 retrieval point, defined as $= 10 / (\text{total number of images} - 1)$ or 1.0, whichever is smaller. The last column shows the numbers of cases for the tissue types.

4Table 6.3. CBIR performance confusion matrix

Tissues	A	B	C	D	E	F	G	H	I	J	K	L	M
A	2.3	1.4	0.0	1.4	0.4	4.4	0.0	0.0	0.0	0.0	0.1	0.0	0.1
B	0.0	0.3	0.3	0.3	0.5	8.0	0.0	0.0	0.0	0.5	0.0	0.0	0.3
C	0.0	0.0	0.8	0.0	1.1	3.9	0.1	0.3	0.1	2.6	0.1	0.0	0.9
D	1.7	0.3	0.0	2.2	1.1	4.2	0.1	0.0	0.0	0.1	0.2	0.0	0.3
E	0.6	0.0	0.3	0.9	0.8	5.0	0.5	0.2	0.0	0.6	0.2	0.0	0.8
F	0.1	0.0	0.3	0.1	0.4	7.4	0.3	0.3	0.0	0.4	0.0	0.0	0.8
G	0.0	0.0	1.0	0.0	0.6	6.4	0.2	0.0	0.0	1.0	0.2	0.0	0.6
H	0.0	0.0	0.3	0.0	1.0	4.3	0.3	0.7	1.7	0.7	0.3	0.3	0.3
I	0.0	0.0	0.4	0.0	0.5	3.0	0.9	1.5	1.4	1.4	0.0	0.9	0.1
J	0.0	0.0	0.8	0.0	0.5	4.6	0.3	0.4	0.3	1.9	0.0	0.0	1.1
K	0.0	0.0	0.4	0.1	0.6	6.0	0.1	0.0	0.0	1.6	0.3	0.0	0.9
L	0.0	0.0	1.0	0.0	0.5	5.0	0.0	0.5	0.5	1.5	0.0	0.0	1.0
M	0.0	0.0	0.6	0.0	0.6	5.9	0.0	0.3	0.0	1.6	0.1	0.0	0.9

Only top 10 retrieved images are counted. The number is the average of all query sessions of the tissue type. Tissue labels: A) Autolysis; B) Autolyzed Renal Tubules; C) Benign Prostate; D) Brain; E) Fibrous Tissue; F) Heart Muscle; G) Heart Muscle and Fat; H) Kidney; I) Kidney Tubules; J) Prostate Cancer Gleason 3; K) Prostate Cancer Gleason 4; L) Prostate Cancer Gleason 5; M) Prostate Fibrous Tissue



17 Figure 6.7. Content-based Image Retrieval Performance

Box-whisker plot of class-wise P_{10} . Tissue types: 1) Autolysis; 2) Autolyzed Renal Tubules; 3) Benign Prostate; 4) Brain; 5) Fibrous Tissue; 6) Heart Muscle and Fat; 7) Heart Muscle; 8) Kidney Tubules; 9) Kidney; 10) Prostate Cancer Gleason 3; 11) Prostate Cancer Gleason 4; 12) Prostate Cancer Gleason 4; 13) Prostate Fibrous Tissue.

6.7 PERFORMANCE ANALYSIS AND INTERPRETATION

6.7.1 Interpretation of Performance Metric

Precision and recall are two canonical performance evaluation metrics that have been long accepted in text information retrieval as well as in image information retrieval. One routine to measure precision and recall is to compute percentage of correctly retrieved documents and the

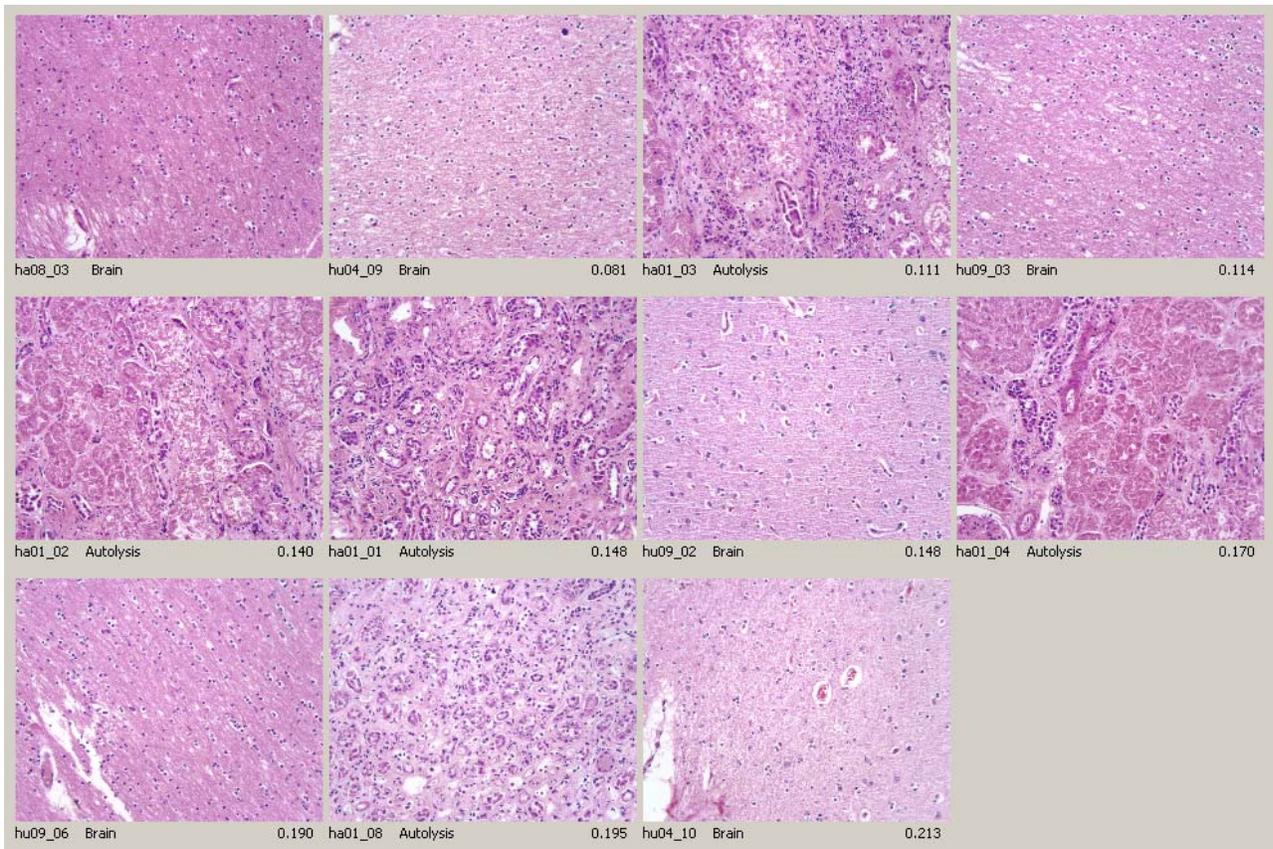
percentage of qualified documents present in the top ten retrievals. This method was adopted for it was not appropriate to plot the precision/recall curves in extended test ranges with remarkable variation in total number of images available in each of the tissue types.

In this particular research, extra attention should be paid when interpreting the final results of the retrieval performance. Due to the scarce availability of certain tissue samples, the collected tissue micro-array slides fail to provide enough cases for some of the tissue types or subtypes (see Table 6.2). The size of the tissue spots on the tissue micro-array slides imposed restriction in selection of a representative field of the intended tissue type. The captured views often consisted of more than one tissue. Some organs (from autopsy) might have undergone autolysis to various degrees with changed morphological feature at the histological and/or cellular level, before the tissue sample was processed. The data collection, including imaging and diagnosis, was completed well before the beginning of algorithm development and testing. In fact, the data was used to test the algorithm design rather than to train the algorithm or tune the parameters. All these would affect the behavior of the retrieval system and its performance evaluation metrics, which are not commonly seen in other retrieval systems with a sufficiently large corpus.

One important consequence of the distribution of sample images is that several tissue types with a small number (sometimes less than 10) of total cases would never see a high P_{10} value. One example is brain images, which displays one of the higher precision scores averaged over 21% for all 12 cases, 3 of which reach $P_{10} = 50\%$, another three reach $P_{10} = 33\%$. Considering the low sample density of brain tissue in the population, the precision measure is a positive proof of retrieval performance for the brain tissue. The R_{10} is averaged at 18%.

