APPLICATION OF LAPLACIAN MIXTURE MODEL TO IMAGE AND VIDEO RETRIEVAL

Tahir Amin
B.Sc. Electrical Engineering
University of Engineering and Technology Lahore, Pakistan

Faculty of Engineering and Applied Sciences
Department of Electrical and Computer Engineering

Submitted in partial fulfillment
of the requirements for the degree of
Master of Applied Science

School of Graduate Studies
Ryerson University
Toronto, Ontario

January 2004

Author's Declaration

I hereby declare that I am the sole author of this thesis.
I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Borrower's Page

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Name	Signature	Address	Date

ABSTRACT

APPLICATION OF LAPLACIAN MIXTURE MODEL TO IMAGE AND VIDEO RETRIEVAL

©Tahir Amin 2003

Master of Applied Science
Department of Electrical and Computer Engineering
Ryerson University

In this study, we present a new approach to feature extraction for image and video retrieval. A Laplacian mixture model is proposed to model the peaky distributions of the wavelet coefficients. The proposed method extracts a low dimensional feature vector which is very important for the retrieval efficiency of the system in terms of response time. Although, the importance of effective feature set cannot be overemphasized, yet it is very hard to describe image similarity with only low level features. Learning from the user feedback may enhance the system performance significantly. This approach known as the relevance feedback is adopted to further improve the efficiency of the system. The system learns from the user input in the form of positive and negative examples. The parameters of the system are modified by the user behavior.

The parameters of the Laplacian mixture model are used to represent texture information of the images. The experimental evaluation indicates the high discriminatory power of the proposed features. The traditional measures of distance between two vectors like city-block or Euclidean are linear in nature. The human visual system does not follow this simple linear model. Therefore, a non-linear approach to the distance measure for defining the similarity between the two images is also explored in this work. It is observed that non-linear modelling of similarity yields more satisfactory performance and increases the retrieval performance by 7.5 per cent.

Video is primarily multi-modal, i.e., it contains different media components like audio, speech, visual information (frames) and caption (text). Traditionally, visual information is used for the video indexing and retrieval. The visual contents in the videos are very important, however, in some cases visual information is not very helpful for finding clues to the events. For example, certain action sequences such as goal events in a soccer game and explosion in a news video are easier to identify in the audio domain than in the visual domain. Since the proposed feature extraction scheme is based on the shape of the wavelet coefficient distribution, therefore, it can also be applied to analyze the embedded audio contents of the video. We use audio information for indexing video clips. A feedback mechanism is also studied to improve the retrieval performance of the system.

Acknowledgement

I would like to thank my supervisor Dr. Ling Guan and co-supervisor Dr. Mehmet Zeytinoğlu for their encouragement, guidance, and continuous support throughout my research work and writing of this manuscript. This work would have been impossible without their feedback, patience and kindness. My thanks are also due to Dr. Dimitrios Hatzinakos for his valuable input to my research as well as to the production of this document.

I would also like to thank Canada Foundation for Innovation (CFI) and the Department of Electrical and Computer Engineering for providing a very well equipped and technically supported Ryerson Multimedia Laboratory. My thanks are due to the School of Graduate Studies of Ryerson University for providing Graduate Student Scholarship.

I would like to acknowledge my supervisors' funding resources National Sciences and Engineering Research Council of Canada (NSERC) and Canada Research Chair Program, for financial support to this research work.

My thanks are due to Hua Yuan for having useful and informative discussion on Gaussian mixture model. Finally, I would like to thank my colleagues and members of the Ryerson Multimedia Laboratory for creating a friendly and congenial environment in the Lab. It was my pleasure to work with such a great team.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Information retrieval has become an active area of research due to the emergence of information superhighway. The amount of information available in the digital format is increasing at an exponential rate. This huge amount of information is accessible through public (such as Internet) and private networks as well as by stand alone systems. The cost of storing the digital information has been reduced significantly due to the development of new technologies such as Compact Discs (CDs) and Digital Video Disks (DVDs). The generation of data in different formats has also become very easy in recent years. Therefore, the information is now available in different media such as video and images. In order to utilize this tera bytes of information in a meaningful way, effective and efficient search methodologies are required.

The primary objective of information retrieval is to provide access to the recorded knowledge. In this chapter, we will discuss fundamentals of the information retrieval (IR) systems. The traditional IR paradigm developed primarily for the text documents has limitations when applied to the other data formats such as images, audio, graphics or video. The transformation of the IR systems from simple text-based annotation to content-based analysis is also described.

1.1 Retrieval of Text Documents

1.1.1 Cataloguing

The purpose of cataloguing is to facilitate the search of a particular document from the library. A book is described by a few standard textual terms such as book

title, author, subject etc. The users are able to search through the catalogue to find out a particular book and to know what books are contained in the library. The users can also find out all the documents/books on a particular subject. After the invention of computers, the traditional catalogues have been replaced by a more sophisticated electronic catalogue system. It is now possible to attach a variety of descriptors characterizing a document. The search is performed by computers and the users have more options to choose from. For example, the user can find out the documents published in a particular year. Keyword search is also provided which adds further flexibility to the retrieval process. Another advantage of the modern electronic catalogues is that if the library does not have a particular item the user is looking for, it may suggest some documents that possess similar characteristics. The cataloguing process is however manual, i.e., carried out by a person. It is not only very time consuming but also subjective to the personal bias.

1.1.2 Modern IR of Text Documents

The manual indexing of the electronic documents is a very tedious operation. The speed at which the electronic documents are being created and stored in digital format makes it almost impossible to index them through manual operation. An automatic algorithm is necessary to produce indexes and search engines should be developed to retrieve the relevant information. The modern IR does not rely on the controlled vocabulary like traditional cataloguing. The indexes are derived from the contents of the documents automatically. The pioneer work in this area was done by Luhn [1]. The documents are encoded by submitting to a mechanized process to extract the significant words from the document. These words are then compared with the query words. The relevancy of the documents is decided on the basis of the occurrence of these words. Luhn developed a statistical approach in which the *significance* factor of a sentence is derived from an analysis of its words. The frequency of word occurrence in a document is a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnishes a useful measurement for determining the significance of sentences. The

significance factor of a sentence will therefore be based on a combination of these two measurements. Luhn contended that the words are significant if their frequency of occurrence is within a range. If the frequency of occurrence increases beyond a higher cut off limit, the word becomes insignificant such as in case of commonly used words 'is', 'are', 'am' etc. On the other hand, if the frequency of occurrence is below the lower cut off limit then the word is also not significant. In this case the document does not contain sufficient instances of the word to make it significant. This principle is depicted in Figure 1.1.

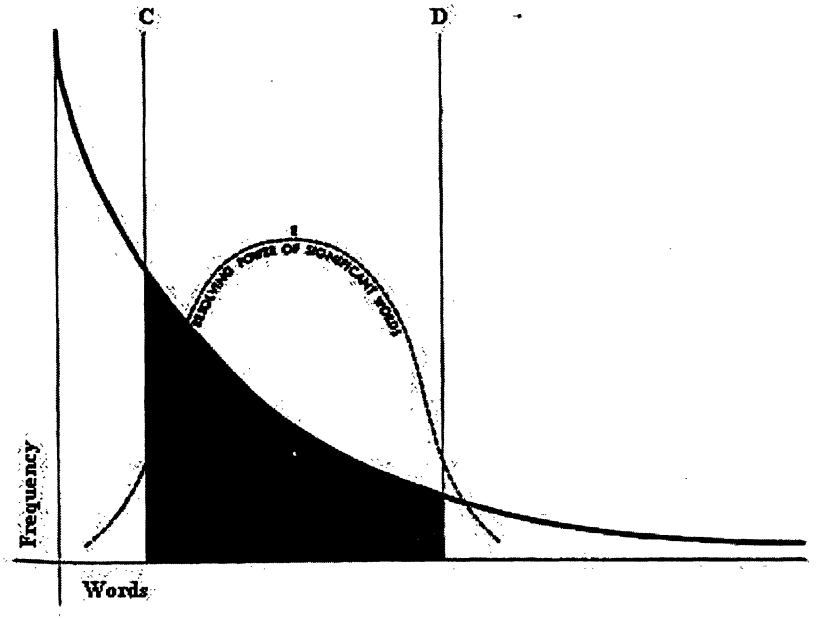

Figure 1.1: Luhns word rank-frequency diagram [1].

1.2 Text-Based Retrieval of Audiovisual Documents

The text-based retrieval of the audiovisual documents is an extension of the modern IR. The attributes of the audiovisual documents known as the meta-data are also stored along with the data. Meta-data is the data about the data that describes characteristics of the audiovisual documents such as file name, file type, file size, author's name etc. When the users submit textual queries, the system searches

through the meta-data to find the relevant set of documents. The meta-data is added to the documents by a manual or semi-automatic process. Text-based retrieval of audiovisual data has many limitations from which some are enumerated below:

- The manual indexing of audiovisual information is highly subjective and reflects the personal viewpoint of the author.
- It is very difficult to find suitable textual annotations for the description of audiovisual documents.
- The users have to provide textual description of the query which is very difficult to establish. The users cannot search the database by providing an example image or video sequence.

In spite of the above noted limitations, text-based search of the audiovisual documents is a popular way of finding information on the World Wide Web (www). The popular web search engines such as Google [2], Lycos [3] and Altavista [4] have extended their capabilities to include key-word based search of the audiovisual data. Google search engine has image search that allows users to retrieve images by text queries. Figure 1.2 shows the Advanced Image Search Interface of Google. Lycos search engine has provided multimedia search including video, music and images. Similarly, Altavista meta search engine has options of image, MP3/Audio and video search.

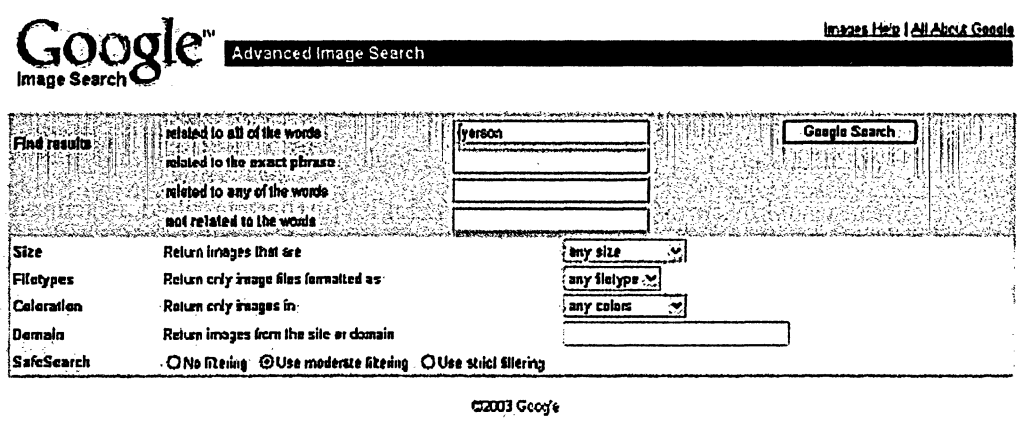

Figure 1.2: Advanced Image Search Interface of Google [2].

1.3 Content-Based Retrieval

The inefficiencies and limitations of the text-based retrieval of the audiovisual documents prompted the researchers to look into alternative ways of searching these documents. Indexing and retrieving based on the contents was the natural outcome of this investigation. The idea of content-based searching has already been matured for the text documents. The key-word search based on the word frequency paradigm is content-based in essence. The visual and aural contents of the audiovisual data such as texture, color, shape, pitch and loudness. are used as clues for retrieving similar documents from the database. A low level representation of these characteristics is to be extracted from the contents of the audiovisual data. This extracted information is then used to build content-based indices for the documents. The development in the content-based retrieval systems is due to the contributions from many areas such as computer vision, human computer interaction, psychology, pattern recognition, signal processing and multimedia database systems.

1.3.1 Content-Based Image Retrieval

Problems with text-based access to images have prompted increasing interest in the development of content-based solutions. This is most often referred to as content-based image retrieval (CBIR). Content-based image retrieval relies on the characterization of primitive features such as color, shape, and texture that can be automatically extracted from the images. Feature extraction is the basis of the content-based image retrieval. In broad sense, features may include both textual (keywords, annotations etc.) and visual descriptors (color, texture, shape and faces etc.). Within the visual feature scope, the features can be further classified as general features and domain-specific features. The former includes color, texture and shape and the latter is application dependent and may include, for example, human faces and finger prints. Because of the human perceptual subjectivity, there is no single best representation for a given feature. Multiple representations characterize the same visual feature from different perspectives.

Users seeking images come from a variety of domains, including law enforcement, journalism, education, entertainment, medicine, architecture, engineering, publishing, advertising, and art. They may look at the images with different perspectives. Learning high level semantic concepts is a challenging task for the content-based retrieval systems. Finding the objects of interest and high level segmentation of the images similar to that performed by the human visual system still needs a lot of research.

1.3.2 Content-Based Video Retrieval

By the word video, we refer to the image sequence and its accompanying audio track. The embedded audio content in videos consists of speech as well as non-speech segments. As the information revolution continues into the new millennium, the generation and dissemination of digital multimedia content continues to witness phenomenal growth. However, this rate of growth has not been matched by the simultaneous emergence of technologies that can process the multimedia content efficiently. State of the art systems for content management lag far behind the expectations of the users of these systems. The users expect these systems to perform analysis at the same level of complexity and semantics that a human would employ while analyzing the content. Herein lies the reason why no commercially successful systems for content management exist. Humans assimilate information at a semantic level and do so with remarkable ease. We do not recall the movie we watch in its entirety but through a small number of scenes that leave an impression on our mind and through the story-line that we grasp. In fact the human ability to apply knowledge to the task of sifting through large volumes of multimodal data and extracting only relevant information is amazing. The troika of sensory organs, short term and long term memory and ability to learn and reason based on sensory inputs (through supervised or unsupervised training) are the mainstay of the human ability to perform semantic analysis on multimodal data.

The task of automatic analysis is to reduce the tremendous volume of multimodal data to concise representations, which capture the essence of the data. Tools for efficient storage, retrieval, transmission, editing, and analysis of multimedia content are

absolutely essential for the utilization of raw multimedia content. Examples include television and satellite broadcast of news, sports, weather, politics, entertainment, personalized entertainment like video on demand, search and retrieval of content on the World Wide Web, electronic commerce; human-computer intelligent interaction, communication through wired and wireless devices; advanced collaboration like video-conference, chat rooms, recreational activities like planning travel, content sharing and personalization of content indexing. Video databases serve as a perfect example of how the acute need for tools has severely constrained the use of multimedia content. Research in speech recognition is now over 3 decades old. Filtering of multimedia content can enable automatic rating of Internet sites and restrict access to violent content. Semantic understanding could mean better and natural interfaces in human computer interaction.