Recall value could be very different if the tissue samples are abundant in the image

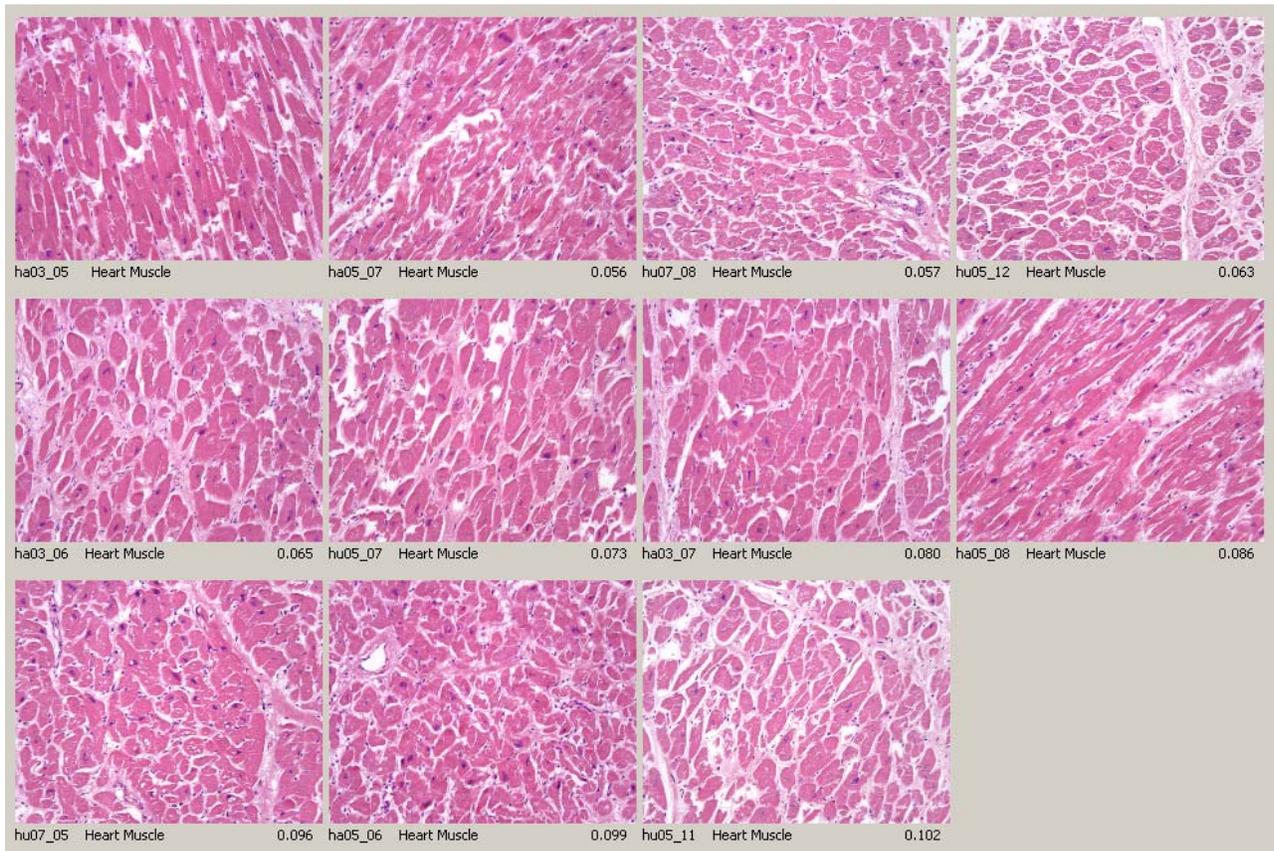
collection. Quite the opposite of the case with brain tissue, over 1/3 of the all the images are heart muscle (total 68 cases), which means assuming random distribution, the background precision would be around 30% at any given point. While the precision metric shows an average value of 73%, 15 heart muscle images measure 100% for P_{10} , another 12 images see a P_{10} value of 90%, etc., which is proof of the effectiveness of the retrieval algorithm. However, due to the density of heart muscle images in the population, even with a P_{10} value of 100%, the biggest possible R_{10} value is still below 15%. In the performance test, the average R_{10} is only 11%, much lower than that of brain images. In this particular case, the relatively low R_{10} doesn't reflect the actual retrieval performance of the algorithm.



18 Figure 6.8. Query session: brain

This is a screen crop of the GUI showing a query session using a brain image as query. The

query image is always at the upper left corner. Top ten retrieved images with the smallest distance values to the query are displayed from top to bottom, left to right. Under the image, the first string shows the image ID, which is assigned at the image capturing time and unique to each individual image. The next string is the tissue type tag from the diagnoses made by pathologist at UPMC (kind help from Dr. John Gilbertson) right after image capturing, well before the algorithms were developed. For those retrieved images, there is a third number, which is the summation of the Kulback-Liebler distances between the corresponding features from the query image and those from the retrieved image. This particular query session shows $P_{10} = 50\%$, $R_{10} = 45\%$. The top 10 retrieved images see contamination of 5 autolysis images besides 5 brain images.



19 Figure 6.9. Query session: heart muscle

$P_{10} = 100\%$, $R_{10} = 11\%$. In this case, the relatively low R_{10} doesn't reflect the algorithms' performance. Total 15 heart muscle images achieve 100% for P_{10} , and 90% for another 12 images.

6.7.2 Pathology Classification and Specificity of Image Feature

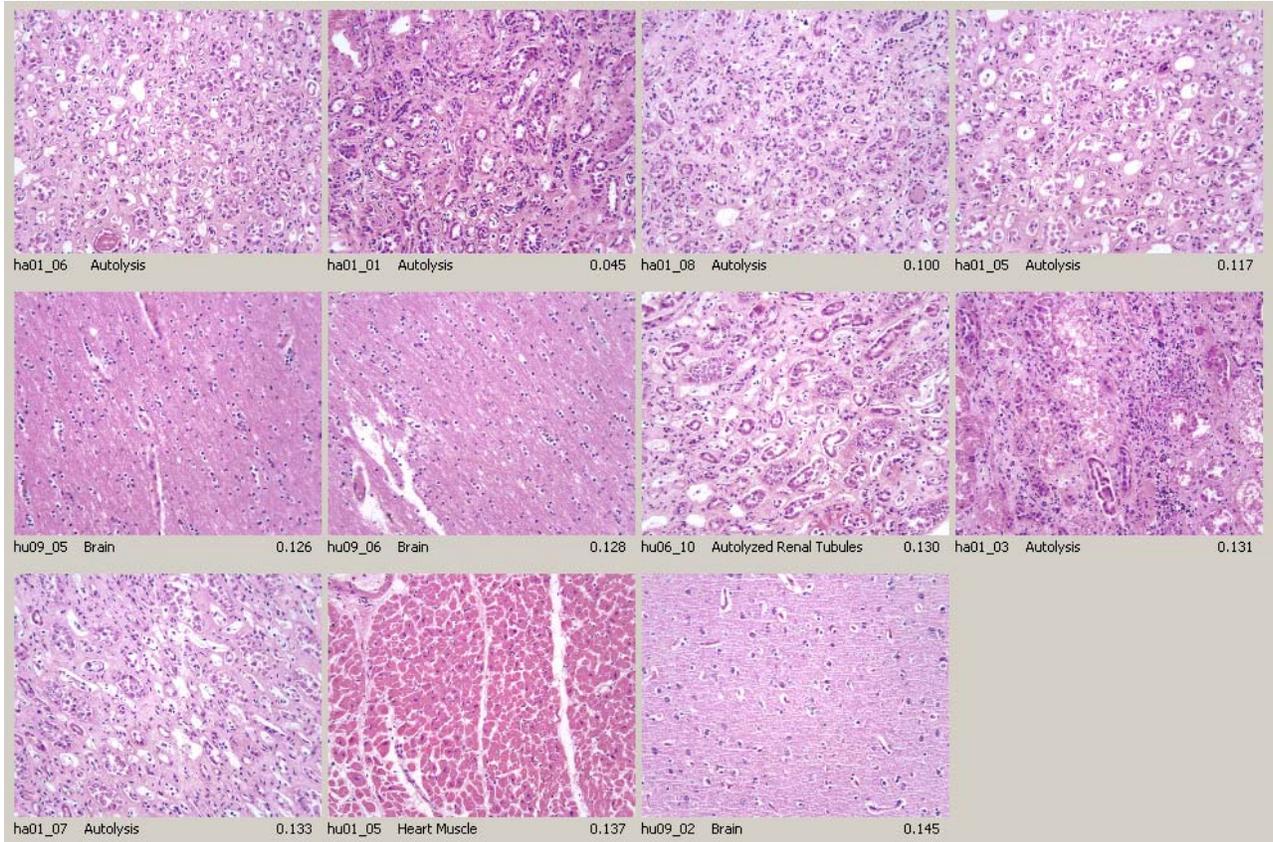
As a matter of fact, pathologists are trained to find diagnostic evidences from the complication and multiplication of visual features, while individual machine vision algorithms are either highly specific for one kind of image content or effective across a wider range of image contents but with low accuracy when evaluate the performance individually. Algorithms can be developed to recognize one kind of object, such as to identify cell nucleus, or to measure only one visual feature from objects across multiple types, such as to measure the chromaticity of the stained tissues of various sources and processing protocols.

In the problem of medical image classification, one example of such difficulty is that many diagnostically different tissue cases, or images may actually look very similar, and different terms may actually refer to almost the same thing, so that the machine vision algorithms tend to identify two or three visually similar tissue types as one group with satisfactory accuracy, and yet, the performance evaluation could still indicate disappointing metrics.

One possibility is that the image feature could actually be of high accuracy and specificity and effective when handling several tissue types as a super-class. Further research is needed to find extra, second features for finer grain classification. Another possibility is that, sometimes, it is quite reasonable to combine two classes with different name tags and treat them as one, considering the fact that ambiguity is not uncommon among human pathologists dealing with similar cases. It is often important to look into the individual cases to better understand the behavior of machine vision algorithms.

In figure 6.8, when searching with the brain image, the particular P_{10} and R_{10} values, as in this case, are 50% and 45% respectively, which indicates that half of the top 10 matches are correctly brain images and about half of all the brain images in the collection are recognized.

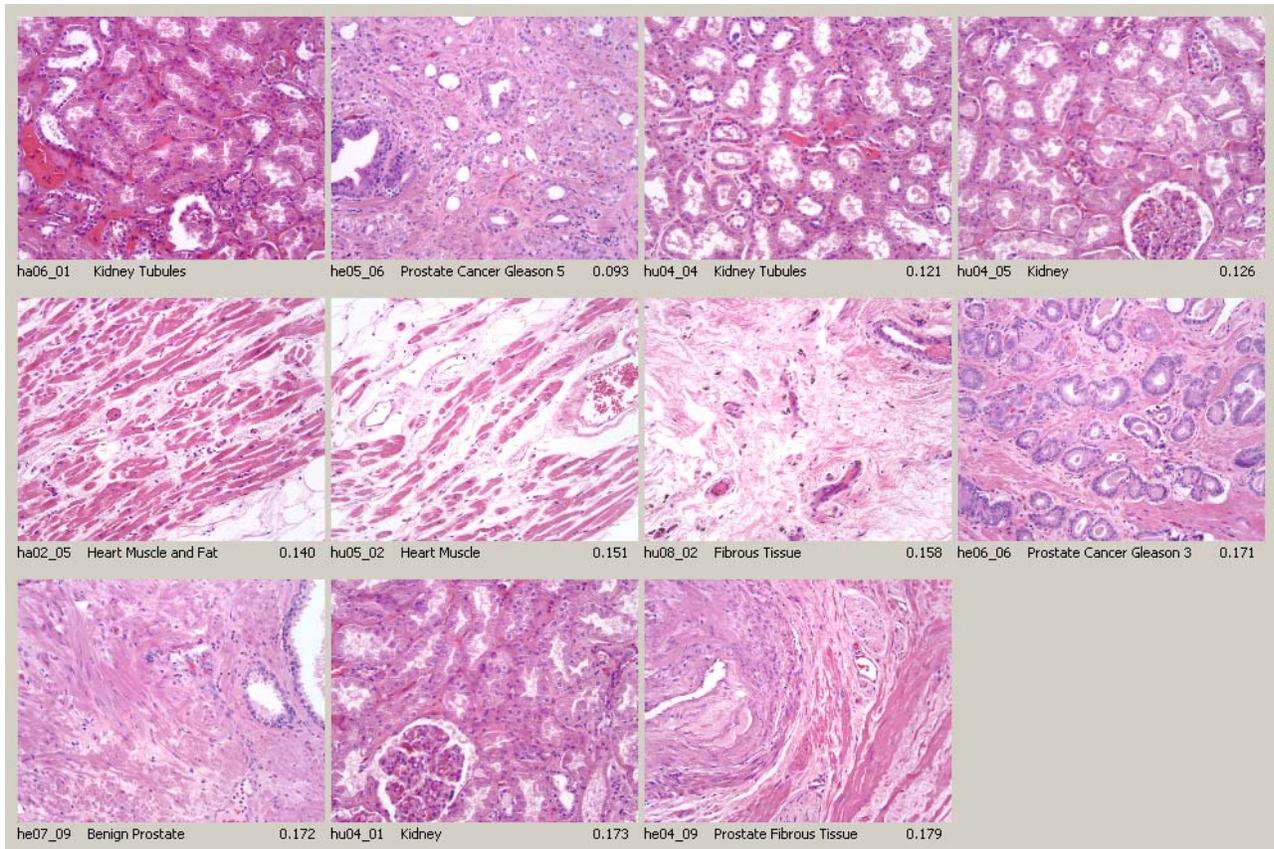
Taking a closer look at those incorrect matches, we see that the other five mismatches are all images labeled as “autolysis”, with P_{10} and R_{10} values of 50% and 62% respectively, if the query session were to be treated as an autolysis session. All the images labeled as autolysis are clearly not of brain origin, and the morphology of autolyzed tissues as in this study doesn’t represent any natural existence of biological structure. However, the chromaticity and the scale of some structural elements, both of which the k-coding feature is expected to measure, of both brain images and autolysis images do bear certain similarity as oppose to other tissue types. The retrieval performance analysis indicates that the algorithm shows certain degree of specificity and accuracy for two of them but is not sufficient to tell them a part when used alone. Autolysis images also appeared with certain probability in other query sessions employing brain image as query, and this appears to be a common phenomenon.



20 Figure 6.10. Query session: autolysis

$P_{10} = 50\%$, $R_{10} = 70\%$ (retrieved 5 out of total 7 autolysis images).

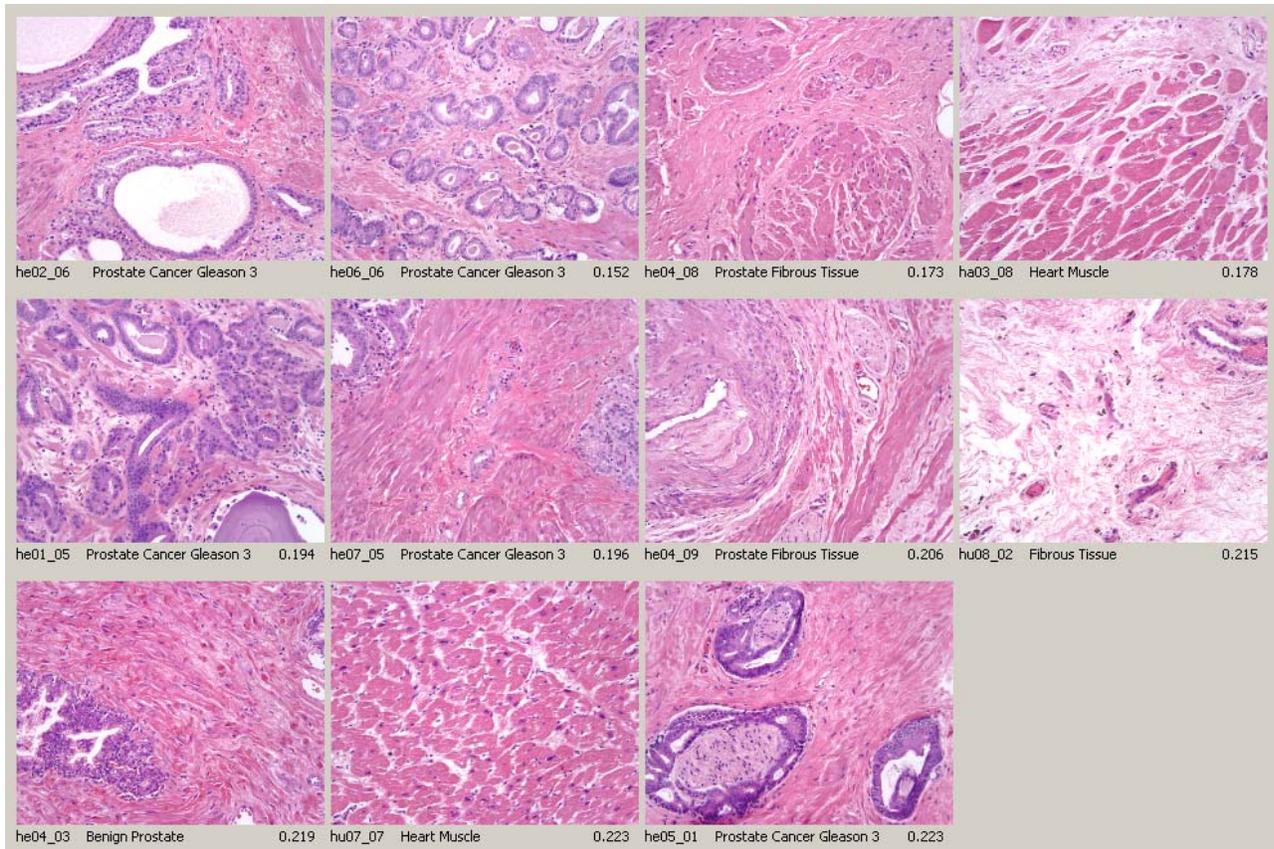
Figure 6.11 shows a session using an autolysis image as query. The retrieved images include 5 autolysis images, 1 autolyzed renal tubules, 3 brain images, and 1 heart. This is in accordance with the observation and interpretation discussed previously.



21 Figure 6.11. Query session: kidney tubules

$P_{10} = 10\%$, $R_{10} = 14\%$, (retrieved 1 out of total 7 kidney tubules cases), plus 2 out of total 3 kidney images.

Another less obvious example is one kidney tubules query session as shown in Figure 6.11. According to the result automatically computed by the Python script, only one image is labeled as “kidney tubules” as the query among the top 10 retrieved images. However, there are two other images labeled as “kidney” (out of total three kidney images in the entire collection, $R_{10} = 67\%$), which are morphologically very similar to kidney tubules images and also of the same origin. The fact that four other images are of prostate origin prompts that some morphological features are shared among the tissues of these two origins.



22 Figure 6.12. Query session: prostate cancer Gleason's grade 3

$P_{10} = 40\%$, $R_{10} = 20\%$, (retrieved 4 out of total 20 kidney tubules cases), plus 4 other tissue samples of prostate origin.

The query session as in figure 6.12 shows that although 4 Gleason's grade 3 images bring the P_{10} value to 40%, there are 4 extra tissue samples of prostate origin that bear similar features as the query image. During the preprocessing step, any white spaces were intentionally ignored. Prostate fibrous tissue images are also found in other prostate cancer query sessions.

6.7.3 Within-Class Divergence

Just like what has been observed as can be called between-class convergence, in which two or more kind of images with different class labels share similar image features, it is no surprise to find examples of within-class divergence, where images with the same diagnostic label actually

show very different responses to the same feature detection algorithm. The fact is that the visual features of some of the images bearing one particular tissue type tag are actually very different from the rest of the images from the same class. This contributes to the varying retrieval performance.

Two possible situations may contribute to the ambiguity. The boundary between two tissue types, as seen in the collection of images in this dissertation, is not defined well enough. This brings arbitrariness to the determination of some difficult cases. Sometimes, the meaning of two different terms may not be very different after all, such as shown in “kidney” and “kidney tubules”. Another part of the ambiguity arises from the fact that many tissue images are not pure with only one single type of cells, and/or one single type of arrangement of cells. One organ may consist of multiple tissue types. Every spot on a tissue micro-array slide, every image field may contain one diagnostically significant tissue type that the sample is labeled after, and other auxiliary tissues, such as fat tissue, and fibrous tissue, both of which are present in many of the images.