1.4 MPEG-7 - Multimedia Content Description Interface

MPEG-7, formally known as Multimedia Content Description Interface, is the current ISO (International Organization for Standardization) standard developed by MPEG (Motion Pictures Expert Group). While the prior MPEG standards focus on coding and representation of audio-visual content, MPEG-7 focuses on description of multimedia content. It addresses content with various modalities including image, video, audio, speech, graphics and their combinations. MPEG-7 complements the existing MPEG standards and aims to be applicable to many existing formats. The development of MPEG-7 standard is driven by critical needs for indexing, searching, filtering and managing audio-visual data. It is important not only for the end users, but also for the providers of audiovisual content or services. In order to achieve the maximum inter-operability and facilitate the creation of innovative applications, MPEG-7 intends to be an inter-operable interface. It defines the syntax and semantics of various description tools. Each tool may be designed for specific or generic modalities (e.g., audio, visual or multimedia), aspects (e.g., media, meta, structural or semantic)

and applications (e.g., search engine, filtering agent or navigation) [5]. The purpose of MPEG-7 is to provide inter-operability among systems and applications used in generation, management, distribution, and consumption of audiovisual content descriptions. Such descriptions of streamed or stored media help users or applications to identify, retrieve, or filter audio-visual information. Example applications include broadcast media selection, radio, TV channels, digital libraries, image catalog, musical dictionary, multimedia directory services and multimedia editing. The use of MPEG-7 descriptions is expected to result in a flexible and scalable framework for designing services that can be accessed from a variety of terminals such as mobile devices, set top boxes, and personal computers. This is shown in Figure 1.3.

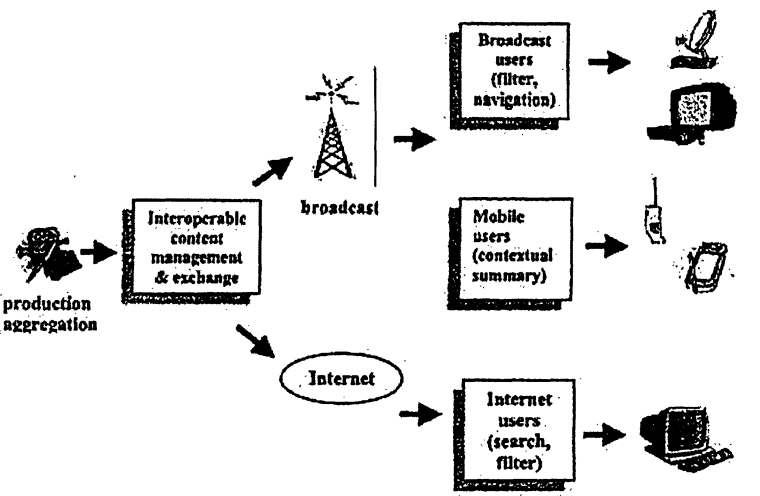

Figure 1.3: Role of MPEG-7 in Facilitating Inter-operable Services and Applications [5].

MPEG-7 provides several normative elements such as *Descriptor* (D), *Description Scheme* (DS) and *Description Definition Language* (DDL) to handle a wide range of applications. *Descriptors* define syntax and semantics of features of audiovisual content. Different levels of abstraction are addressed by MPEG-7. At the low abstraction level, descriptors may include shape, motion, texture, color, and camera motion for images/videos, energy, harmonicity and timbre for audio. At higher abstraction level, *descriptors* may include events, abstract concepts and content genres. Audio

and visual descriptors represent specific features related to audio and visual content respectively. Generic *descriptors* address generic features. *Description schemes* allow construction of complex descriptions by specifying the structure and semantics of the relationships among the constituent *descriptors* or *description schemes.* For example, the description scheme for a video segment may specify the syntax and semantics of the component elements such as underlying segment decomposition, individual segment attributes (such as segment length and textual annotations), and relationships between component segments. As in the case of descriptors, description schemes can be categorized to audio, visual, or generic. Generic description schemes usually represent generic meta information related to all kinds of media such as audio, visual, text and graphic. MPEG-7 also includes descriptors and description schemes related to creation, production, management and access of audio-visual content. Such meta-data may include information about the coding scheme used for compression (e.g., JPEG, MPEG-2), the overall data size, conditions for accessing the material (e.g., intellectual property rights information and financial information), classification (include parental rating, and content classification into a number of pre-defined categories), and links to other relevant material (this information may help the user speeding up the search).

The MPEG-7 description definition language allows flexible definition of MPEG-7 description schemes and descriptors based on XML Schema. The current *descriptors* and *description schemes* are application independent. When it is required to describe content for specific domains (e.g., news, films), there is often a need to extend and specialize the generic MPEG-7 tools and use DDL to define specialized or additional tools. MPEG-7 allows descriptions of audiovisual content at different perceptual and semantic levels. It is expected that low-level features (such as color and structural features) can be extracted in fully automatic ways, whereas high-level features that describe semantic information might need more human interaction.

Figure 1.4 depicts the MPEG-7 processing chain to explain the scope of the MPEG-7 standard. It is important to understand that for descriptors and description schemes, the MPEG-7 standard does not specify how to extract these descriptions.

However, as a normative requirement, the representation of these descriptions must conform to the MPEG-7 standard. Compliant MPEG-7 binary or non-binary descriptions can be accessed, understood, and consumed by applications that are able to decode and process MPEG-7 descriptions. How the MPEG-7 descriptions ought to be used for further processing—i.e., for search and filtering of content is not standardized in MPEG-7 to leave maximum flexibility to applications. It should also be noted that MPEG-7 descriptions may be physically located with the associated audio-visual material in the same data stream or on the same storage system. Alternatively, the descriptions could also be located anywhere else, as long as it is possible to link audio-visual material and their MPEG-7 descriptions efficiently.

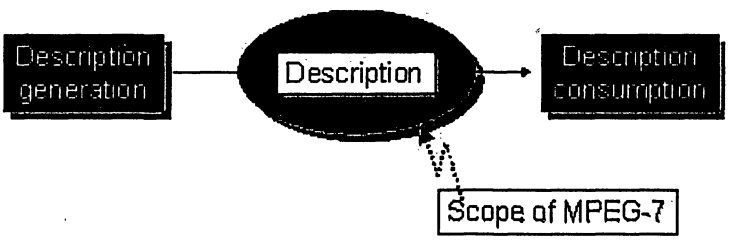

Figure 1.4: Scope of MPEG-7 [5].

1.5 Contribution of the Thesis

1.5.1 Modelling of Wavelet Coefficient Distributions

Wavelet coefficients have very peaky distributions. Taking into account this feature and the retrieval application in mind, we develop a statistical approach for modelling the shape of the wavelet coefficient distributions. The finite mixture modelling is an acknowledged approach in the statistical pattern recognition. We investigate Laplacian mixture model to model the shape of the wavelet coefficient distributions. We demonstrate that a mixture of only two components can approximate the wavelet coefficient distribution efficiently. The parameters of the model are used to index the image and embedded audio in the video retrieval. The proposed feature extraction approach is tested by experimental evaluation. The results indicate that the proposed features provide a good representation of the texture content of the images.

Since the proposed features are based on the characteristics of the wavelet coefficient distribution, the technique is also very successful for audio-based video retrieval.

1.5.2 Application to Image Retrieval

A statistical content analysis approach is developed for application to image indexing and retrieval. The proposed features are very effective for the description of texture information contained in the images. We have also shown the power of human centered interaction in improving the efficiency of the retrieval engine. The low dimensionality of the proposed feature vector reduces the complexity of the retrieval process.

1.5.3 Application to Video Retrieval

An audio based indexing scheme for video data is proposed in this work. The work in this area has been scarce in the past. Most of the research in the video retrieval is directed to the utilization of visual information. We have used audio to describe the video content because the users are often interested in the auditory similarity such as in case of songs and sports videos. The system can be utilized for the audio retrieval or for building audio-based indexes for the video data. The main advantage of the proposed technique is its efficiency for video retrieval at the scene and clip levels. A graphical user interface is developed through which users can browse the database, play a video clip, search the database by providing a query clip and provide feedback.

1.6 Organization of the Thesis

The rest of the thesis is organized as follows. **Chapter 2** presents a survey of several techniques used for image and video retrieval. Image indexing and retrieval techniques based on color, texture and shape characteristics are described. Video indexing and retrieval schemes based on key-frames, audio, motion and caption text are also mentioned. The concept-based retrieval of audio-visual documents is a new development in multimedia content description. This search paradigm is also discussed in this chapter.

The first part of **Chapter 3** provides the background information for the two main ideas used throughout this work. Wavelet transform was introduced during 1980s. Since then, it's use have become very extensive in image and audio analysis. This chapter provides the necessary mathematical framework for the one-dimensional and two-dimensional Wavelet transform from an electrical engineering perspective. In the second part of the chapter, a mixture of Laplacian model for modelling the shape of the wavelet coefficient distributions is developed. The procedure for calculating the model parameters using the Expectation Maximization (EM) algorithm is explained explicitly. Experimental evaluation is presented to prove the validity of the model for our particular application.

In **Chapter 4**, we apply the statistical approach developed in chapter 3 for extraction of texture features from the images. The radial basis function is adopted for similarity measurement which is a non-linear similarity criteria. The feedback mechanism for the improvement of retrieval performance and to know users' notion of similarity is discussed in this chapter. The experimental evaluation of the algorithm for image retrieval is also presented. A comparison study between some existing schemes and the proposed approach is also described.

We explore the utility of the embedded audio content for finding certain types of video shots in **Chapter 5**. Video is a multi-modal data type. The embedded audio contents of the video is a rich source of information that should be tapped for video content description. A comprehensive experimental evaluation of the audio-based approach is presented in this chapter.

We conclude this thesis with **Chapter 6**. The possible future research extensions are detailed. The fusion of the multi-modality for video indexing is a challenging task for the possible future research work. This chapter also discusses the extension of the image retrieval system to the color image databases.

Chapter 2

Literature Review

2.1 Content-Based Image Retrieval

The traditional information retrieval systems relied on the manual indexing of the documents. When this indexing paradigm was extended to image retrieval, it was noted that it is very hard to describe the contents of images by textual descriptors. Moreover, the human subjectivity is also an important issue. The early image retrieval techniques were focused on transforming the manual indexing with a fully automatic process without any human interference. The image retrieval systems are based on the low level features representing color, shape and texture. The indexing of the images based on the objects of interest, however, requires very efficient image segmentation algorithms. Researchers have taken different approaches and have proposed several techniques for extraction of features from the images.

2.1.1 Color Based Image Retrieval

Color is a very important visual feature that is immediately perceived by the viewer. Image retrieval using color information endeavors to model the human color perception. Color feature is very robust and is independent of the size of the image. There are many different techniques proposed in the literature for image indexing based on the color information. The arrangement of colors is connected to psychological effects. The psychological and artistic studies show that the color sensations in an observer follow certain rules [6]. The color is usually represented by points in 3-dimensional color space. There are various color models based on different studies. The extraction

and performance of color features is related to the color space used.

The color histograms are the most traditional way for representation of color by low-level features. The idea of color histogram was first proposed by M. J. Swain and D.H. Ballard [7]. The color properties of the images can be defined by 3 independent color distribution for the three primary colors (such as in RGB color space). It may also be represented by one distribution over three primaries. The images are usually subsampled before the extraction of color features. The distribution of the colors is obtained by discretizing the image colors. Swain and Ballard also introduced the notion of *Histogram Intersection* to compare the histograms of two images to define similarity. The advantage of this similarity measure is that the colors that are not present in the query image do not contribute towards the similarity measurement. Another advantage is that its performance is not affected by the number of bins in the histogram. They also used the histogram intersection technique to locate objects. Histograms are invariant to translations and rotation about the viewing axis, and change only slowly under change of angle of view, change in scale and occlusion. Also histograms, by themselves, do not include spatial information. So the images with very different layouts may have the same histogram.

The quantization of the color space is essential to reduce the number of colors and increase the computational efficiency. Due to this quantization, histograms may have problems in representing the color content of the images. The number of quantization levels should be selected in such a way so that the different colors does not fall into the same bin. A uniform quantization of the color space may be appropriate for the perceptually uniform color spaces. But in case of perceptually non-uniform color spaces such as HSV or RGB, a non-uniform quantization method should be applied. The choice of the quantization levels thus affects the efficiency of a particular color feature extraction algorithm. Researchers have proposed different ways to quantize the color space. Smith and Chang [8] have proposed to partition the HSV color space into 166 bins. They place more importance on hue (18 levels) than on the value of saturation (only 3 levels). They also preprocess the images by passing it through a median filter. This preprocessing eliminates the outliers and enhances the prominent

color regions.

Several techniques have been proposed to integrate the spatial information with the color histograms. Gong et.al. model the color-spatial information of an image by splitting it into nine equal sized sub-images [9]. They represent each sub-image by a different color histogram. This approach is simple but is expensive both to compute and to store. W. Hsu, T.S. Chua and H.K. Pung propose the method of maximum entropy discretization with event covering method to include the spatial information [10]. Smith and Chang propose the usage of back-projection of binary color sets to extract color regions from images [11]. This technique provides for both the automated extraction of regions and representation of their color content. They have implemented this technique in the VisualSEEk content-based image/video retrieval system designed for the World Wide Web [12]. M. Stricker and A. Dimai first split the image into 5 partially overlapping fuzzy regions. Then they combine the color feature similarity of each of these sub-images by attributing more weight to the central region. However, this solution is highly domain dependent. It may be effective for an archive of photographs but it might not work well in other application areas [13, 14].

Pass and Zabih propose histogram refinement technique that splits the pixels in a given bucket into several classes, based upon some local property. From every bucket, only pixels in the same class are compared. They calculate a split histogram called a color coherence vector (CCV). CCV partitions each histogram bucket based on spatial coherence [15] [16]. The pixels are classified as coherent or incoherent to a given color. A region is defined as coherent if it is about 1 per cent of the total image. The total number of coherent and incoherent pixels are determined which then form the CCV. Cinque et.al. propose image color indexing scheme based on spatial chromatic histograms (SCH). SCH combines information about the location of pixels of similar color and their arrangement within the image [17]. M. Mitra, J. Huang and S.R. Kumar propose new color features called color correlograms for color content of the images. Color correlograms include the spatial correlation of colors, and can be used to describe the global distribution of the local correlations [18]. Stricker

and Orengo use the first three central moments of the probability distribution of each color. In order to compare two images according to their color moments, they propose a similarity function that consists of a weighted sum of the absolute differences of the moments summed over all color channels [19]. However the color moments are very sensitive to the variations in the intensity.