The scale of the image features that are important for tissue classification sometimes may be bigger than the other so that the dimension of the image used for this study is not sufficient for the features to be statistically stable. All images were captured at the same magnification which might overlook some important features that were better examined at different detail levels.

Artifacts were also introduced when the tissue sample was not processed correctly as in the case of tissues with autolysis. All these bring in noise to the features of interest so as to blur the classification boundary of different tissue types as well as make the images of the same classification display significant variation in their response to feature detection. The within-class

variation of image features is shown as the within-class variation of P_{10} values. Some images allow retrieval of far less relevant images than some other from the same class.

All the issues and interpretation discussed above have been observed in other image retrieval system performance evaluation analysis using precision and recall as metrics, and may reflect the deficiency in higher level reasoning capability of machine vision.

6.8 SUMMARY

The image retrieval performance analysis is performed to evaluate the efficacy of k-coding and run-length probability distribution as a complex image content description feature across multiple tissues with various levels, different natures of pathology. Some of tissue types give better performance than the other. Generally, the feature set works better with heart muscle, brain, works reasonably with kidney (including kidney and kidney tubules) tissues, and with sub tissue types or special morphology as in tissue with autolysis (including autolysis and autolyzed renal tubules) and Gleason's grade level 3 prostate cancer. In some cases, the performance is considered good for an image content description algorithm using one single image feature and the simplest distance measure. Overall, these show a satisfactory coverage of tissue morphology that is beyond what have been observed with the two-class classification task in the first stage of this research.

The better performance with tissues such as heart muscle, brain, and even autolysis indicates that the feature set shows its advantage dealing with images with relatively consistent cell type and homogeneous morphology. The histology and pathology of the rest of the tissue

types form relatively complex structures with more boundaries, and hierarchy of regions. This may complicate the behavior of the feature detection in two ways: 1) the complicated image patterns are of bigger scale than the dimension of individual cells, and thus, require a bigger image dimension to maintain the statistical stability of the run-length probability distributions; 2) complex patterns in biological structure form hierarchy, while the simplest run-length model treats an image as “a bag of pixels”. High level morphological change that is important for the tissue histology and pathology are not reflected in the run-length feature. It is possible to build up extra run-length features at multiple scale levels similar to a tree-like structure or a pyramidal representation as in wavelet transform.

The performance also degraded in tissues with irregular morphology or tissues with surrounding tissue contamination, such as fibrous tissue, heart muscle and fat. Cancerous tissues suffer from dramatic morphological change tend to confuse with other tissue types with unrelated origin.

Several factors may contribute to the performance degradation:

- 1) The stability of feature suffers from the restriction imposed by the selected (or unselected) field of view of the tissue. The tissue micro-array samples didn't allow any flexibility for selection of view as one spot is only slightly larger than a captured field at 20x magnification with >1 mega pixels. Some tissue samples contain surrounding tissues that are not meant to be a part of image content, as well as artifacts before and during the course of tissue processing. The size of the image doesn't support enough statistical stability and the scale and magnification is fixed so that some important features for tissue classification are not easily identifiable.
- 2) The pathology classification system used to label the images doesn't reflect the

underlying connections. While two or more class labels may connect to one tissue origin, there could be morphologically very different sub-types in one class. Both could hurt the performance metrics greatly.

- 3) The morphological feature that the algorithm captures may be shared by two or more tissue types. The result is that the algorithm does a good job in picking out two different tissues but the performance measure of this mixed retrieval doesn't reflect the actual efficiency. Single image feature can never be enough to solve all the problems.

Machine vision is hampered by general deficiency of reasoning capability compared with human visual perception, which allows us to solve the above issues and to deal much more complicated situations.

7.0 CONCLUSION AND FUTURE WORK

7.1 PATHOLOGY MICROSCOPIC IMAGES AND RUN-LENGTH FEATURE

The goal of this thesis is to research on a new kind of feature extraction and similarity measure that is suitable for pathology microscopic images, adaptable to multiple image types, and yet simple and efficient thus make it suitable for the task of content-based image retrieval. Instead of customizing for specific image content for the purposes such as morphometric analysis, object recognition, content-based retrieval is oriented to design algorithms that take few assumptions about the image content so that they can be applied to a relatively broad range of image types. Besides, like other information retrieval systems, computational efficiency is important for content-based image retrieval in order to handle large data sets.

The design is based on the fact that the tissue morphology consists of the scale and arrangements of cellular components and of cells. With H&E stained tissue samples, the image components are artificially colored according to their physic-chemical properties, which determine their affinity to two chemicals, Hematoxylin and Eosin. The components stained with Hematoxylin appear purple, and those with Eosin, red. Many important pathology diagnoses are based on the morphology change by examining the H&E stained tissue samples. Color quantization and color code run-length probability distribution feature is designed to target such displayed morphology in a global scale. Code run-length distribution is a way to model objects, and regions in image, the scale and distribution of them, in a one-dimensional fashion.

7.2 FEATURE EXTRACTION AND SIMILARITY MEASURE

Image feature extraction is, to a significant degree, tied to the particular hardware platform, and the targeted image content. The performance often varies when there is a change in the imaging system, or the algorithm is to be used to handle a different kind of images. To automatically adapt to a broader range of tissue types and different capturing systems, the same k-means unsupervised clustering technique is applied to all tissue types in the collection in the same fashion. As an unsupervised algorithm, k-means clustering is able to adapt to the data distribution of individual images when generating an essential palette as the codebook. The only system-wide parameter chosen with certain arbitrariness is the number of total clusters in the k-means.

It is assumed that different types of images should have different codebooks with color codes representing different image components. Because of this, it is also assumed that, in most cases, the resulting color code distributions would be very different across different tissue types, as they are not strictly comparable. The images from the same class with similar coloration and geometry are most likely to display similar patterns in codebook composition and in code run-length probability distributions. Thus, average Kulback-Liebler distance is used as a measure for the matching of the codebook composition and the similarity of the corresponding probability distributions. This is defined, rather than “learned” from the distribution of the data itself, based on the theory of mutual information measure of the distributions.

The combination of an unsupervised feature extraction algorithm and a mathematically defined distance measure has the implication that the whole process is automated without much human intervention or data dependent training. Without a training stage, the data collected was used for the purpose of testing the performance of the system. Such an approach imposes few restrictions in applying the algorithms to include more tissue types, or even to other artificially

colored images, such as microscopic medical images stained with other biochemical dyes or probes.

7.3 RUN-LENGTH FEATURE FOR TISSUE CLASSIFICATION AND CONTENT-BASED IMAGE RETRIEVAL

Overall, the described run-length feature performed well with certain tissue types as shown in the classification test with brain tissue and thyroid tissue, and also in the content-based retrieval test with brain tissue, and heart muscle, both of which are relatively homogeneous in the morphology of differentiated cells – neuronal cells and heart muscle cells with few contamination of other cell types – and the cells are laid out in regular patterns with few complication with extra structures such as sinusoid, gland, and fibrous tissue. The feature also performed reasonably well with kidney and prostate cancer tissues.

Besides the complication of tissue structures, it is possible that when the number of classes grows larger, due the problem of dimensionality curse in dimensional feature space, the efficiency of the distance measure may also degrade.

The most important observation in the performance analysis is the variation of response of the images from the same tissue type to the feature detection algorithm. While the retrieval performance of some images was very satisfactory, the rest was poor. The reasons for such a variation were attributed partially to the particular nature of pathology diagnosis, partially to the limitation of the system, such as the chosen image standard, the limitation of the feature extraction algorithm when dealing with complex histological structures. Further research is necessary to address this problem in order to improve performance and stability.

7.4 FUTURE WORK

This dissertation represents an initial effort in solving the complex content-based image retrieval problem along a new path. Many aspects of the new approach are worth further investigation. Especially, three major issues deserve further research effort in order to achieve better performance.

7.4.1 Image quality control

Image quality is important to almost every image processing and content analysis research project. Digital image analysis techniques are sensitive to the quality and standard the images are recorded and digitized. The statistical stability of the feature values is an important quality measure of the features. Larger image size improves the statistical quality of the features without changing other parts of the feature extraction algorithms by reducing the within-group variation. The results as shown in the previous chapter indicate that tissue types that are better characterized with simpler morphological features see better retrieval performance than tissue types that are known for more complex morphological features. This suggests that it is possible to obtain better results for some of the tissue types by increase image dimension. This is based on the morphological nature of the content of pathology microscopic images.

With the advances in imaging hardware, it is possible to use bigger CCD chip to capture a bigger field of view at a time. This is important in order to compensate for the increased complexity in morphology in tissues other than the simplest ones, such as brain and heart muscle. Software tools can also be used to stitch small fields into a big image. This is expected to make the features statistically more stable and also allows for extraction of complex features to

be explored based on the basic run-length features. In this research, a magnification of 20 x, which translates to about 1 micron per pixel, was used exclusively for all images. While there have been no obvious problems with this, different magnifications can also be tested to find the best parameter combinations for each of the different tissue types. An organized diagnostic classification scheme can also help to better interpret the results and maintain consistency.

7.4.2 Rotation invariant and two-dimensional features

One-dimensional single run-length probability distribution, as described in this thesis, is the simplest form of a class of possible morphological feature descriptors for the purpose. It takes into account only the one-dimensional run-length of the codes. This is a good characteristic in many ways and suitable for many tasks such as with medical images. However, many other images, such as natural scenic images, are directional, which means the run-length features would be very different if images are analyzed in different directions. This can be used to design a simple method to show the effect, with which run-length features are extracted from the same image but rotated by a series of degrees from the original. It can be anticipated that if the image content is directional and is not rotation invariant in a global sense, the run-length feature will show periodic changes with magnitude corresponding to the degrees of the rotation. This is not expected from the majority of pathology microscopic images in this research, as neither the tissue morphology (with only a few exceptions, such as muscle tissues) at the magnification level they are imaged nor the preparation and imaging procedures imposes any directional properties to the image.

One way to extend run length feature to image content that is not rotation invariant is to align images to be compared along the same canonical direction. The canonical direction can be

defined according to the direction of the features so as to allow maximum regularity patterns and least random interruptions in that direction, such as a direction that maximizes overall code run-lengths. This can be carried out in the compressed domain using a frequency domain representation of the image. After the canonical direction has been identified, images are then rotated to align with that direction and run-length features are analyzed. All the feature extraction and feature comparison techniques as described before can be applied to the rotated images without change. If necessary, the images can be rotated to align with more than one directions and multiple sets of the run length features can be analyzed to reflect the two dimensional nature of the image content.

Based on the same color quantization methodology, it is also possible to extend the scheme to model two-dimensional or even three-dimensional distributions of color code blocks. This allows for better approximation of objects and regions using code runs and/or code blocks. However, by defining higher dimensional features, extra assumptions have been made about the geometrical and topological properties of the image content which may not hold with the actual image. For example, if a method is used to measure the geometry of a region, then it is implied that the object has a defined close boundary. With one-dimensional run-length feature, it doesn't require any definition of boundary or shape for the features to be meaningful and operable. Still it represents a very attractive approach to expand the original idea of run-length feature and deserves in-depth research and innovation. It will be a challenge to improve the algorithm while maintaining the computational efficiency, theoretic simplicity, and generality, which have been important goals of this research.

First paragraph.

7.4.3 Feature extraction and modeling

K-means clustering is capable of adapting to the distribution of the image pixels in the color space. As an unsupervised algorithm, it is data dependent, which means, it is optimized for the data population rather than for the classification. Final clustering result is affected by the distribution of global population. It is possible to use other clustering/classification algorithms for the initial codebook design to achieve better coding efficiency. It is also possible to pre-process the image so that the subsequent clustering is less subject to the distribution of the data, but better suited for the purpose of classification. One possible technique is to use an alternative color representation than RGB color space.

Although Kulback-Liebler distance is solid in theory and is one of the most widely used measure for mutual information, there have been a few criticisms about the original version of Kulback-Liebler distance when applied to solve real world problems. A most common one is that Kulback-Liebler distance is not symmetric, which may not be critical when used as a distance measure for retrieval. Theoretically, the relationship between the query and the retrieved documents can be viewed as asymmetric, too. There are several modified versions of Kulback-Liebler distance that are commonly used to address the issue. Modifications have also been made to make the Kulback-Liebler distance smoother when dealing with a sparse distribution. This can potentially help to statistically stabilize the content-based retrieval performance.

The contribution of this research is that it has developed and tested a new methodology that provides a measure for such features that are difficult to approach using conventional methodology. From a greater perspective, k-means clustering and run-length probability distribution is a simplest form of a novel image feature extraction and comparison approach that is particularly suited for images that are artificially colored. It aims to adaptively model the scale

and distribution of objects and regions in the image based on the coloration in an efficient way. When used alone for the purpose of pathology microscopic image classification and content-based retrieval, it has demonstrated good performance for some of the tested tissue types, while the performance is less satisfactory and stable for other tissue types. This is comparable to the general performance of many basic image feature extraction algorithms. It is greatly anticipated that future work will improve the feature extraction algorithm, extend it to two-dimensional block pattern and multi-code co-occurrence, and apply it to solve more real world problems.

BIBLIOGRAPHY

- [IBM99] IBM Almaden Research Center. Technical summary of color descriptors for MPEG-7. MPEG-7 proposal P165, MPEG-7 Seoul Meeting, Mar. 1999.
- [Iowa State CS page] <http://www.cs.iastate.edu/jva/jva-archive.shtml>
- [SCL page] <http://www.scl.ameslab.gov/ABC>
- [antonini92] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," IEEE Trans. on Image Processing, 1(2):205-220, April 1992.
- [ardizzoni99] S. Ardizzoni, I. Bartolini and M. Patella. Windsurf: Region-Based Image Retrieval Using Wavelets. In Proceedings of the 1st Workshop on Similarity Search (IWOSS'99 - DEXA'99), Florence, Italy, September 1999.
- [arkin91] Esther M. Arkin, L. Chew, D. Huttenlocher, K. Kedem, and J. Mitchell. An efficiently computable metric for comparing polygonal shapes. IEEE Trans. Patt. Recog. and Mach. Intell., 13(3), March 1991.
- [bang02] Hoon Yul Bang and Tsuhan Chen, "Feature Space Warping: An Approach to Relevance Feedback", to appear in IEEE Int'l Conf. on Image Proc. (ICIP 2002), Rochester, New York, U.S.A., September 2002
- [bartolini00] I. Bartolini, P. Ciaccia and M. Patella. WINDSURF: A Region-Based Image Retrieval System. CSITE-011-00 Technical Report, July 2000.
- [beckmann90] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: An efficient and robust access method for points and rectangles." Proc. ACM SIGMOD, pp. 322-331, 1990.
- [besser95] H. Besser and J. Trant. 1995. Introduction to imaging: issues in constructing an image database. Santa Monica, CA: Getty Art History Information Program. 48 pp. [For Web version see: <http://www.getty.edu/research/institute/standards/introimages/> <<http://www.getty.edu/gri/standard/introimages/>>]
- [boggess01] A. Boggess and F. Narcowich. A First Course in Wavelets With Fourier Analysis. Upper Saddle River: Prentice Hall. New Jersey, 2001.

- [brandt99] S. Brandt. Use of shape features in content-based image retrieval. Master's thesis. Department of Engineering Physics and Mathematics, Helsinki University of Technology, 1999
- [brinkhoff93] T. Brinkhoff, H. Kriegel, and B. Seeger, "Efficient processing of spatial joins using R-trees," Proc. ACM SIGMOD, pp. 237-246, 1993.
- [brunelli99] R. Brunelli, O. Mich. On the use of histograms for image retrieval. Proceedings of IEEE Multimedia Systems '99 International Conference on Multimedia Computing and Systems [IEEE ICMCS99], Florence, Italy, June 7-11, 1999, pp. 143-147
- [buckley95] C. Buckley and G. Salton, "Optimization of relevance feedback weights," Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 351-357, Seattle, WA, July 1995.
- [callan99] Jamie Callan. IR class notes. 2000. (need to find out if they have published the book based on the syllabus.)
- [carson97] Chad Carson, Serge Belongie, Heyit Greenspan, and Jitendra Malik, "Region-based image querying," IEEE CVPR '97 Workshop on Content-Based Access of Image and Video Libraries. 1997.
- [chang89] S.-K. Chang. Principles of Pictorial Information Systems Design. Prentice Hall Intern. Editions, 1989.
- [chang93] T. Chang and C. C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," IEEE Trans. On Image Processing, 2(4):429-441, Oct. 1993.
- [chang93] T. Chang and J. Kuo. Texture analysis and classification with tree-structured wavelet transform. IEEE Trans. Image Proc. 2:429-441, Oct. 1993.
- [chang95] S.-F. Chang, "Compressed-domain techniques for image/video indexing and manipulation," IEEE Int'l Conf. on Image Proc. (ICIP95), Special Session on Digital Library and Video on Demand, Vol. I, pp. 314-316, 1995.
- [chen01] Y. Chen, X. S. Zhou, T. S. Huang, "One-class SVM for learning in image retrieval," IEEE Int'l Conf. on Image Proc. (ICIP'2001), Thessaloniki, Greece, Oct. 7-10, 2001.
- [chen02] Yixin Chen and James Z. Wang, "A region-based fuzzy feature matching approach to content-based image retrieval," IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(9):1-16, Sept. 2002.
- [chen94] J. L. Chen and A. Kundu, "Rotation and gray scale invariant texture identification using wavelet decomposition and hidden Markov model," IEEE Trans. on Pattern Analysis and Machine Intelligence, 16(2):208-214, Feb. 1994.
- [chuang96] Gene C.-H. Chuang and C.-C. Jay Kuo. Wavelet descriptor of planar curves: Theory and applications. IEEE Trans. Image Proc., 5(1):56-70, January 1996.