The techniques discussed above are based on the spatial domain processing of the images. Jacobs, Finklestein and Salesin have proposed the application of a truncated, quantized two-dimensional wavelet decomposition of the images [20]. The images are first decomposed using the Haar wavelet transform. The similarity is measured by calculating the number of significant wavelet coefficients that are close to each other in the query image and the test images. To reduce the computational complexity, the images are first subsampled at 128×128 and colors are quantized with 6 bins for each of the RGB component. For painted queries, 60 largest coefficients are used for each color, while for scanned images, 40 largest coefficients are compared. Color based image retrieval schemes based on wavelet transform are also proposed in [21, 22].

2.1.2 Texture Based Retrieval

Texture is a very powerful visual characteristic of the images that is very hard to define. The most salient features with respect to the human perception are periodicity, coarseness, preferred direction and degree of complexity. The texture describes the orientation and the spatial depth between the overlapping objects. Texture similarity is more difficult to describe with low level features compared to the color similarity. Two images can be considered to have similar texture when they show similar spatial arrangements of colors (or gray levels), but not necessarily the same colors (or gray levels). There are many techniques to extract the texture information from the images based on certain models. Some of the techniques are discussed in the following paragraphs.

The statistical moments of the gray-level histogram are used for the texture representation. The variance is commonly used as the feature representing the texture. Histogram information can also provide additional texture measures such as unifor-

mity and average entropy. The use of descriptors such as energy, entropy, contrast and homogeneity derived from the image's gray-level co-occurrence matrix are proposed by Haralick et.al. [23]. Gotlieb and Kreyszig evaluate the performance of the statistics proposed by Haralick. Their experiments indicate that contrast, inverse deference moment and entropy have the biggest discriminatory power among the proposed statistical measures [24]. Tamura et.al. have developed computational approximations to the visual texture features based on psychological studies. The six visual textual properties are coarseness, contrast, directionality, linelikeness, regularity and roughness [25].

After the introduction of the wavelet transform in early 1980, many researchers have used the wavelet transform for the texture feature extraction. Smith and Chang employ the mean and variance of the wavelet coefficients as texture features. They achieved a high accuracy of retrieval [26]. Tree structured wavelet transform is used by Chang and Kuo to further improve the classification accuracy [27]. Gross et. al. apply wavelet transform with KL expansion and Kohonen maps to perform texture analysis [28]. In [29] and [30] the texture features are extracted by wavelet transform with co-occurrence matrix. Ma and Manjunath perform a comparison for the texture representations by various wavelet transforms including the orthogonal, bi-orthogonal, tree-structured and Gabor wavelet transform. Their experimental evaluation showed that the Gabor wavelet transform performed the best among the tested wavelet kernels [31].

2.1.3 Shape Based Retrieval

The shape is a very important characteristic for describing the objects. The human beings can identify the shape of the objects by only a few clues and can associate even the broken edges. The human notion of similarity among the shapes is based on the topological closeness of edges and lines in space. The color and texture are both global attributes of an image. However, the representation of shape requires some kind of region identification. Shape representations can be divided into two categories: boundary-based and region-based. The boundary-based representation

uses only the outer boundary of the shape while region-based representation uses the entire shape region [33]. The most successful representations for these two categories are Fourier Descriptors and Moment Invariants.

The early work on Fourier Descriptors can be found in [33] and [34]. Rui et. al. modified the Fourier Descriptors to make it robust to noise and invariant to geometric transformations [35]. Hu identified seven moment invariants for the description of the shape [36]. Many improved versions based on his work have been developed by other researchers. The recent work on the shape description includes Finite Element Method, Turning Functions [37] and Wavelet Descriptor [38].

2.1.4 Concept Based Retrieval

Concept based retrieval is a recent development that tries to eliminate the inconvenience brought by the popular query by example (QBE) paradigm. Images and audiovisual documents incorporate more than just natural language contents. Semantic retrieval will need to embrace mental entities communicated in non verbal languages. A document is treated here as a piece of information representing thoughts expressed in a certain concept language. The key to concept-based retrieval is to allocate the lexicon and grammar of that concept language and to built index and query structures upon them. We seek to allocate all distinct words of a concept language and the mechanism by which a more specific concept phrase is expressed by using words in that language. Subsequently the words which we call *elecepts* are used to index the documents while the concatenation mechanisms (generative grammar) such as the concurrency (AND) and adjacency (ADJ) operators are supported in the query operation. The basic design of this methodology is to derive the *elecepts* and generative grammar of the concept language. The documents are indexed with the *elecepts* and generative grammar rules are embedded into the concept query operation. First the *relecepts*-aspects of relevance under which documents are retrieved such as perceptual similarity of the document collection are identified. Each *relecept* is then subjected to a generative concept analysis where *elecepts* and generative grammar of the concept language are derived. Documents are then indexed with *elecept* indices, whereas

generative grammar is used to operationalize the query operation.

The use of *elecept* indices allows a database to be indexed more economically. As *elecepts* are finite discrete entities, a concise description such as by using the semantic description scheme of MPEG-7 can be devised. Once the *elecept* indices are built, extensive query-ability support may be operated through the post-coordinate indexing scheme. The generative grammar comprising rules by which *elecepts* are synthesized to produce a compound concept in the language is derived and made accessible to the query operation. By rendering accessible the *elecepts* and the generative grammar, a large number of concept queries can be post-coordinately posed by using *elecepts* and the grammar operators. The approach has been successfully applied to retrieval of artistry documents [39, 40].

2.1.5 Relevance Feedback in Image Retrieval

The focus of the early attempts in the field of CBIR was to develop fully automated, open loop systems without any user feedback. There is a big semantic gap between high level concepts understood by the human perception and the low level features used for image representation. The human perception is also subjective which means that different human beings may interpret the same visual content differently. The subjectivity of the human perception is another factor in limiting the success of fully automated systems. This scenario served as a motivation to include the human user in the loop. The process of gathering feedback information from the users by presenting partial retrieval results is called *Relevance Feedback.* Different approaches have been adopted to incorporate the feedback information provided by the users. Some of the approaches are reviewed in the following paragraphs.

The relevance feedback can be used to update the weights associated with the component features [41]. This approach of is used in the MARS project [42]. A Bayesian learning based on a probabilistic model of a user's behavior is proposed by Cox et.al. [43]. The predictions of the model are combined with the selections made during a search to estimate the probability associated with each image. These probabilities are then used to select images for display. T.P. Minka and R.W. Picard pre-compute

many plausible groupings of the data. The system then selects and combines these groupings based on the positive and negative examples from the user. The relevance information can be feed back to modify these groupings or influence future grouping generation. In this way, the system is not only trained during individual example based sessions with the user but also trained across sessions [44]. Peng proposes a locally adaptive technique for CBIR that enables relevance feedback to take on multi-class form. He estimates local feature relevance based on Chi-squared analysis using information provided by multi-class relevance feedback. Local feature relevance is used to compare a flexible metric that is highly adaptive to the query locations [45]. Ashwin et.al. introduce the idea of negative feedback to estimate the parameters of the model. The proposed algorithm iteratively updates the parameters of the similarity metric so as to fit relevant examples while excluding the irrelevant ones. This is achieved by modifying the weights associated with the relevant examples [46].

Vasconcelos and Lippman present a Bayesian learning algorithm that relies on belief propagation to integrate user feedback [47]. Bayesian retrieval leads to a criteria for evaluating local image similarity without segmentation. This type of retrieval system requires users to provide image regions, or objects, as queries. Zhang et.al. investigate the application of support vector machines (SVM) in relevance feedback for region-based image retrieval in [48]. Both the one class SVM as a class distribution estimator and two class SVM as a classifier are taken into account. For the latter, two representative display strategies are studied. A new kind of kernel that is a generalization of Gaussian kernel is proposed. However, Chen et.al. introduce a modified SVM algorithm for one-class model to deal with the small samples in image retrieval applications with positive examples [49]. Wu, Tian and Huang investigate the possibility of taking advantage of unlabelled images in the given image database to make feasible a hybrid statistical learning. Assuming a generative model of the database, the proposed approach casts image retrieval as a transductive learning problem in a probabilistic framework [50]. Wang et. al. cast the image retrieval problem in the optimal filtering framework. They employ an optimal filter based on LMS and RLS algorithms to implement relevance feedback [51]. Zhuang, Liu, and Pan developed

a network of semantic templates to allow semantic search via a key-word language [52]. La Cascia et.al. proposed a system that combines textual and visual statistics in a single index vector for content-based search of a WWW image database. Textual statistics are captured in vector form using latent semantic indexing (LSI) based on text contained in the HTML document. Visual statistics are captured in vector form using color and orientation histograms. By using an integrated approach, it becomes possible to take advantage of possible statistical couplings between the content of the document (latent semantic content) and the contents of images (visual statistics) [53].

The unification of keywords and visual feature content was considered by Zhou and Huang. They propose a joint querying and relevance feedback scheme based on both keywords and low-level visual contents incorporating keyword similarities. An algorithm is also developed for the learning of the word similarity matrix during user interaction, namely word association feedback [54]. Laaksonen, Koskela and Oja, introduce the application of tree-structured self-organizing feature map (SOM) for the implementation of relevance feedback. They use the standard visual descriptors defined in MPEG-7 standard [55]. Muneesawang, and Guan have adopted a radial basis function (RBF) method for implementing an adaptive metric which progressively models the notion of image similarity through continual feedback from the users. They apply the proposed approach on image database compressed by wavelet transform and vector quantization coders [56]. Rui and Huang present a vigorous optimization formulation of the learning process and solve the problem in a principled way. By using Lagrange multipliers, they have derived explicit solutions, which are both optimal and fast to compute [57].

2.2 Content-Based Video Retrieval

A number of techniques developed primarily for CBIR were extended to content-based video retrieval (CBVR). However, this extension is not straight forward as video contains huge amount of data. Video has both spatial and temporal dimensions and video index should capture the spatio-temporal contents of the scene. In order to achieve this, a video is first segmented into smaller units. This is the first stage

which decomposes the huge video data into smaller and manageable units called shots, episodes and scenes. The key-frames are then identified and used for indexing and retrieval. This process is known as video parsing. The second stage is to find suitable attributes to the video content. The efficient access to the video data requires the tools for both of the concepts. There are many applications for video parsing such as generation of highlights and video summarization.

2.2.1 Key-Frame Based Video Retrieval

The indexing process is the next step to the segmentation of video signals. Arman et.al. propose to automatically extract a reference frame from each shot segment to facilitate efficient video browsing comparable to the fast-forward and fast-rewind functionality of a conventional video cassette player. This reference frame called the key-frame is used as the representative frame of the video segment. The results of the retrieval are displayed in the shape of key-frames [58]. Zhang et.al. perform the content-based indexing of a video documents by the CBIR operation over the key-frames. This scheme allows a video to be characterized by using a manageable sum of images representing perceptually distinct shot segments of the video content [59, 60]. Pickering et.al. propose a complete video parsing and retrieval system. They use color histogram for the detection of short boundaries. For the detection of abrupt transitions, the histograms of the two consecutive frames are compared. Each frame (image) is divided into 9 blocks. The histogram of all the nine blocks is calculated for each of the RGB component. They also detect the gradual transitions by taking the difference in histograms over a number of consecutive frames [61]. Each of the video shot is represented by a key-frame. The indexing is done using a variety of features including HSV histogram and text descriptors. In [62], Nagasaka and Tanaka attempt to search for objects by using local color histograms and color map based on the QBE paradigm.

2.2.2 Motion Based Video Retrieval

The motion is a very important characteristic that provides dynamic content analysis of the video data. Many events are easily defined by describing it with the amount of motion present in a sequence of frames. Temporal processing of the video is required to extract motion vectors/descriptors from the videos. The motion between the consecutive frames as well as over a number of frames is used for indexing. In compressed domain processing of the videos such as MPEG encoded video sequences, the motion vectors can be obtained directly from the compressed bitstream. This has the advantage of saving the de-compression overhead.

There are two main approaches to motion analysis. The first one employs segmentation, tracking and characterization of moving elements in order to determine a spatio-temporal representation of the video shot [63]. This approach utilizes either parametric motion models or dense optical flow fields. The description of motion content generally relies either on the extraction of qualitative pertinent features for the entities of interest such as the direction of the displacement or to the trajectory of the center of gravity of the tracked objects [64] or on the computation of global histograms of estimated dense optical flow fields [65]. However, there may be cases where the entities of interest are not single objects. In those cases, video cannot be handled in such a way. In the second approach, the interpretation of dynamic content is achieved without any explicit prior motion segmentation. The pioneering work presented in [66] leads to the definition of temporal textures. The features extracted from spatial co-occurrences of normal flows are used to classify the sequences either as simple motions (rotation, translation, divergence) or as temporal textures. In [67], features are extracted from the characterization of surfaces derived from spatio-temporal trajectories. A probabilistic modelling of dynamic content and an associated statistical scheme for motion-based video indexing and retrieval is presented in [70].

2.2.3 Audio Based Video Retrieval

Audio is a rich source of content information in videos. The analysis of the audio may provide useful clues about the video content. In certain cases, it is easy to find

the events in audio domain than in the visual domain such as explosions, goal events in a soccer game or news items with a particular broadcaster. Most of the research in content-based video retrieval is focussed on the visual properties of the video data. A variety of audio characteristics have been used for retrieval such as loudness, pitch, brightness, bandwidth, and harmonicity.

Saunders classified the audio into speech and music based on zero crossing rate and audio energy [71]. Wold et. al. classified the audio into 10 different classes [72]. The speech recognition and related techniques have progressed greatly in the past few years. But processing of non-speech audio has witnessed a little progress. Some systems exploit speech recognition for the extraction of features. Speech is converted to a text document after speech recognition. The text document can be indexed and retrieved using the standard text retrieval methods [73, 74]. Some of the work on the segmentation and classification of audio streams in video has been reported in [75] [76]. The result of audio segmentation and classification can be integrated into video classification and retrieval system as an important factor. However, the techniques to classify the audio or the embedded audio may not be useful for the retrieval of news video. In this type of application, music, speech, noise and crowd voice may be found together in the same video clip. Hence we need features that represent the global similarity of the audio content.

2.2.4 Caption Based Video Retrieval

The caption text in the videos can also serve a useful descriptor for video indexing and retrieval. An example work based on the caption text can be found in [77] where Tang, Gao, Liu, and Zhang present a video caption detection and recognition system based on a fuzzy-clustering neural network (FCNN) classifier. They develop a caption-transition detection scheme to locate both spatial and temporal positions of video captions with high precision and efficiency.

2.2.5 Retrieval of Semantics

The integrations of the multi-modal features is a challenging task in the representation of the video data. The important issue is the retrieval of sequences that carry a high level concept. In other words, understanding the semantic contents of the video clips is very important for the video classification from the users' point of view. Most of the work in CBVR is concentrated in the area of shot level retrieval and segmentation. In recent days, attempts are underway for providing higher level access to the video data. A few works have been summarized below.