- [cox98] I. J. Cox, M. Miller, T. Minka, P. Yianilos, "An optimized interaction strategy for Bayesian relevance feedback," IEEE Conf. Computer Vision and Pattern Recognition (CVPR '98), 1998
- [cross89] G. R. Cross and A. K. Jain. Markov random field texture models. IEEE Trans. Pattern Analysis and Machine Intelligence, 5(1):25-39, 1989.
- [das99] M. Das, R. Manmatha, and E. Riseman. Indexing flower patent images using domain knowledge. IEEE Intelligent Systems, pp. 24-36. September/October 1999
- [deng01] Yining Deng, B. S. Manjunath, Charles Kenney, Michael S. Moore, and Hyundoo Shin, "An efficient color representation for image retrieval." IEEE Transactions on Image Processing, 10(1):140-147. January 2001.
- [deng99] Y. Deng and B.S. Manjunath, An efficient low-dimensional color indexing scheme for region-based image retrieval, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), p. 3017-20, 1999
- [eakins96] J.P. Eakins. Automatic image content retrieval - are we getting anywhere? Proceedings of Third International Conference on Electronic Library and Visual Information Research (ELVIRA3), pp. 123-135, De Montfort University, Milton Keynes, May 1996.
- [eakins96] J.P. Eakins. Automatic image content retrieval ?are we getting anywhere? Proceedings of Third International Conference on Electronic Library and Visual Information Research (ELVIRA3), pp. 123-135, De Montfort University, Milton Keynes, May 1996.
- [eakins98] J.P. Eakins. Techniques for image retrieval. Library and Information Briefings, 85. London: South Bank University, Library Information Technology Centre, 1998
- [eakins98] J.P. Eakins. Techniques for image retrieval. Library and Information Briefings, 85. London: South Bank University, Library Information Technology Centre, 1998.
- [eakins] John Eakins. <http://www.jtap.ac.uk/reports/htm/jtap-039.html>
- [faloutsos94] Christos Faloutsos, Ron Barber, Myron Flickner, Jim Hafner, Wayne Niblack, Dragutin Petkovic and Will Equitz. Efficient and effective querying by image content. Journal of Intelligent Information Systems, 3, 3/4, July 1994, pp. 231-262.
- [flickner95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: the QBIC system. IEEE Computer, pp. 23-32, Sept. 1995.
- [foley92] J. D. Foley, A. v. Dam, S. K. Feiner, and J. F. Hughes. Computer Graphics: Principles and Practice in C, chapter 13. Achromatic and Colored Light, pages 563-604. Addison-Wesley Publishing Company, Inc., 2nd edition, June 1992.

- [fukunaga90] Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, Academic Press, San Diego, CA.
- [gaede98] V. Gaede and O. Gunther, "Multidimensional Access Methods," ACM Comp. Surveys, 30(2):170-231, June 1998.
- [galloway75] Galloway MM. "Texture analysis using grey level run lengths," Comp. Graph. and Image Proc. 1975; 4: 172-179
- [gettyedu] <http://www.getty.edu/research/institute/standards/introimages/>
- [gevers00] T. Gevers and A. W. M. Smeulders, "PicToSeek: Combining Color and Shape Invariant Features for Image Retrieval," IEEE Trans Image Processing, 9(1):102-119, 2000.
- [glass96] Gene V Glass and Kenneth D. Hopkins, "Statistical methods in education and psychology," third edition, page 92-93. Allyn & Bacon, Needham Heights, MA 02194, 1996.
- [gonzalez92] Gonzalez, R. C., and Woods, R. E., Digital Image Processing, Addison-Wesley, Reading, MA 1992.
- [gotlieb90] Calvin C. Gotlieb, and Herbert E. Kreyszig. Texture descriptors based on co-occurrence matrices. Comput. Vis., Graphics, and Image Proc., 51:70-86, 1990.
- [graham98] M. Graham and J.A. Eakins. A prototype retrieval system for trade mark images. VINE, No. 107, pp. 73-80, 1998
- [guttman84] A. Guttman, "R-trees: adynamic index strcutre for spatial searching," Proc. of ACM SIGMOD, pp. 47-57, June 1984.
- [hafner95] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Effecient color histogram indexing for quadratic form distance functions," IEEE Trans. Pattern Anal. Machine Intell., 17:729-736, July 1995.
- [haralick73] R.M. Haralick, K. Shanmugam, and I. Dinstein. Texture features for image classification. IEEE Trans. On Sys. Man. And Cyb. SMC3(6):1345-1350, 1973.
- [heckbert82] P. Heckbert. Color image quantization for frame buffer display. Computer Graphics, Vol. 16, No. 3, pp. 297-304, 1982.
- [hirata93] K. Hirata and T. Kato. Rough sketch-based image information retrieval. NEC Research and Development. 34(2):263-273, April 1993.
- [hu62] M.K. Hu. Visual pattern recognition by moment invariants, computer methods in image analysis. IEEE Transactions on Information Theory. Vol IT-8, pp. 179-187, 1962

- [huang01] Thomas S. Huang and Xiang Sean Zhou, "Image retrieval with relevance feedback: from heuristic weight adjustment to optional learning methods." IEEE Int'l Conf. on Image Proc. (ICIP2001), Thessaloniki Greece, October 2001.
- [huang96] T.S. Huang, S. Mehrotra, and K. Ramachandran. Multimedia analysis and retrieval system (MARS) project. Proc. of 33rd Annual Clinic on Library Application of Data Processing-Digital Image Access and Retrieval, 1996.
- [idris95] F. Idris and S. Panchanathan, "Image indexing using vector quantization", In W. Biblack and R. C. Jain, editors, Storage and retrieval for image and video database III, Proceedings of SPIE, Vol. 2420, Bellingham, WA, 1995. SPIE.
- [idris95] F. Idris and S. Panchanathan, "Image indexing using wavelet vector quantization," SPIE Proceedings: Digital Image Storage Archiving Systems, 2606:269-275, Oct. 1995
- [idris96] F. Idris and S. Panchanathan, "Algorithms for indexing of compressed images," Proceedings of International Conference on Visual Information Systems. Melbourne, pp. 303-308, 1996.
- [ikeda00] T. Ikeda and M. Hagiwara. Content-based image retrieval system using neural networks. International Journal of Neural Systems, vol. 10, no. 5 (2000) 417-424. World Scientific Publishing Company
- [ishikawa98] Yoshiharu Ishikawa, Ravishankar Subramanya, and Christos Faloutsos, "MindReader: Querying database through multiple examples," Proceedings of the 24th VLDB Conference, pp. 218-227, New York, USA, 1998.
- [jacobs95] C. E. Jacobs, A. Finkelstein, and D. H. Salesin, "Fast multiresolution image querying," Proc. of ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques, pp. 277-286, Los Angeles, Aug. 1995.
- [jain86] A.K. Jain. Fundamentals of Digital Image Processing. pp342-430. Prentice Hall, 1986.
- [jermyn02] Ian H. Jermyn, Cian W. Shaffrey, and Nick G. Kingsbury, "Evaluation methodologies for image retrieval systems," Proceedings of Advanced Concepts for Intelligent Vision Systems (ACIBS 2002), Ghent, Belgium, September 9-11, 2002.
- [jiang01] Armstrong A, Jiang J. 'An efficient image indexing algorithm in JPEG compressed domain', IEEE International Conference on Consumer Electronics, USA, August 2001.
- [jiang02] Feng G.C. and Jiang J (2002): "JPEG compressed image retrieval via statistics features" Accepted to Pattern Recognition.
- [jiang02] Jiang J. (2002): "Content based image indexing and retrieval in compressed domain", in Computer Graphics International 2002, Edited by J. Vince and R. Earnshaw, Springer-Verlag, London Ltd, 2002.

- [jiang02] Jiang J. and Feng G.C. (2002): "The spatial relationship of DCT coefficients between a block and its sub-blocks" IEEE Transactions on Signal Processing, IEEE, 50 (5): 1160-1169. (http://www.inf.brad.ac.uk/staff/profiles/pubs_profile.php3?usercode=jjiang1) (<http://imaging.comp.glam.ac.uk/publist.htm>)
- [jiang02] Jiang J., Armstrong A.J. and Feng G.C. (2002): "Direct content access and extraction from JPEG compressed images" Pattern Recognition, ELSEVIER.
- [jiang02] Liu M. G, Jiang J. and Hou C. H (2002): "Combination of image indexing and compression", in Accepted to ICASSP'02, IEEE Annual International Conference on Acoustics, Speech and Signal Processing
- [kashyap81] R. L. Kashyap and R. Chellapa. Decision rules for choice of neighbors in random field models of images. Comp. Graph. and Image Proc., 15:301-318, 1981.
- [kato92] T. Kato. Database architecture for content-based image retrieval. Image Storage and Retrieval System. (Jambardino, A A and Niblack, W R, eds), Proc SPIE 1662, 112-123, 1992
- [koc95] Ut-Va Koc and K. J. Ray Liu, "DCT-based motion estimation", Technical Research Report. T.R. 95-1, Electrical Engineering Department and Institute for Systems Research, University of Maryland at College Park, 1995
- [koegel94] John F. Koegel Buford. Multimedia Systems. Addison-Wesley Publication Co. - New York:ACM Press: Reading, Mass., 1994.
- [kohonen95] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, and K. Torkkola, "LVQ pak: the learning vector quantization program package, version 3.1" Laboratory of Computer and Information Science, Helsinki University of Technology, Finland, 1995
- [korfhage97] R.R. Korfhage. Information Storage and Retrieval. pp. 84-85. John Wiley & Sons, Inc. 1997.
- [laaksonon99] Jorma Laaksonon, M. Koskela, and E. Oja, "PicSOM: Self-Organizing Maps for content-based image retrieval," Proc. of IJCNN'99. Washing, DC, July 1999.
- [lambrou98] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domain." International Conference on Acoustics, Speech, and Signal Processing (ICASSP-98), vol. 6, (Seattle, WA), pp. 3621-3624.
- [leung02] Wing Ho Leung and Tsuhan Chen, "Trademark retrieval using content-skeleton stroke classification", submitted to ICME 2002.
- [leung02] Wing Ho Leung and Tsuhan Chen, "User-independent retrieval of free-form hand-drawn sketches", to appear in ICASSP 2002, Orlando, FL, May 2002.

- [li94] Bingcheng Li and Song De Ma. On the relation between region and contour representation. Proc. IEEE Int. Conf. on Patt. Recog., 1994.
- [li97] J. Li, "Hybrid wavelet-fractal image compression based on a rate-distortion criterion," SPIE Proceedings: Visual Communications and Image Processing, 3024:1014-1025, Feb. 1997.
- [lin97] Tsong W Lin, " Fixed attribute-length linear quadtree representations for storing similar images", Proc. SPIE, Multimedia Storage and Archiving Systems II, 3229:278-287. C.-C. J. Kuo; Shih Fu Chang; Venkat N. Gudivada; Eds. Oct. 1997.
- [liuy] Y. Liu, F. Dellaert, and W. Rothfus. Classification driven semantic based medical image retrieval, submitted.
- [lu99] G. Lu and S. Teng, "A novel image retrieval technique based on vector quantization," Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation, pp. 36-41. Feb. 17-19, 1999, Viana, Atria
- [mallat89] S. G. Mallat, "A theory for multiresolution signal representation: the wavelet decomposition," IEEE Trans. on Pattern Analysis and Machine Intelligence, 11(7):674-693, July 1989.
- [mandal96] M. K. Mandal, T. Aboulnasr, and S. Panchanathan, "Image indexing using moments and wavelets," IEEE Trans. on Consumer Electronics, 42(3):557-565, Aug. 1996.
- [mandal97] M. K. Mandal , F. Idris, and S. Panchanathan. Image and video indexing in the compressed domain: a critical review. Proc. of SPIE: Multimedia Storage and Archiving Systems, 3229:2-13, Dallas, Texas, Nov 3-4, 1997. (<http://www.ee.ualberta.ca/~mandal/publish/publication.html>)
- [mandal99] M. K. Mandal , F. Idris, and S. Panchanathan, "A Critical Evaluation of Image and Video Indexing Techniques in the Compressed Domain," Image and Vision Computing Journal-special issue on Content Based Image Indexing, Vol. 17, Issue 7, pp. 513-529, May 1999
- [mandal99] M. K. Mandal, T. Aboulnasr, and S. Panchanathan, "Fast wavelet histogram techniques for image indexing," Computer Vision and Image Understanding (CVIU), 75(1-2):99-110, 1999.
- [manjunath96] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," IEEE Trans. on Pattern Analysis and Machine Intelligence, 18(8):837-841, Aug. 1996.
- [marr83] D. Marr. Vision: A computational investigation into the human representation and processing of visual information. New York: W. H. Freeman, Sept. 1983
- [marr93] D. Marr. Early processing of visual information. Philosophical Transactions of the Royal Society of London B Vol.275, pp 483-524.

- [mehre97] Babu M. Mehre, M. Kankanhalli, and Wing Foon Lee. Shape measures for content based image retrieval: A comparison. *Information Processing & Management*, 33(3):319-337, 1997
- [miano93] John Miano. *Compressed Image File Formats, JPEG, PNG, GIF, XBM, BMP*. Addison-Wesley, 1993.
- [muler01] Henning Müller, Wolfgang Müller, David McG. Squire, Stéphane Marchand-Maillet and Thierry Pun, "Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals," *Pattern Recognition Letters*, special issue on Image/Video Indexing and Retrieval, 22(5):593-601, 2001.
- [natsev99] A. Natsev, R. Rastogi, and K. Shim, "WALRUS: a similarity retrieval algorithm for image database". *SIGMOD '99*, page 395--406.
- [nes97] Niels Nes, Carel van den Berg, Martin Kersten. Database support for image retrieval using spatial-color features. *Image Databases and Multi-Media Search*, pages 293-300. World Scientific, August 1997. also http://www.cwinl/~niels/Pub/image_retrieval/image_retrieval.html. Monet system
- [ng92] I. Ng, T. Tan, and J. Kittler, "On local linear transform and Gabor filter representation of texture," *Proc. of the 11th IAPR Intl. Conf. on Pattern Recognition*, pp.627-631, 1992.
- [niblack93] W. Niblack, R. Berber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, and P. Yanker. The QBIC project: Querying images by content using color, texture and shape. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*. San Jose, pp 172-187, Feb. 1993.
- [niblack97] W. Niblack, X. Zhu, J. Hafner, T. Breuel, D. Pondeleon, D. Petkovic, M. Flickner, E. Upfal, S. Nin, S. Sull, B. Dom, B. L. Yeo, S. Srinivasan, D. Zivkovic, and M. Penner. Updates to the QBIC system. *Proc. SPIE Storage and Retrieval for Image and Video Database VI*, 3312:150-161, Dec, 1997.
- [ortega97] M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra, and T.S. Huang. Supporting similarity queries in MARS. *Proc. of ACM Conf. On Multimedia*, 1997
- [pan00] Zhengjun Pan, Alistair G. Rust, and Hamid Bolouri, "Image redundancy reduction for neural network classification using discrete cosine transforms", *Proc. of the IEEE-INNS-ENNS International Joint Conf. on Neural Networks (IJCNN2000)*, Vol. III, 149-154, Como, Italy, 2000.
- [pass99] Greg Pass and Ramin Zabih, "Comparing images using joint histograms." *ACM Journal of Multimedia Systems*, 7(3):234-240, May 1999.
- [pentland84] A. Pentland. Fractal-based description of natural scenes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(6):661-674, November 1984.

- [pentland94] A. Pentland, R. W. Picard, and S. Sclaroff, "PhotoBook: tools for content-based manipulation of image databases," Proc. of SPIE: Storage and Retrieval for Image and Video Databases II, 2185:34-47, Feb. 1994.
- [pentland96] A. Pentland, R.W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. Int. J. Comput. Vis., 18(3):233-254, 1996.
- [persoon77] Persoon, E., and Fu, K.-S., "Shape discrimination using Fourier Descriptors," IEEE Trans. Systems, Man, and Cybernetics, vol. SMC-7, no. 3, pp. 170-179, 1977.
- [petrakis02] Euripides G.M. Petrakis, Christos Faloutsos, and King-Ip Lin: "ImageMap: An Image Indexing Method Based on Spatial Similarity", IEEE Trans. on Knowledge and Data Engineering (IEEE-TKDE), 14(5):979-987, Sept./Oct. 2002.
- [petrakis]Euripides G.M. Petrakis, "Fast retrieval by spatial structure in image database," Journal of Visual Languages and Computing (accepted).
- [podilchuk96] C. Podilchuk and X. Zhang, "Face recognition using DCT-based feature vectors," Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, 4:2144-2147, 1996.
- [prasad97] L. Prasad. Morphological analysis of shapes. CNLS Research Highlights, Los Alamos National Laboratory. July 1997
- [press92] Press, W. H., S. A. Teukolsky, et al. (1992). Numerical Recipes in C, Cambridge University Press.
- [rao93] A.R. Rao and G. L. Lohse. Towards a texture naming system: identifying relevant dimensions of texture. IEEE Conference on Visualization. pp. 220-227. Oct. 1993.
- [rashkovskiy92] P. Rashkovskiy and S. Mallat, "Second generation compact image coding with wavelets," in Wavelets: A tutorial in Theory and Applications, Ed: C. K. Chui, Academic Press, Inc., 1992.
- [rasmussen97] E. M. Rasmussen. 1997. Indexing images. Annual review of information science and technology. vol. 32, pp 169-196
- [ravishankar90] A. Ravishankar Rao. Taxonomy for texture description and identification. Springer-Verlag, New York, 1990.
- [ray98] K.J. Ray Liu and Ut-Va Koc, "DCT-Based Motion Estimation," US Patent 5,790,686, Aug. 4, 1998. http://www.isr.umd.edu/ISR/accomplishments/012_DCT/
- [reeves77] R. Reeves, K. Kubik and W. Osberger, "Texture characterization of compressed aerial images using DCT coefficients," Proc. of SPIE: Storage and Retrieval for Image and Video Databases V, 3022:398-407, Feb. 1977.

- [rocchio71] JJ Rocchio, "Relevance feedback in information retrieval," The SMART Retrieval System, pp. 313-323, Prentice Hall, 1971.
- [rowley98] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38. January 1998
- [rui00] Y. Rui, T. S. Huang, "Optimizing learning in image retrieval," Proc. IEEE Int. Conference On Computer Vision and Pattern Recognition (CVPR), pp. 236-243, Hilton Head, South Carolina, June 2000.
- [rui98] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool in interactive content-based image retrieval," IEEE Trans. on Circuits and Systems for Video Technology, 8(5):644-655, Sept. 1998.
- [rui99] Y. Rui, T.S.Huang, and S.F. Chang. Image Retrieval: Current techniques, promising directions and open issues, Journal of Visual Communication and Image Representation, vol. 10, pp. 39-62, March 1999
- [rui99] Y. Rui, T.S.Huang, and S.F. Chang. Image Retrieval: Current techniques, promising directions and open issues, Journal of Visual Communication and Image Representation, vol. 10, pp. 39-62, March 1999.
- [salton83] G. Salton and M.J. McGill: Introduction to Modern Information Retrieval, McGraw-Hill Book Company, New York (1983)
- [sellis87] T. Sellis, N. Roussopoulos, and C. Faloutsos, "The R+-tree: A dynamic index for multi-dimensional objects," Proc. 12th VLDB, pp. 507-518, 1987.
- [shapiro01] Linda G. Shapiro and George C. Stockman, "Computer vision," 1st edition, Chapter 8. Prentice Hall, January 23, 2001.
- [shen94] F. Qi, D. Shen, and L. Quan, "Wavelet transform based rotation invariant feature extraction in object recognition," Proc. of Intl. Symp. on Information Theory & its Applications, pp. 221-224, Nov. 1994.
- [shen96] B. Shen and I.K.Sethi, "Direct feature extraction from compressed images," Proc. of SPIE, 2670:404-414, 1996.
- [shneier96] M. Shneier and M.A. Mottaleb, "Exploiting the JPEG compression scheme for image retrieval," IEEE Trans. on Pattern Analysis and Machine Intelligence, 18(8):849-853, Aug. 1996.
- [smeulders00] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, Ramesh Jai, "Content-based image retrieval at the end of the early years," IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(12):1349-1380. December 2000.