Wang, Naphade, and Huang purpose a post-integration model that integrates low-level media types to identify visually and auditorily similar video segments. Relevance feedback is used to improve the speed and accuracy of searching in a video database. The model first treats the underlying media as independent processes and then combines distance scores from each of the underlying media at the later stage [78]. Naphade et.al. propose a dual probabilistic framework called multiject and multinet to respectively model the semantic concepts and their contextual constraints in the feature space. In the framework, semantic content of a video segment is conceived as objects (man, helicopter) and events (explosion, ball-game, man-walking) that occur at certain sites (outdoor, beach). The retrieval for a concept was operationalized as a pattern recognition task over the multimodal feature indices. As with semantic visual template, this work also attempts to bridge the semantic-gap by assuming the existence of a consistent correlation between a semantic concept and its perceptual features. The fusion of multimodal features is attempted in [79]. Haering, Qian, and Sezan propose a three-level video-event detection methodology and apply it to animal-hunt detection in wildlife documentaries. The first level extracts color, texture, and motion features, and detects shot boundaries and moving object blobs. The mid-level employs a neural network to determine the object class of the moving object blobs. This level also generates shot descriptors that combine features from the first level and inferences from the mid-level. The shot descriptors are then used by the domain-specific inference process at the third level to detect video segments that

match the user-defined event model [80].

Jeong et.al. used fuzzy triplets for indexing video shots and formulating queries. The fuzzy triplets and rules define generic spatio-temporal patterns in video streams by specifying the relative spatial relationships between salient objects and their change with respect to time [81]. In [82], Smith and Kanade attempt to characterize video document by using a combination of keywords and perceptual features. Keywords are derived through speech recognition, while detection algorithms are used to allocate frames containing human face and text objects. Nakamura and Kanade propose the spotting by association method to enhance the detection of semantics in news segments by using context hints derived from visual and natural language information. The method is based on the assumption that there exists a consistent pattern of correlation between visual scene and natural language usage on the staging of certain news topics [83].

Chapter 3

Concepts and Laplacian Mixture Model

3.1 Introduction

Mathematical transformation is a very powerful tool in signal analysis. The signals are transformed from one domain into another to extract certain signal properties that are not easily observable. Another purpose of the transformation is to simplify mathematical operations. Most of the signals that we come across are time-domain signals. In the time-domain representation, the variation of certain quantity such as amplitude, energy and power are represented as a function of time. This representation of the signals may not be the best representation for a signal processing application. Therefore, in many cases we transform the signal into a different domain by applying a mathematical transformation. Many of the signal properties may be more visible in the frequency domain [84]. Fourier transform is the tool that transforms the signal into frequency domain. Fourier transform and its inverse are defined as [85]:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) \exp(-\jmath\omega t) dt \tag{3.1}$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \exp(\jmath\omega t) d\omega \tag{3.2}$$

where $F(\omega)$ is the Fourier transform of the signal $f(t)$. $exp(\jmath\omega t)$ is defined using the Euler's formula

$$exp(\jmath\omega t) = \cos(\omega t) + \jmath \sin(\omega t) \tag{3.3}$$

It is obvious from the above equations that the Fourier transform represents the signal in terms of *sine* and *cosine* basis functions. These basis functions are however limited in time only. The Fourier transform is a reversible transform which allows to go back and forth between the time domain and the frequency domain. However, the signal is a function of time (in time domain representation) or a function of frequency (in frequency domain representation). This is sufficient for the stationary signal analysis. The Stationary signals are signals whose frequency contents do not change with time. Hence, in this type of analysis we are interested in knowing the total frequency content of the signal. But most of the signals in practice are non-stationary meaning that their frequency contents change with time. The human auditory system depends on both the time and frequency parameters. A time-frequency representation of such signals is required for analysis. The Fourier transform is extended to achieve this time-frequency representation. This is known as Short-Time-Fourier Transform (STFT) and is defined as:

$$STFT(\tau, \omega) = \int s(t) g(t - \tau) \exp(-\jmath \omega t) dt \tag{3.4}$$

STFT is the Fourier transform of the signal $s(t)$ after applying a window function $g(t-\tau)$ around the time τ. This window function $g(t-\tau)$ is shifted to cover the whole signal. Consecutive Fourier transforms are applied on overlapped data. In this way, we obtain a time-frequency representation of the signal. STFT assumes that the signal is stationary for the duration of the window. The choice of the window size is very critical. Choosing a short window size will achieve a good resolution in time. But the number of samples used in the Fourier transform will also be reduced. This reduces the number of discrete frequencies that can be represented in the frequency domain. On the other hand, a large window will increase the frequency resolution at the expense of the time resolution. Therefore, there is a trade off between the frequency and time resolutions. STFT is suitable for applications where high resolution is not required. The STFT covers the time-frequency plane with a uniform array of resolution squares as shown in Figure 3.1. The dimension of the time-frequency tile represents the minimum time and frequency intervals over which separate signals can

be differentiated.

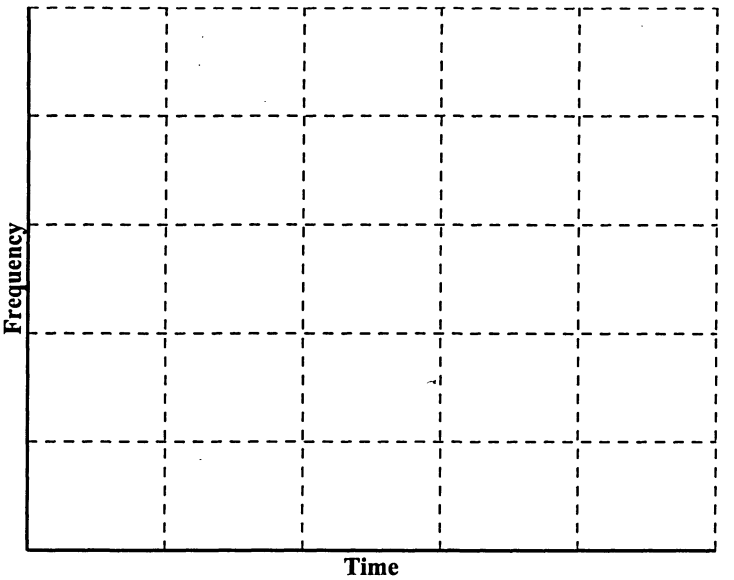

Figure 3.1: STFT Coverage of Time-Frequency Plane

3.2 Wavelet Transform

3.2.1 Continuous Wavelet Transform

The continuous wavelet transform (CWT) was developed as an alternative to the STFT. The approach is similar to STFT in the sense that the function is multiplied with a *wavelet* similar to the window function in the STFT. However, there are several important differences between the two techniques:

1. The width of the window used by the Wavelet transform is not constant and changes as the transform is calculated for every single spectral component. This is the most significant characteristic of the wavelet transform.

2. Negative frequencies are not computed.

The continuous wavelet transform is defined as [86]:

$$CWT(\tau, s) = \Psi(\tau, s) = \int_{-\infty}^{+\infty} x(t)\psi^*_{\tau,s}(t)dt \tag{3.5}$$

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{|s|}}\psi(\frac{t-\tau}{s}) \tag{3.6}$$

where τ and s are translation and scale parameters respectively and the superscript $*$ refers to the complex conjugate. $\psi(t)$ is known as the *Mother wavelet* that serves as a prototype for generating other window functions $\psi_{\tau,s}(t)$ known as the *daughter wavelets.* The daughter wavelets are obtained by shifting and scaling the mother wavelet. It must be noted here that the wavelets are finite length oscillatory functions. The translation process is similar to that of STFT; where the window function is moved over the entire signal. In the Wavelet transform, the scale parameter (s) replaces the frequency parameter. Large scales represent the global view of the signal or the low frequencies; while low scales represent the high frequencies. Mathematically speaking, large scales dilate the signal while low scales correspond to the compressed signals. CWT is a continuous transform and, therefore, τ and s are incremented continuously. Since the transform is to be computed using a digital computer, both parameters are increased by a sufficiently small step size. This means that the time-scale plane is sampled and becomes discrete.

Continuous wavelet transform is a reversible transform subject to the constraint:

$$\{2\pi \int_{\infty}^{-\infty} \frac{|\hat{\psi}(\xi)|^2}{|\xi|} d\xi\}^{\frac{1}{2}} < \infty \tag{3.7}$$

where $\hat{\psi}$ is the Fourier transform of ψ. The above equation implies that $\hat{\psi}(0) = 0$ or:

$$\int \psi(t)dt = 0 \tag{3.8}$$

This is not a very restrictive condition and wavelet functions can be found that satisfy the above condition.

In the case of CWT analysis of the signal, the discretization process may be performed in any desired way. However, the *Nyquist* sampling rate is the minimum sampling rate required for the reconstruction/synthesis of the signal. Mathematically, the discretization process is defined by the following equation:

$$\psi_{n,k}(t) = s_0^{-n/2}\psi(s_0^{-n}t - k\tau_0) \tag{3.9}$$

where $s_0 > 1$ and $\tau_0 > 0$ are the discrete versions of scale and translation parameters while n and k are the step sizes for scale and translation respectively. The *scale* parameter is discretized on a logarithmic grid while the time parameter is discretized based on the *scale* parameter. This means that the sampling rate for the time parameter is dependent on the value of the *scale*; and is different for different *scales.* Figure 3.2 shows the sampling grid which is *dyadic* in nature: The base of the log-

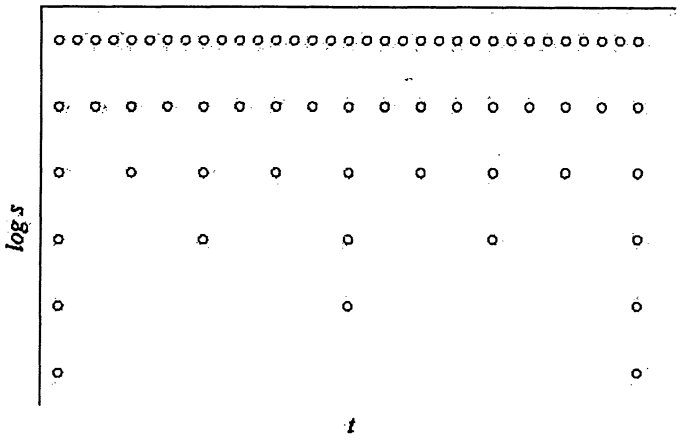

Figure 3.2: Sampling grid for time-scale plane

arithm depends on the user, 2 being the most common one. Only a finite number of points are taken. For example, if the base of logarithm is 2 then *scale* will have values 2,4,8,16 and so on.

3.2.2 Discrete Wavelet Transform

A discrete version of the transform is required so that it can be computed using digital computers. We can calculate the CWT using a digital computer by discretizing the time-scale plane as shown in the previous section. Discretized CWT is the sampled version of the continuous wavelet transform and gives a lot of redundant information consuming a large amount of computational resources. To reduce the computational complexity, a discrete wavelet transform (DWT) is defined. DWT provides sufficient

information for signal analysis and reconstruction.

The discrete wavelet transform is obtained by passing the signal through a series of low pass and high pass filters. When a signal is passed through a filter, the signal is convolved with the impulse response of the filter to produce the output signal. The filtering operation for a discrete signal is defined as follows:

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \tag{3.10}$$

where $y(n)$ is the output of the system with impulse response $h(n)$ and $x(n)$ is the input signal. The filters are half band digital low pass filters. The output of these filters therefore contains only the frequencies up to half of the maximum frequency of the original signal. The frequencies higher than half of the maximum frequency in the original signal are removed by the low pass half band filter. Therefore, we can eliminate half of the samples by subsampling without any loss of information. This subsampling operation doubles the scale of the signal since half of the samples are now removed. The filtering operation, on the other hand, reduces the frequency resolution by removing half of the spectral components from the signal. Mathematically, the procedure is defined as:

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)x(2n-k) \tag{3.11}$$

If $h(n)$ is a low-pass filter then its corresponding mirror filter $g(n)$ is defined as:

$$g(n) = (-1)^n h(N-1-n) \tag{3.12}$$

These filters $g(n)$ and $h(n)$ are called quadrature mirror filters (QMF). The discrete wavelet transform is implemented by a quadrature mirror filter bank. Discrete wavelet transform is given by:

$$\psi_{j,k} = a_0^{j/2}\psi(a_0^j t - k) \tag{3.13}$$

Taking $a_0 = 2$, the scaling function is defined as:

$$\phi(t) = \sqrt{2}\sum_n h(n)\phi(2t-n) \tag{3.14}$$

and the wavelet function $\psi(t)$ is:

$$\psi(t) = \sqrt{2}\sum_n g(n)\phi(2t-n) \tag{3.15}$$

The decomposition and reconstruction of the signal using the above procedure is shown in the Figures 3.3 and 3.4. $\bar{h}(n)$ and $\bar{g}(n)$ are inverse filters of $h(n)$ and $g(n)$ respectively.

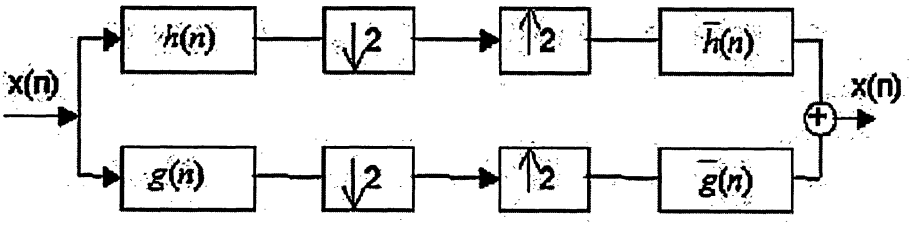

Figure 3.3: One-level decomposition and reconstruction of a signal by DWT

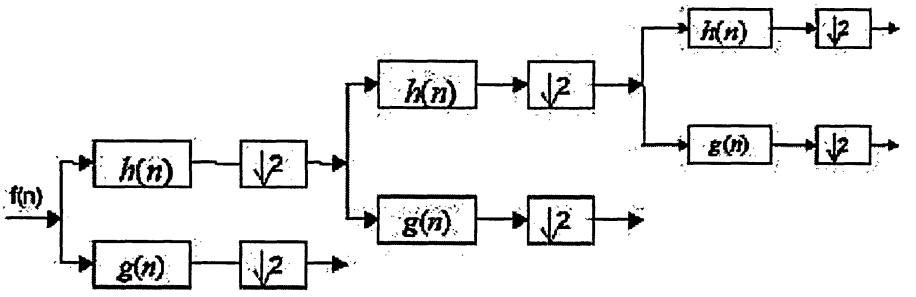

Figure 3.4: Multilevel decomposition of a signal by DWT

The image data is two dimensional. Therefore 2-D version of the discrete wavelet transform is required. The 2-D version of the DWT is obtained by performing the 1-D wavelet transform first on rows and then on columns of the data. This is illustrated in Figure 3.5. The H and G represent the QMF filters described above and $x(m,n)$ is the 2-D input signal. The LL (Low-Low) subband contains the low spatial frequencies in both horizontal and vertical directions. That means the edge information is not present in this subband. The LH (Low-High) subband contains low spatial frequencies in vertical direction and high frequencies in the horizontal direction. The information about the horizontal edges is obtained from this subband. The HL (High-Low) subband contains low spatial frequencies in horizontal direction and high frequencies in the vertical direction. The information about the vertical edges

is present in the HL subband. The HH (High-High) subband contains both vertical and horizontal high frequencies, and provides the diagonal edge information.