- [smith94] J.R. Smith and S.F. Chang, "Transform features for texture classification and discrimination in large image databases," Proc. of IEEE Intl. Conf. on Image Processing, 3:407-411, 1994.
- [smith94] John R. Smith and Shih Fu Chang, "Quad-tree segmentation for texture-based image query." Proceedings ACM Multimedia '94. pp. 279-286, ACM, October 1994.
- [smith95] J.R. Smith and S.F. Chang. Automated image retrieval using color and texture. Columbia University Technical Report, TR# 414-95-20, July 1995
- [squire00] David McG. Squire, Wolfgang Müller, Henning Müller, Thierry Pun, "Content-based query of image databases: inspirations from text retrieval," Pattern Recognition Letters 21:1193-1198, 2000.
- [squire01] David McG. Squire, Henning Müller, Wolfgang Müller, Stéphane Marchand-Maillet and Thierry Pun, "Design and Evaluation of a Content-based Image Retrieval System, chapter 7. Idea Group Publishing, 2001.
- [squire95] D.M. Squire, and T.M. Caelli. Shift, rotation and scale invariant signatures for two-dimensional contours, in a neural network architecture. 1st International Conference, Mathematics of Neural Networks and Applications, July 1995, Lady Margaret Hall, Oxford
- [squire99] David McG. Squire, Henning Muller, Wolfgang Muller, "Improving response time by search pruning in a content-based image retrieval system, using inverted file techniques," IEEE Workshop on Content-based Access of Image and Video Libraries, pp.45-49, June 22-22, 1999, Fort Collins, Colorado.
- [srihari95] R.K. Srihari. Automatic indexing and content-based retrieval of captioned images. IEEE Computer Magazine. 28(9):49-56, 1995
- [srikanth99] M. Srikanth, Image indexing and retrieval using the cross-entropy measures. Proceedings of the HKK Conference, Waterloo, Ontario, Canada, June 1999
- [stillings95] Neil A. Stillings et al. Cognitive Science, An Introduction. 2nd Edition. The MIT Press, Cambridge, Massachusetts, 1995.
- [su00] Zhong Su, Hongjiang Zhang, and Shaoping Ma, "Using Bayesian classifier in relevance feedback of image retrieval," IEEE International Conference on Tools with Artificial Intelligence, Vancouver, Canada, November 2000.
- [su01] Zhong Su, Hongjiang Zhang, and Shaoping Ma, "Relevance feedback using a bayesian classifier in content-based image retrieval," SPIE Storage and Retrieval for Media Database 2001, San Jose, January 2001.
- [suematsu02] Nobuo Suematsu, Yoshihiro Ishida, Akira Hayashi, Toshihiko Kanbara, "Region-Based Image Retrieval using Wavelet Transform," The 15th International Conference on Vision Interface, pp9-16, May 27-29, 2002, Calgary, Canada.

- [swain91] Michael Swain. and Dana Ballard. Color indexing. *Int. J. Comput. Vis.* 7(1):11-32, 1991.
- [swanson96] M. D. Swanson, S. Hosur and A. H. Tewfik, "Image coding for content-based retrieval," *Proc. of SPIE, VCIP*, 2727:4-15, 1996
- [tamura78] H. Tamura, S. mori, and T. Yamawaki. Texture features corresponding to visual perception. *IEEE Trans. On Sys. Man. And Cyb.* SMC:8(6):780-786, 1978.
- [tang98] Tang, Xiaou, "Texture information in run-length matrices," *IEEE Transaction on Image Processing*, Vol. 7, p1602-1609, No. 11, November 1998.
- [tieu00] Kinh Tieu and Paul Viola, "Boosting image retrieval with relevance feedback," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)*, Hilton Head, South Carolina.
- [tseng95] D. C. Tseng, Y.F. Li and C.T. Tung, Circular histogram thresholding for color image segmentation. *Document Analysis and Recognition. Proceedings of the Third International Conference*, Vol. 2, pp. 673-676, Aug. 1995.
- [uler01] Henning Müller, Wolfgang Müller, Stéphane Marchand-Maillet, David McG. Squire and Thierry Pun, "An automatic benchmark for content-based image retrieval, *International Conference on Multimedia and Exposition*," ICME 2001, Tokyo, Japan, 2001
- [unser95] M. Unser, Texture classification and segmentation using wavelet frames. *IEEE Trans. Image Proc.*, 4:1549-1560, 1995
- [vasconcelos00] Nuno Vasconcelos and Andrew Lippman, "Bayesian Relevance Feedback for Content-Based Image Retrieval", *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'00)*, p. 63, Hilton Head, South Carolina, June 16 - 16, 2000
- [vasconcelos98] N. Vasconcelos and A. Lippman, "A Bayesian framework for content-based indexing and retrieval," *Proc. of DCC '98, Snowbird, Utah*, 1998.
- [vellaikal95] A. Vellaikal, C. C. J. Kuo, and S. Dao, "Content-based retrieval of remot-sensed images using vector quantization," *Proc. of SPIE*, 2388:178-189, 1995.
- [vellaikal95] A. Vellaikal, C.-C. Kuo and S. Dao, "Content-Based Retrieval of Color and Multispectral Images Using Joint Spatial-Spectral Indexing," *Proc. SPIE Digital Image Storage and Archiving Systems*, 2606:232-243, 1995
- [verevka95] O. Verevka and J.W. Buchanan, "Local k-means algorithm for color image quantization," *In Proceedings of Graphics Interface*, pages 128-135, 1995.
- [voorhees86] Ellen M. Voorhees, "The efficiency of inverted index and cluster searches," *Proceedings of 1986 ACM Conference on Research and Development in Information Retrieval*, pp. 164-174, ACM Press, New York, NY, USA, 1986

- [wallace91] Gregory K. Wallace. The JPEG still picture compression standard. *Communic. ACM*, 34(4):31-45, Apr. 1991.
- [wan96] X. Wan and C.-C.J.Kuo, Color distribution analysis and quantization for image retrieval. *Proc. SPIE Storage Retrieval Still Image Video Database IV 2670*, pp.8-16, February, 1996.
- [wang96] H. Wang and S. F. Chang, "Adaptive image matching in the subband domain," *Proc. of SPIE: VCIP*, 2727:885-896, 1996.
- [wang97] J. Z. Wang, G. Wiederhold, O. Firschein, and S. X. Wei, "Wavelet-based image indexing techniques with partial sketch retrieval capability," *Proc. of the Forum on Research and Techniology Advances in Digital Libraries*, pp.13-24, Washington DC, May 1997.
- [wang98] J. Z. Wang, G. Wiederhold, O. Firschein, and X. W. Sha, "Content-based image indexing and searching using Daubechies' wavelets," *Int'l J. Digital Libraries*, 1(4):311-328, 1998.
- [webster93] Webster's 3rd International Dictionary, Merriam Webster Inc. Publishers, Springfield, Massachusetts, USA. 1993.
- [weeds02] Julie Weeds, "The Reliability of a Similarity Measure." *Proceedings of the Fifth UK Special Interest Group for Computational Linguistics (CLUK5)*. Leeds. January 2002.
- [white96] D. White and R. Jain, "Similarity indexing: algorithms and performance," *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 2670:62-73, 1996.
- [wouwer98] Gert Van de Wouwer. *Wavelets for multiscale texture analysis*. PhD thesis. Departement Natuurkunde, Universiteit Antwerpen, 1998.
- [yang98] Z. Yang, X. Wan, and C.-C. J. Kuo, "Interactive Image Retrieval: concept, prodedure and tools," *IEEE 32nd Asilomar Conference, Monterey, CA*, pp. 313-317, Nov. 1998.
- [yates99] Richardo Baeza-Yates & Berthier Ribeiro-Neto. *Modern Information Retrieval. Glossary*. Addison Wesley Longman Publishing Co. Inc. May, 1999. (web: <http://www.sims.berkeley.edu/~hearst/irbook/glossary.html>)
- [yoon01] J. Yoon and N. Jayant, "Relevance feedback for semantics based information retrieval," *IEEE Int'l Conf. on Image Proc. (ICIP'2001)*, pp. 42-45, Thessaloniki, Greece, Oct. 7-10, 2001.
- [zahn72] Zahn, C. T., and Roskies, R. Z., "Fourier descriptors for plane closed curves," *IEEE Trans. on Computers*, vol. C-21, no. 3, pp. 269-281, 1972.
- [zheng] Lei Zheng, Arthur Wetzel, John Gilbertson and Michael Becich, "Design and analysis of a content-based pathology image retrieval system," *IEEE Transaction on Information Technology in Biomedical Sciences (TITBS)*, accepted.

- [zheng_thesis] Lei Zheng, "Design and analysis of a content-based pathology image retrieval system." Master's thesis, Department of Pathology, University of Pittsburgh, School of Medicine, 2002.
- [zhu00] Lei Zhu, Aibing Rao, Aidong Zhang, "Advanced feature extraction for keyblock-based image retrieval," Proceedings of International workshop on multimedia information retrieval (MIR2000), pp.179-183, Los Angeles, California, USA, November 4, 2000
- [zhu02] Lei Zhu, Aibing Rao, and Aidong Zhang, "Theory of Keyblock-based image retrieval," ACM Transactions on Information Systems (TOIS), 20(2):224-257, April 2002.