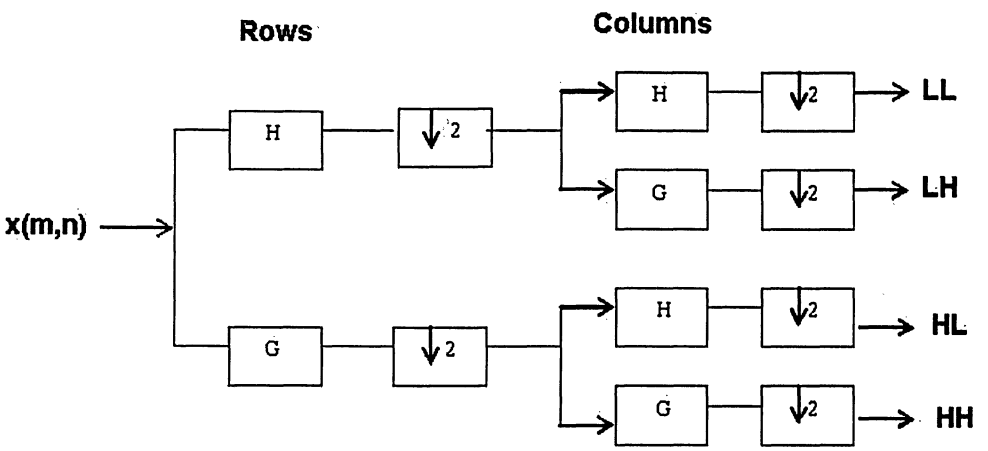

Figure 3.5: 1-level decomposition of a 2-D signal by DWT.

3.3 Finite Mixture Models

Finite mixture models are widely used in the statistical modelling of data. It is a very powerful tool for probabilistic modelling of the data produced by a set of alternative sources. Finite mixtures represent a formal approach to unsupervised classification in statistical pattern recognition. The usefulness of this modelling approach is not limited to clustering. They are also able to represent arbitrarily complex probability density functions (pdf) [87]. The distributions of the wavelet coefficients are very peaky due to the energy packing property of the wavelets. Their modelling using fixed shaped distributions such as Gaussian or Laplacian gives rise to mismatches. The mixture modelling provides an excellent and flexible alternative for this kind of complex distributions.

Let $\mathbf{X} = [X_1, ..., X_d]^T$ be a d-dimensional random variable, while $\mathbf{x} = [x_1, ..., x_d]^T$ is a particular observation of this random variable. We suppose that the data has been generated by a finite mixture of k components. The probability density function

of this random variable $p(\mathbf{x}|\theta)$ is then defined as [88]:

$$p(\mathbf{x}|\theta) = \sum_{m=1}^{k} \alpha_m p_m(\mathbf{x}|\theta_m) \tag{3.16}$$

where $\alpha_1, ..., \alpha_k$ are the *mixing probabilities* and θ_m is the parameter set representing the m-th component. Also

$$\theta = \theta_1, \theta_2, ..., \theta_k \tag{3.17}$$

Therefore the complete set of model parameters $[\theta_1, ..., \theta_k, \alpha_1, ..., \alpha_k]$ is to be calculated to specify the mixture. Since α_m are probabilities:

$$\alpha_m \geq 0 \; and \; \sum_{m=1}^{k} \alpha_m = 1 \; for \; m = 1, ..., k \tag{3.18}$$

The mixtures can be built with different type of components. However, it is usually assumed that the components of the mixture have the same functional form such as Gaussian or Laplacian. There are two fundamental questions

1. How to calculate the parameters of the model?

2. How to find the number of components?

The *Expectation Maximization* (EM) is a standard approach to solve the first question. It is more difficult to find out the number of components. Once the number of components are fixed, we can apply EM algorithm to find out the model parameters.

3.4 Parameter Estimation

The maximum likelihood principle is widely used in the statistical analysis for estimation of the parameters of a probability density function. This principle was originally developed by R.A. Fisher and is based on the likelihood function. The likelihood of a set of data is defined as the probability of obtaining that particular set of data, given a probability distribution model. This probability is a function of the parameters of the model. This function is known as the likelihood function. The values of the parameters that maximize this sample likelihood function are known as the *Maximum Likelihood Estimators* or MLE's.

Given a data set of N points, $\mathcal{X} = [\mathbf{x_1}, \cdots, \mathbf{x_N}]$, which are i.i.d. (Independent identically distributed), we suppose that the underlying distribution from which these values were drawn is $p(\mathcal{X}; \theta)$. The likelihood function is then defined as:

$$\ell(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_{i=1}^{N} p(\mathbf{x_i}|\theta) \tag{3.19}$$

We want to maximize this likelihood function with respect to the parameters of the model. Instead of maximizing this function directly, we maximize the log likelihood which is the log of the above equation:

$$\ln(\ell(\theta|\mathcal{X})) = \ln \prod_{i=1}^{N} p(\mathbf{x_i}|\theta) \tag{3.20}$$

$$\ln(\ell(\theta|\mathcal{X})) = \sum_{i=1}^{N} \ln p(\mathbf{x_i}|\theta) \tag{3.21}$$

3.4.1 EM Algorithm

EM algorithm is based on the maximum likelihood principle for the estimation of the parameters of the underlying distribution from a given data. There are two main applications of the EM algorithm [89]:

- Estimation of the model parameters when the data has missing values.
- When the analytical optimization of the likelihood function is intractable. The function can be simplified by assuming the existence of a hidden variable whose values are missing.

Our objective is to determine the underlying probability distribution. We suppose that the data set was not produced by a single distribution but by a mixture of densities. We assume the existence of a hidden variable $\mathcal{Z} = [\mathbf{z_1}, \cdots, \mathbf{z_N}]$ whose values are not known. This M dimensional variable indicates which component has generated a particular data value. The graphical representation of this model is shown in Figure 3.6 assuming the constituent components as Gaussian with mean μ_m and covariance Σ_m. We try to maximize the log likelihood of the joint distribution $\ln p(\mathcal{X}, \mathcal{Z}|\theta)$. This quantity is called the complete log-likelihood. Since, we cannot

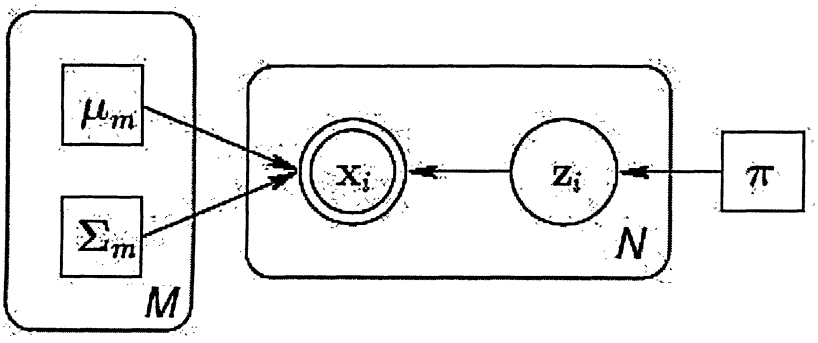

Figure 3.6: Graphical model for maximum likelihood density estimation using a mixture of Gaussian

observe the values of the random variables $\mathbf{z_i}$, we must work with the expectation of this quantity w.r.t some distribution $Q(\mathbf{z})$. The log of the complete likelihood can be written as follows [90]:

$$\ell_c(\theta) = \ln p(\mathcal{X}, \mathcal{Z}|\theta) = \ln \prod_i^N p(\mathbf{x_i}, \mathbf{z_i}|\theta) = \ln \prod_i^N \prod_m^M (p(\mathbf{x_i}|m, \theta)p(m))^{z_{im}} \tag{3.22}$$

$$\ell_c(\theta) = \sum_i^N \sum_m^M z_{im} \ln(p(\mathbf{x_i}|m, \theta) + z_{im} \ln \pi_m) \tag{3.23}$$

where $\prod$ denotes the product and $p(m)$ is the prior probability of $m-th$ component. The EM algorithm consists of two steps:

- The Expectation step or E-step calculates the expectation of the complete log-likelihood function.
- The Maximization step or M-step maximizes the expectation calculated in E-step.

E-Step:

Taking the expectation of Equation 3.23 with respect to (w.r.t) $Q(\mathbf{z})$, we get:

$$\langle \ell_c(\theta) \rangle_{Q(\mathbf{z})} = \sum_i^N \sum_m^M \langle z_{im} \rangle \ln(p(\mathbf{x_i}|m, \theta)) + \langle z_{im} \rangle \ln \pi_m \tag{3.24}$$

where $\langle\rangle$ is the expectation operator.

M-Step:

This step performs the maximization of the expectation value calculated above. Therefore differentiating the above equation w.r.t θ and putting it equal to zero, we can calculate the maximum value of expected complete log-likelihood as:

$$\frac{\partial \langle \ell_c(\theta) \rangle_{Q(\mathbf{Z})}}{\partial \theta} = \sum_i^N \sum_m^M \langle z_{im} \rangle \frac{\partial}{\partial \theta} \ln[p(\mathbf{x_i}|m, \theta] + \frac{\partial}{\partial \theta} \langle z_{im} \rangle \ln \pi(m) \tag{3.25}$$

$$\frac{\partial \langle \ell_c(\theta) \rangle_{Q(\mathbf{Z})}}{\partial \theta} = \sum_i^N \langle z_{im} \rangle \frac{\partial}{\partial \theta} \ln p(\mathbf{x_i}|m, \theta) = 0 \tag{3.26}$$

3.4.2 Gaussian Mixture Model

Let us assume that the underlying distribution is a mixture of M d-dimensional Gaussian components. Then the probability function $p(x_i|m, \theta)$ is simply a conditional probability of generating x_i given that the m-th model is chosen.

$$p(x_i|m, \theta) = \frac{1}{(2\pi)^{d/2}|\Sigma_m|^{1/2}} \exp(-\frac{1}{2}(x_i - \mu_m)^T \Sigma_m^{-1}(x_i - \mu_m)) \tag{3.27}$$

where μ_m is the mean value and Σ_m is the covariance matrix. A 1-dimensional Gaussian distribution with $\mu = 0$ and $\sigma = 1.1$ is shown in Figure 3.7. The Gaussian

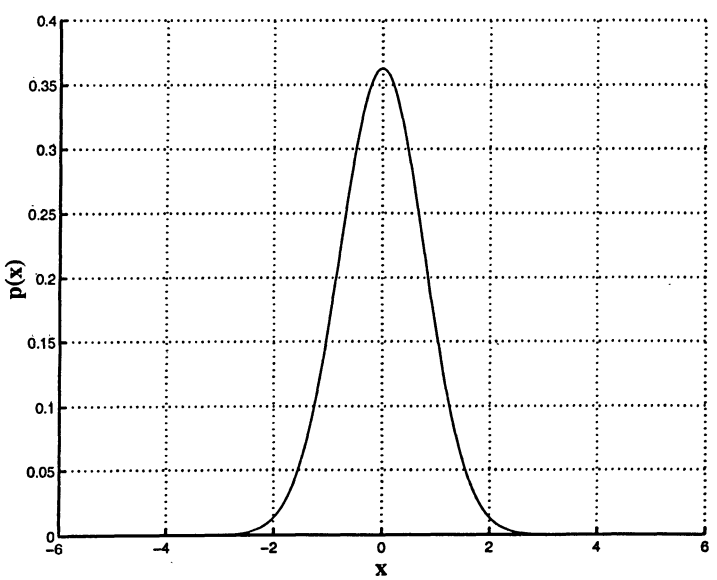

Figure 3.7: Gaussian Distribution, $\mu = 0$ and $\sigma = 1.1$

mixture model has been successfully applied for image retrieval by H. Yuan, X. Zhang and L. Guan in [91]. Although mixture of Gaussians is very widely used to model the shape of the unknown distributions, we still do not have the basis to show that the Gaussians are the best solution for concrete problems. If an infinite number of components were available in the Gaussian mixture then we could have modelled any arbitrary shaped distribution. This is however practically infeasible . This is the motivation to apply mixture of Laplacians approach to model the distribution of the wavelet coefficients.

3.4.3 Laplacian Mixture Model

The univariate Laplacian distribution is defined as:

$$p(x) = \frac{1}{2b}\exp(-\frac{|x-\mu|}{b}) \tag{3.28}$$

A Laplacian distribution with $\mu = 0$ and $b = 1$ is shown in Figure 3.8. Here we assume

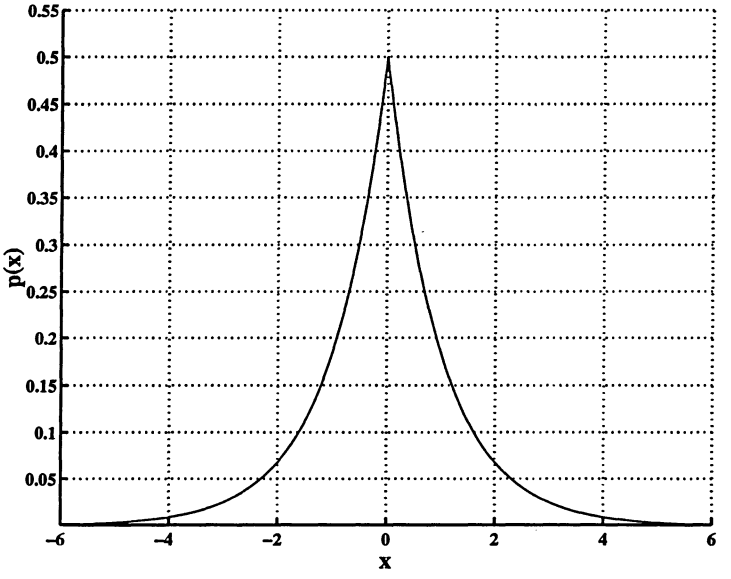

Figure 3.8: Laplacian Distribution, $\mu = 0$ and $b = 1$

that the underlying distribution is a mixture of M Laplacian components centered at zero ($\mu = 0$). Then the probability function $p(x_i|m, \theta)$ is simply a conditional

probability of generating x_i given that the m-th model is chosen.

$$p(x_i|m,\theta) = \frac{b_m^{-1}}{2}\exp(-|x_i|b_m^{-1}) \tag{3.29}$$

M-Step:

There is only one parameter b_m in the above equation i.e. $\theta_m = [\pi_m, b_m]$. Taking the *logarithm* and differentiating the above equation w.r.t. b_m^{-1}, we get:

$$\begin{aligned}
\frac{\partial}{\partial b_m^{-1}}\ln p(x_i|m,\theta) &= \frac{\partial}{\partial b_m^{-1}}\ln(\frac{b_m^{-1}}{2}\exp(-|x_i|b_m^{-1})) \\
&= \frac{\partial}{\partial b_m^{-1}}(\ln(2)) + \frac{\partial}{\partial b_m^{-1}}\ln(b_m^{-1})\frac{\partial}{\partial b_m^{-1}}(-|x_i|b_m^{-1}) \\
&= \frac{1}{b_m^{-1}} - |x_i| \\
&= b_m - |x_i|
\end{aligned} \tag{3.30}$$

Substituting this result in Equation 3.26, we have:

$$\sum_i^N \langle z_{im}\rangle(b_m - |x_i|) = 0 \tag{3.31}$$

and the following update equation:

$$b_m = \frac{\sum_i^N \langle z_{im}\rangle |x_i|}{\sum_i^N \langle z_{im}\rangle} \tag{3.32}$$

In order to maximize the expected log-likelihood in Equation 3.23 w.r.t. π_m, the constraint $\sum_m^M \pi_m = 1$ should also be enforced. This is achieved by using the Lagrange multiplier λ and augmenting Equation 3.23 as follows:

$$\acute{L}(\theta) = \langle \ell_c(\theta)\rangle_{Q(\mathbf{Z})} - \lambda(\sum_m^M \pi_m - 1) \tag{3.33}$$

Differentiating w.r.t. each π_m, we get:

$$\frac{\partial}{\partial \pi_m}\langle \ell_c(\theta)\rangle_{Q(\mathbf{Z})} - \lambda = 0,\ for 1 \le m \le M \tag{3.34}$$

Using Equation 3.23:

$$\frac{1}{\pi_m}\sum_i^N \langle z_{im}\rangle - \lambda = 0,\ for\ 1 \le m \le M \tag{3.35}$$

or equivalently

$$\sum_{i}^{N}\langle z_{im}\rangle - \lambda\pi_m = 0, \ for\ 1 \leq m \leq M \tag{3.36}$$

Summing Equation 3.36 over all M models we get:

$$\sum_{m}^{M}\sum_{i}^{N}\langle z_{im}\rangle - \lambda\sum_{m}^{M}\pi_m = 0 \tag{3.37}$$

But since $\sum_m^M \pi_m = 1$ we have:

$$\lambda = \sum_{m}^{M}\sum_{i}^{N}\langle z_{im}\rangle = N \tag{3.38}$$

Substituting this result back into Equation 3.38, we get the following update equation for π_m:

$$\pi_m = \frac{\sum_i^N \langle z_{im}\rangle}{N} \tag{3.39}$$

E-Step:

In the M-step, we have derived the update equations that maximizes the expected complete log-likelihood $\langle \ln p(\mathcal{X}, \mathcal{Z}|\theta)\rangle$. Now we have to ensure that we are actually maximizing the incomplete log-likelihood $p(\mathcal{X}|\theta)$ because it is the quantity that we are interested in to maximize. We are guaranteed to maximize the incomplete log-likelihood only when the expectation is taken w.r.t. the posterior distribution of $\mathcal{Z}$ namely $p(\mathcal{Z}|, \mathcal{X}, \theta)$. Therefore each of the expectations $\langle z_{im}\rangle$ that appear in the update equations should be computed as follows:

$$\langle z_{im}\rangle = \frac{p(x_i|m,\theta)p(m)}{\sum_j^M p(x_i|j,\theta)p(j)} = \frac{p(x_i|m,\theta)\pi_m}{\sum_j^M p(x_i|j,\theta)\pi_j} \tag{3.40}$$

3.5 Experimental Results

The experimental evaluation of the Laplacian mixture model (LMM) is carried out by taking two component mixture. Three-level wavelet decomposition of the texture image is performed using Daubechies-2 wavelet kernel. Figures 3.9, 3.10 ,3.11 and 3.12 show the normalized histograms of the wavelet coefficients and estimated pdfs in the corresponding subbands. It is evident that a mixture of only two Laplacian components is able to model the distribution of the wavelet coefficients. The proposed approach is based on the characteristics of the wavelet coefficient distributions.

Therefore, it is a general approach which is equally suitable for images and audio analysis. We will apply these results for indexing texture images in chapter 4. Its application to audio-based video indexing and retrieval is explored in chapter 5.

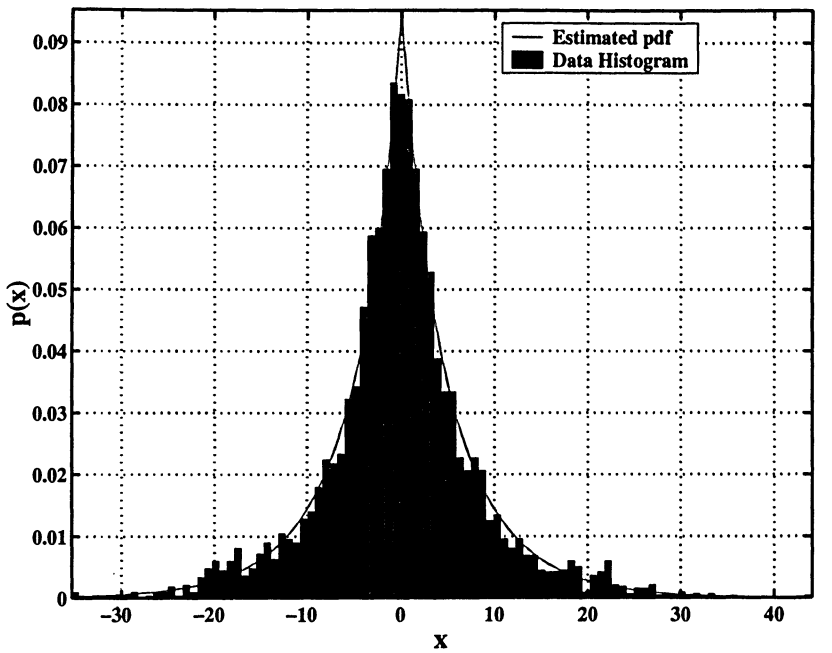

Figure 3.9: Estimated pdf of Wavelet Coefficients in LH subband at Level-1

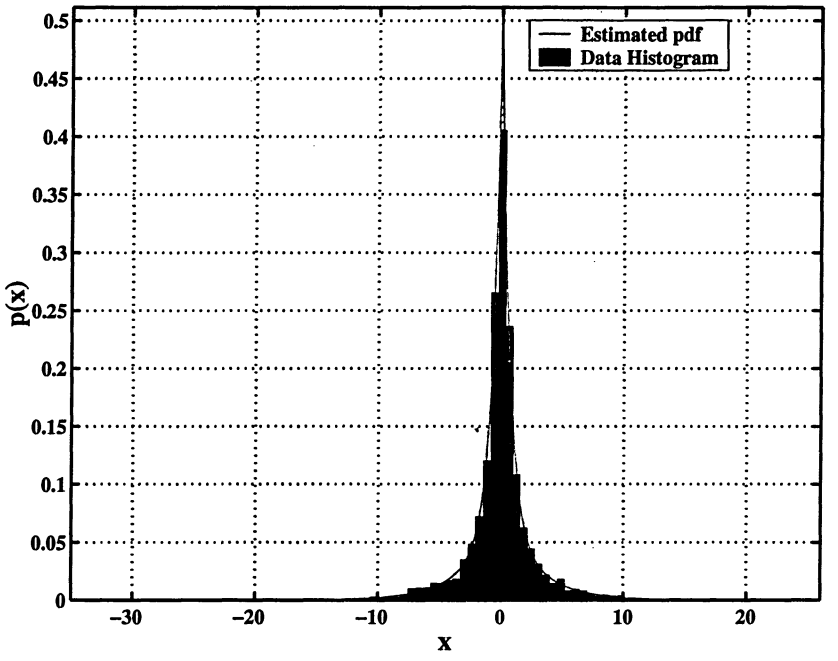

Figure 3.10: Estimated pdf of Wavelet Coefficients in HH subband at Level-1

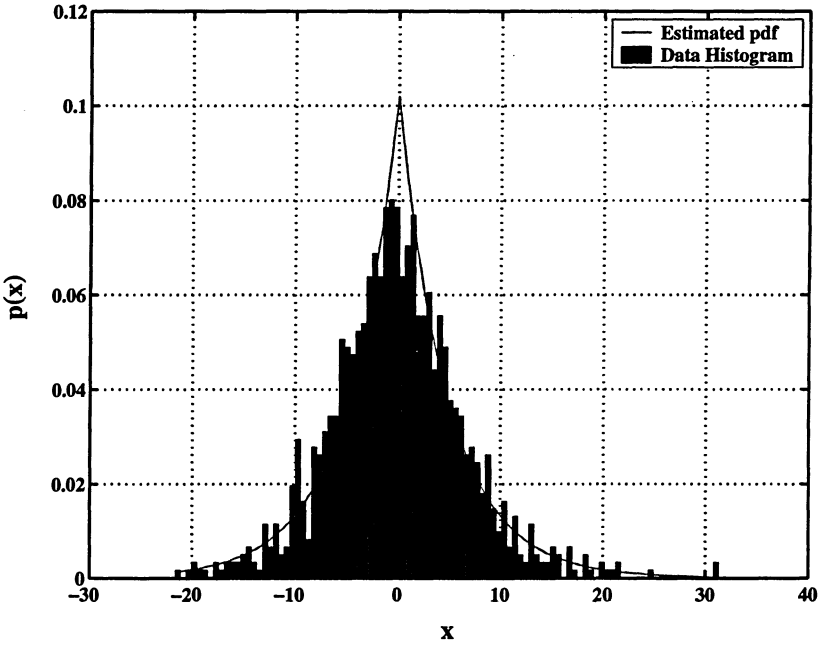

Figure 3.11: Estimated pdf of Wavelet Coefficients in HH subband at Level-2

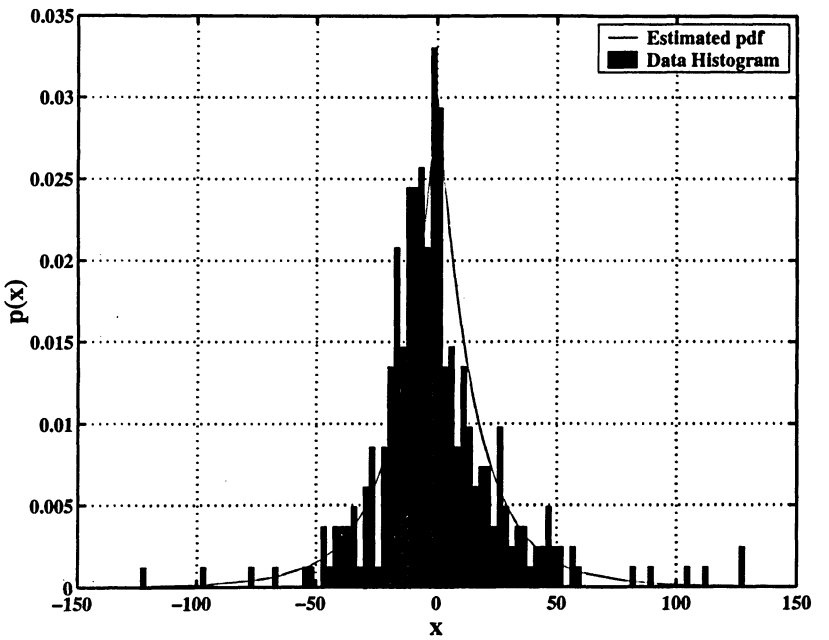

Figure 3.12: Estimated pdf of Wavelet Coefficients in HL subband at Level-3

Chapter 4

Laplacian Mixture Model for Image Retrieval

4.1 Introduction

The main components of Content-Based image retrieval system are shown in Figure 4.1. A brief description is given below:

Content Analyzer: This stage consists of a set of algorithms that perform the image analysis for extraction of visual content descriptors from the image data. These descriptors, also known as features, are used to represent the different visual contents of the images such as color, texture and shape etc. The same visual content may be represented by more than one descriptor. For example, color moments and color histogram are both used to represent the color of the image. The objective here is to find a set of features that model the human visual perception. Ideally, the values of these features should lie very close for similar images and far apart in case of different images.

Search Engine: This component performs the query formulation and ranking tasks. The users can present a query to the system in different forms such as in the form of an example image. The search engine ranks the images according to some similarity measure. The similarity measure actually maps the distance between the feature vectors of the target image and the query image to a real valued number. This number quantifies the visual similarity between the images.

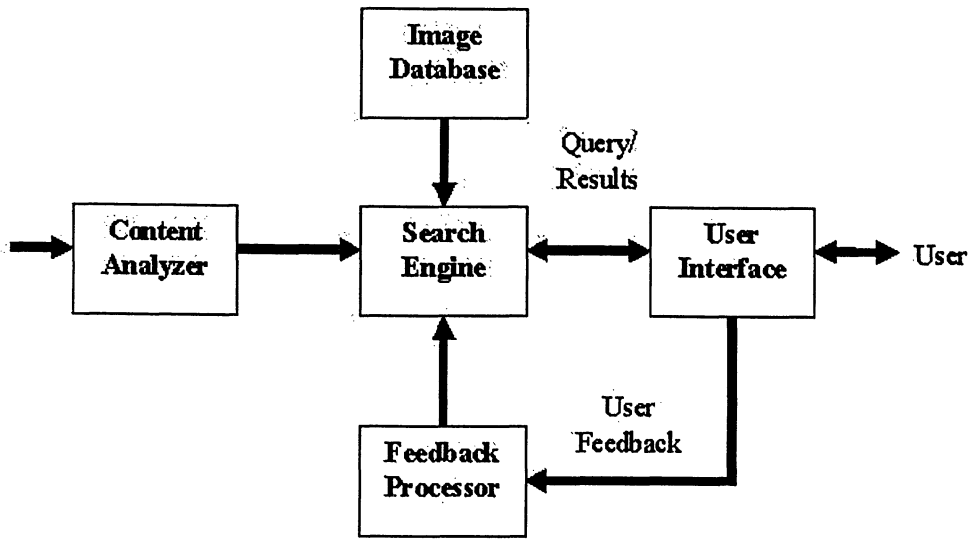

Figure 4.1: Block Diagram of a CBIR system

User Interface: The graphical user interface (GUI) is provided so that the users can interact with the system. The users can search the image database by providing a query through the GUI. The retrieval results are displayed using this GUI. In some CBIR systems, the user can change the parameters of the search process as well as choose the features for retrieval. However, the common users may not be familiar with the low level feature representations of the visual content. The users provide feedback to the system through the graphical user interface.

Feedback Processor: Relevance feedback is an important part of the modern image retrieval systems. This module processes the user feedback provided in different forms. The system parameters are modified and a new search cycle is performed to enhance the system retrieval accuracy. This iterative process is repeated until the user is satisfied with the results.

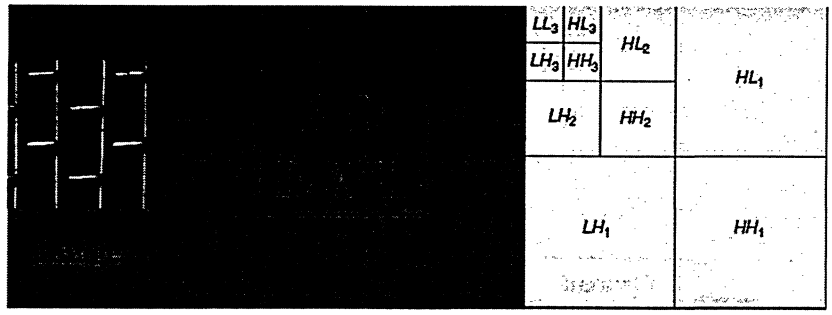

Figure 4.2: 3-Level Decomposition of Image Using DWT

4.2 Feature Extraction

The Laplacian mixture model (LMM) developed in the previous chapter for modelling the shape of the wavelet coefficient's distributions is used for the feature extraction. It is observed that the model parameters are a good representation of the texture information of the images. The texture information is carried by only a few coefficients in the wavelet domain where the edges occur in the original images. The parameters of the model thus obtained describe the distribution of the wavelet coefficients. They are also a good representation of the texture information of the images.

4.2.1 Wavelet Decomposition

The images are first decomposed using the 2-D discrete wavelet transform. The 2-D discrete wavelet transform decomposes the images into 4 subbands representing the horizontal, vertical, diagonal information and a scaled down low resolution approximation of the original image. A 3-level decomposition of an image is illustrated in Figure 4.2.

4.2.2 Modelling Wavelet Coefficient Distribution

We model the wavelet coefficients in each wavelet subband as a mixture of two Laplacians centered at 0 ($\mu = 0$) :

$$p(w_i) = \alpha_1 p_1(w_i|b_1) + \alpha_2 p_2(w_i|b_2) \tag{4.1}$$

$$\alpha_1 + \alpha_2 = 1 \tag{4.2}$$

where α_1 and α_2 are the a priori probabilities of the two classes. The Laplacian component corresponding to the class of small coefficients has relatively small value of parameter b_1. The Laplacian components in the above equations are defined as under:

$$p_1(w_i|b_1) = \frac{1}{2b_1}\exp(-\frac{|w_i|}{b_1}) \tag{4.3}$$

$$p_2(w_i|b_2) = \frac{1}{2b_2}\exp(-\frac{|w_i|}{b_2}) \tag{4.4}$$

The shape of the Laplacian distribution is determined by the single parameter b. The value of this parameter is a good representation of the image texture in the spatial domain. The next step is the estimation of the parameters for this statistical model.

4.2.3 Estimation of Model Parameters

We apply EM algorithm to estimate the parameters of the model. The detailed explanation of the EM algorithm can be found in Chapter 2. The two steps in the EM algorithm for a Laplacian mixture of two components are as under:

E-Step:

The E-step for the n-th iterative cycle is defined as:

$$p_{1i}(n) = \frac{\alpha_1(n)p(w_i|b_1(n))}{\alpha_1(n)p(w_i|b_1(n)) + \alpha_2(n)p(w_i|b_2(n))} \tag{4.5}$$

$$p_{2i}(n) = \frac{\alpha_2(n)p(w_i|b_2(n))}{\alpha_1(n)p(w_i|b_1(n)) + \alpha_2(n)p(w_i|b_2(n))} \tag{4.6}$$

M-Step:

The M-step for the n-th iterative cycle is defined as:

$$\alpha_1(n+1) = \frac{1}{K}\sum_{i=1}^{K} p_{1i}(n) \tag{4.7}$$

$$\alpha_2(n+1) = \frac{1}{K}\sum_{i=1}^{K} p_{2i}(n) \tag{4.8}$$

$$b_1(n+1) = \frac{\sum_{i=1}^{K} |w_i| p_{1i}(n)}{K\alpha_1(n+1)} \tag{4.9}$$

$$b_2(n+1) = \frac{\sum_{i=1}^{K} |w_i| p_{2i}(n)}{K\alpha_2(n+1)} \tag{4.10}$$

Where K is the number of coefficients in the subband. All the images in the database are decomposed by 2-D DWT. The wavelet coefficients in the LH, HL and HH subbands are modelled by two component Laplacian mixture. The parameters $[\alpha_1, \alpha_2, b_1, b_2]$ of the model are obtained from the EM algorithm.

4.2.4 Feature Selection

The relevancy of a feature for defining the texture content of the images is decided on the following basis.

- There should be very small variation in the value of the feature for similar texture.
- The values of the feature for different textures should be significantly different.

It has been observed that the parameters p_1 and p_2 possess very small variance. Their contribution towards the texture contents of the image is very small. Therefore these two parameters are not used for the indexing purpose. The other two parameters b_1 and b_2 are chosen as features at each level of decomposition.

4.3 Feature Vector Normalization

The value of each component in the feature vector has different dynamic range because it represents a different physical quantity. The features having higher values will overshadow the features with lower value in similarity calculation. Therefore the features are normalized before the application of the similarity measure. The normalization process emphasizes each component of the feature vector equally [92].

Let V be the sequence of values to be normalized. One way of normalization to $[0, 1]$ range is to find the maximum and minimum value for the sequence. Then the sequence can be normalized by applying the equation:

$$V_i = \frac{V_i - V_{min}}{V_{max} - V_{min}} \tag{4.11}$$

Where V_{min} and V_{max} are the minimum and the maximum values of the sequence. This normalization procedure is very simple but it does not provide desirable results. For example, let us consider a sequence of values $[1.4, 1.8, 2.7, 2.3, 200]$. By using the above normalization, most of the $[0, 1]$ range will be taken by a single quantity of 200. The other values $[1.4, 1.8, 2.7, 2.3]$ will be wrapped with a very small range. A better way to normalize the sequence is to consider it being generated by a Gaussian distribution. In this procedure, we calculate the mean μ and standard deviation σ of the sequence. The sequence is then normalized as follows:

$$V_i = \frac{V_i - \mu}{\sigma} \tag{4.12}$$

This will map most of the values of the feature V in the range $[-1, 1]$. The advantage of this normalization is that a few abnormal values occurring in the sequence will not bias the importance of other values. Therefore, this normalization technique has been used here.

4.4 Similarity Measures

The similarity criterion is very critical for ranking the images according to their relevancy to the query image. The metric distance between the feature vectors of the query and the test image is commonly used for similarity measurement. The Minkowski-form distance is defined based on the Lp norm:

$$d_p(\mathbf{Q}, \mathbf{T}) = \left(\sum_{i=0}^{N-1} |Q_i - T_i|^p\right)^{\frac{1}{p}} \tag{4.13}$$

where $\mathbf{Q}$ and $\mathbf{T}$ are vectors of dimension N. The above equation is the general form of the distance metric. If $p = 1$, then the distance is known as the *City-block* or *Manhattan* distance:

$$d_1(\mathbf{Q}, \mathbf{T}) = \sum_{i=0}^{N-1} |Q_i - T_i| \tag{4.14}$$

Another famous distance metric is *Euclidean* or L_2 norm defined when $p = 2$:

$$d_2(\mathbf{Q}, \mathbf{T}) = \left(\sum_{i=0}^{N-1} |Q_i - T_i|^2\right)^{\frac{1}{2}} \tag{4.15}$$

Euclidean and *City-block* distance measure only the difference in the lengths of the two vectors. In some cases, the angle between the vectors may be more significant for purpose of similarity. The cosine distance measures the difference in the direction of two vectors irrespective of their length. The *cosine distance* is defined as:

$$d_{\cos}(\mathbf{Q},\mathbf{T}) = 1 - \cos(\theta) = 1 - \frac{\sum_{i=0}^{N-1}(Q_i.T_i)}{\sum_{i=0}^{N-1} Q_i^2 . \sum_{i=0}^{N-1} T_i^2} \tag{4.16}$$

We can see that this is very similar to the correlation coefficient.

χ^2 *Statistic* measure is based on the chi-square test of equality for two sets of frequencies and is defined as:

$$d_{\chi^2}(\mathbf{Q},\mathbf{T}) = 1 - \sum_{i=0}^{N-1} \frac{(Q_i - m_i)^2}{m_i} \tag{4.17}$$

where

$$m_i = \frac{Q_i + T_i}{2} \tag{4.18}$$

The histogram intersection is used for comparing the two histograms. This distance measure between the two histograms was proposed by Swain and Ballard [93]. Their objective was to find known objects within images using color histograms. This is defined as:

$$d_{hist}(\mathbf{Q},\mathbf{T}) = 1 - \sum_{i=0}^{N-1} \frac{min(Q_i, T_i)}{|\mathbf{Q}|} \tag{4.19}$$

Although the above equation defines the degree of similarity between two histograms, yet it is not a metric distance. It can be extended into a metric distance using the formulation:

$$d_{hist}(\mathbf{Q},\mathbf{T}) = 1 - \sum_{i=0}^{N-1} \frac{min(Q_i, T_i)}{min(|\mathbf{Q}|, |\mathbf{T}|)} \tag{4.20}$$

In the following experimental evaluation, we have used the *city-block* distance measure during the initial search. The selection is based on the empirical analysis. *city-block* performed better in terms of ranking of the images according to their perceptual similarity.

4.5 Relevance Feedback

The discriminatory power of the features is highly important for an effective image retrieval system. However, it is very difficult to model the human visual perception

by only a set of features. Also the similarity between the images is a very subjective notion. The visual content of the images may be interpreted differently by different individuals. Some individuals consider color more important while the others may perceive shape as a more relevant characteristic. Even for the same visual descriptor, the human perception is quite varying. The objective of an efficient CBIR system is to model the human visual system. This served as a motivation for the idea of relevance feedback. Relevance feedback is a mechanism of learning from the user interaction. The system parameters are changed depending on the feedback from the user. There may be a variety of ways in which the input from the user can be used. One approach is to modify the query by using the feedback information. Another common approach is to update the feature weights. This method is adopted in the MARS project [94]. The relevance feedback information may be used to construct new features on the fly. This method has been implemented in [95].

At first the images are ranked according to the relevancy to the query image by using a similarity criterion. These initial results are then presented to the user. The user then can mark the images as relevant of irrelevant. The system learns the users's notion of similarity from this positive and negative feedback. The procedure is iterative in nature. At each cycle, the user puts labels to the images and feedback to the system. This process is repeated until the user is satisfied with the results.

Let us suppose that the user presents a query image to the system; the search engine performs the search and K most relevant images displayed for the user. These K images retrieved as relevant in the initial search for the training set. This can be represented as: $T = (\mathbf{x}_k, y_k)_{k=1}^{K}$. Here $\mathbf{x}_k$ denotes the feature vector of the *k-th* image and y_k is the label of the image given by the user. This label have a value of 1 for the relevant images and 0 for the non-relevant. Once the user has labelled all the images presented, we can form two matrices consisting of the features from the two classes. The matrix containing the features of the images marked as relevant is denoted by $R = [x_{m,i}]$ while the matrix containing the features of the non-relevant images is denoted by $D = [z_{n,i}]$. $x_{m,i}$ is the *i-th* component of the feature vector of *m-th* image in the relevant set and $z_{n,i}$ is the *i-th* component of the feature vector

of *n-th* image in the non relevant set. The dimension of R and D is $M \times P$ and $N \times P$ respectively. Where $m = 1, 2, \cdots, M$ and $n = 1, 2, \cdots, N$. P is the dimension of the feature vector, M and N are the number or relevant and non relevant images respectively.

4.5.1 Feature Weight Updating Scheme

The initial search cycle is performed by assigning equal weight to each of the components in the feature vector. However, we do not obtain satisfactory results because of the discrepancy between the system's criteria of similarity and the user's definition. In the feature vector updating scheme, the weights of the feature components are computed based on the importance of the features attached by the user. The density of a feature x_{mi} around q_i is related to the relevancy of the i-th feature. Here q_i refers to the i-th component of the feature vector of query image. A large density usually indicates high relevancy for a particular feature. While a low density implies that the corresponding feature is not relevant to the similarity characterization. This is measured in terms of the standard deviation σ_i of the features in the relevant set. The inverse of the standard deviation is then assigned as the weight for the particular feature [96].

$$B_i = \frac{1}{\sigma_i} \tag{4.21}$$

4.5.2 Query Modification

In the Query by Example (QBE) search paradigm, the user presents a sample image to the system. The initial query is formulated from this example image. Let us consider the scenario where the user cannot find a suitable sample image that represents the relevant class of images. Under this scenario the system is not able to perform satisfactorily. The query modification approach tries to modify the query based on the user feedback. This modified query is then used in the next iteration cycle. The query modification idea was first proposed by Salton for text retrieval [100].

There are different ways in which the query is modified. One way of modifying the query is based on the feature vectors of the images marked as relevant by the user. It

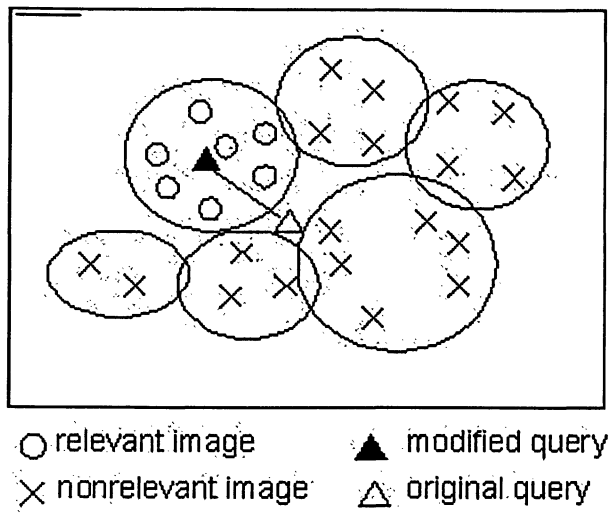

Figure 4.3: Query Modification Model 1

is assumed here that the training subset of relevant images is a good representation of the complete set of relevant images. Based on this assumption the modified query can be taken as the central value of the relevant sequence of feature vectors [96, 97].

$$\hat{\mathbf{q}} = \sum_{m=1}^{M} \mathbf{x_m} \tag{4.22}$$

where $\hat{\mathbf{q}}$ is the modified query and $\mathbf{x_m}$ is the feature vector of the m-th image in the relevant set. This query modification approach is illustrated in the Figure 4.3.

The above noted query modification procedure will not work in case the training set does not represent the complete set of relevant images. In that case, the modified query will not be around the center of the relevant class. The retrieval results by using this new query will contain many non relevant images. This is a very unfavorable situation. This may be rectified by using a large set of training images. However, in practice, it is not possible to increase the training set. Another way to solve this problem is to utilize the non relevant images in the training process. In this approach the new query is obtained by moving the original query towards the center of the relevant class and away from the non relevant class. Let $\mathbf{q}$ represent the feature vector of the original query then the modified query $\hat{\mathbf{q}}$ is defined as:

$$\hat{\mathbf{q}} = \bar{\mathbf{x}} - \alpha_N(\bar{\mathbf{z}} - \mathbf{q}) \tag{4.23}$$

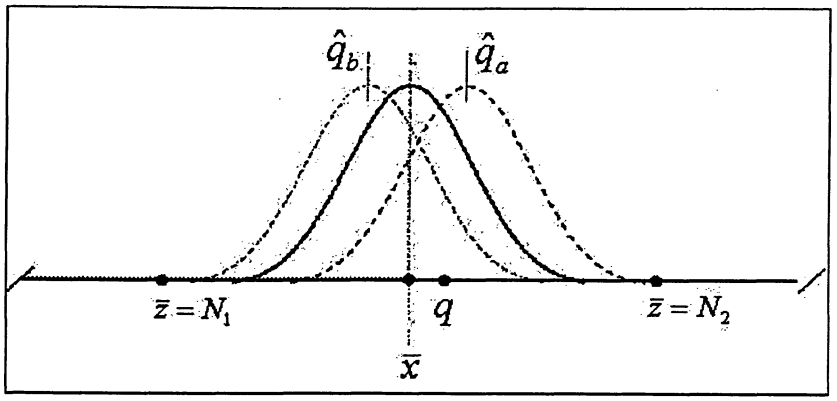

Figure 4.4: Query Modification Model 2

$$\hat{\mathbf{q}} = \bar{\mathbf{x}} - \alpha_N(\bar{\mathbf{z}} - \mathbf{q}) + \alpha_R(\bar{\mathbf{z}} - \mathbf{q}) \tag{4.24}$$

where

$$\bar{\mathbf{x}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x_m} \tag{4.25}$$

$$\bar{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z_n} \tag{4.26}$$

α_N and α_R are small positive constants and $\bar{\mathbf{x}}$ and $\bar{\mathbf{z}}$ are the mean feature vectors of relevant and non relevant image sets respectively. This approach for the query reformulation is shown in the Figure 4.4. The centroid of the relevant class is practically more important because the relevant images are closely clustered in the feature space. On the other hand, the non relevant class is very diverse. The features are distributed randomly over the whole feature space. Hence, the centroid of the non relevant class is not as significant as the centroid of the relevant class. The constant values α_R and α_N are chosen based on this fact. Generally the value of the constant α_R is chosen much higher than α_N.

4.5.3 Radial Basis Function

The traditional distance metrics such as *Euclidean* or *city-block* are restricted to the linear association between the distance and the similarity. The same magnitude of distance is always mapped to the same similarity value. However, the human visual system performs the pattern recognition and classification of visual content on a non-

linear basis. The same magnitude of distance is not mapped to the same similarity value. Therefore, to model the human notion of similarity between the images, a non-linear approach should be adopted. This is achieved by measuring the similarity by a network of Radial basis functions. Radial basis function (RBF) is a kernel function that has excellent non-linear approximation capability [98]. The main property of the radial functions is their response decreases (or increases) monotonically with distance from a central point. A typical example of the radial function is the Gaussian which is defined as under for a scalar input:

$$F(x) = \exp[-\frac{(x-\mu)^2}{2\sigma^2}] \tag{4.27}$$

where μ is the center of the RBF and σ is its width. The Gaussian RBF centered at 0 ($\mu = 0$) with $\sigma = 1.1$ is shown in Figure 3.7.

Gaussian RBF gives a solution to the regularization problem in function estimation subject to a certain smoothness criteria. We employ this property of the Gaussian kernel to approximate the similarity measure. A one dimensional Gaussian RBF is associated with each component of the feature vector is given as [96]:

$$F(\mathbf{x}, \mathbf{q}) = \sum_{i=1}^{P} G_i(x_i - q_i) = \sum_{i=1}^{P} \exp[-\frac{(x_i - q_i)^2}{2\sigma_i^2}] \tag{4.28}$$

where $\sigma_i = 1, ..., P$ are the tuning parameters in the form of RBF widths. The center of the curve is at the query location q_i. The magnitude of the function $F(\mathbf{x}, \mathbf{q})$ represents the similarity between the input vector $\mathbf{x}$ and the query $\mathbf{q}$. The maximum similarity is attained when $\mathbf{x} = \mathbf{q}$. The tuning parameters σ_i reflect the relevance of the *i-th* feature towards the similarity measurement. RBF similarity measurement is illustrated in the Figure 4.5. These tuning parameters σ_i are very critical to the performance of the RBF network. The density of a feature x_{mi} around q_i is related to the relevancy of the i-th feature. This is inversely proportional to the length of the interval [101]. A large density usually indicates high relevancy for a particular feature, while a low density implies that the corresponding feature is not critical to the similarity characterization. Hence the tuning parameters are estimated as:

$$\sigma_i = \eta \max_m |x_{mi} - q_i| \tag{4.29}$$

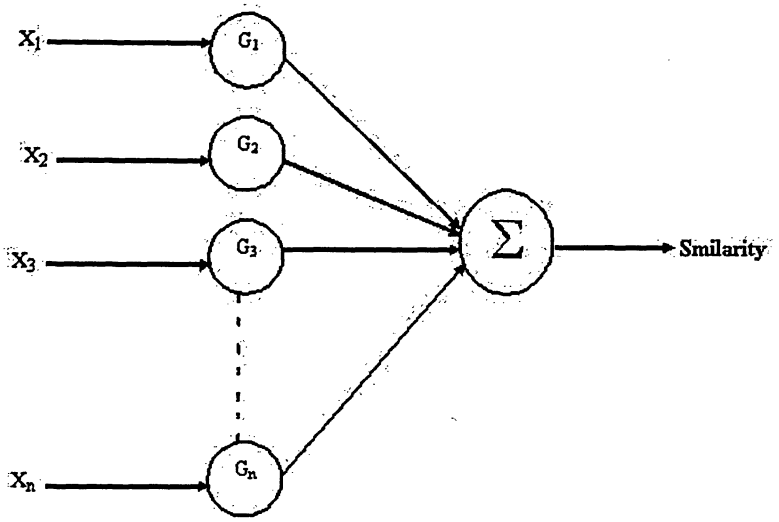

Figure 4.5: Similarity using RBF

where x_{mi} is the *i-th* component of the feature vector of *m-th* relevant image and q_i is the *i-th* component of the feature vector of the query image. The additional factor η is introduced to ensure reasonably high output for the RBF units.

4.6 Experimental Results

4.6.1 Database Description

The proposed approach is tested using the standard Brodatz texture image database. Brodatz image database contains 1856 images divided into 116 classes. Every class has 16 images. Figure 4.6 shows all the texture classes in this database. In some classes the texture remains quite homogenous while in others its non homogenous. Examples from both types of texture classes are given in Figure 4.7

4.6.2 Performance Metrics

The following two performance metrics are used for the comparison and system evaluation. Let the total number of relevant images in the database be T. When the user presents a query to the system, M images are displayed from the top of the

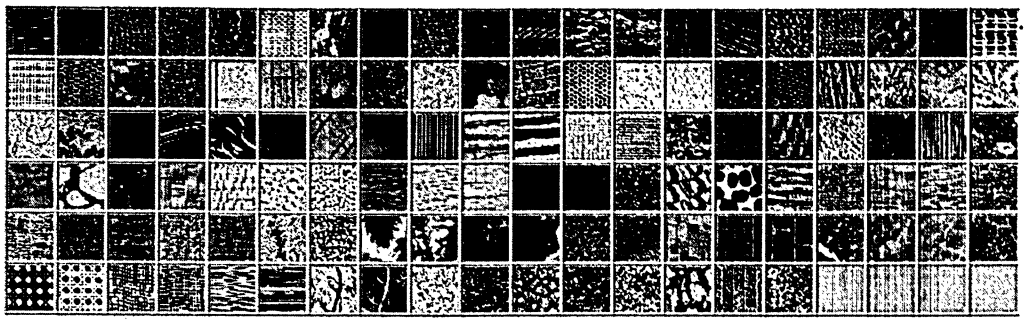

Figure 4.6: 116 Texture Classes in Brodatz Image Database

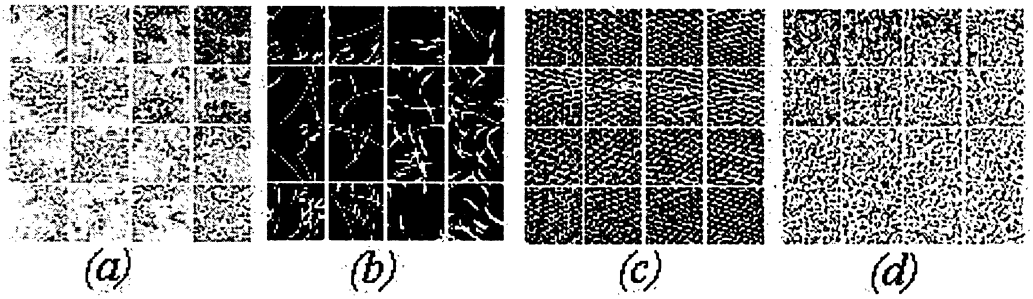

Figure 4.7: *a-b* Non-homogenous Texture; *c-d* Homogenous Texture

ranked list. Out of these M images, R is the number of relevant images while N is the number of irrelevant ones.

$$Recall = (R/M)100 \tag{4.30}$$

$$Precision = (R/T)100 \tag{4.31}$$

In our evaluation, we use a graphical user interface to display 16 top ranked images. The database used in the experimental evaluation contains 16 images in each class. That means both of the above stated measures will have the same value for this particular case. However, the number T is generally larger than R.

4.6.3 Feature Weight Updating

In Table 4.1, we have summarized the results obtained by using the feature weighting scheme in the feedback loop. These results also show the comparison of performance for Daubechies 1(db1) and Daubechies 2 (db2) wavelet kernels. The images are decomposed upto 3-level using db1 or db2 wavelet filters. The wavelet coefficients in

each of the high frequency subbands are modelled as a mixture of two Laplacians. The parameters of the model are used as the features. In case of 3-level decomposition of images, the mean and standard deviation of the wavelet coefficients in the approximate subband and variances of the two Laplacians in each of the 9 high frequency subbands are taken as features. Therefore, the feature vector is 20-dimensional.

Initial search is performed giving equal weights to all the component of the feature vector using *city-block* distance measure. It is observed that the db2 wavelets perform better than db1. The initial recall rate of 67.4 per cent is obtained using db2 wavelet filter which is 3.6 per cent higher than in case of db1. The recall rate is significantly improved by updating the feature vector during feedback. The graphical user interface is implemented to provide easy interaction with the system. The user marks the relevant images using this graphical user interface. The similarity is measured using the city-block distance measure. The most of the performance enhancement is achieved after the first iteration when the recall rate is improved from 63.8 per cent to 71.8 per cent in case of db1 wavelet filters. The recall rate is increased from 67.4 per cent to 74.6 per cent in case of db2 Wavelets. A slight improvement is achieved after the second iteration, however, the system attains stability after two iterations. The performance of the proposed approach is compared with the Wavelet moments method. In this method, the images are decomposed upto 3 levels and mean and standard deviation of the Wavelet coefficients in each subband are used as features [96]. The results indicate that the initial recall rate is 23 per cent higher than the Wavelet moments method. After 4-th iteration, the proposed approach performs 4.2 per cent better than the Wavelet moments method. Figure 4.8 depicts the performance of the system versus the iteration numbers.

Method	Iteration 0	Iteration 1	Iteration 2	Iteration 3
LMM (db1)	63.8	71.8	72.6	72.6
LMM (db2)	67.4	74.6	75.4	75.5
Wavelet Moments	44.4	68.5	70.9	71.3

Table 4.1: Average recall rate (in percentage) for 1856 query images using feature weighting approach (20 features)

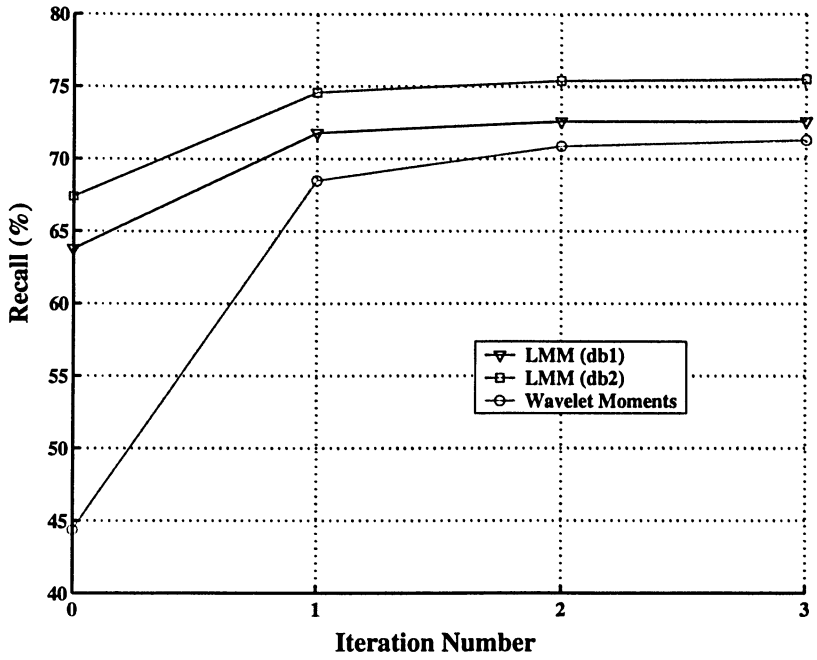

Figure 4.8: Retrieval Performance

4.6.4 Radial Basis Function Method

Table 4.2 summarizes the results obtained from the non-linear RBF approach for relevance feedback. Here the images has been decomposed using db2 wavelet kernel. The three query modification methods discussed in Section 4.5.2 are tested for performance. The initial search is performed using the city-block distance measure.

1. In the RBF1 method, we use the query modification model defined in Equation 4.22. This utilizes only the positive feedback provided by the user in the form of images labelled as relevant. The effect of this query modification approach is moving the query towards the center of the relevant class with each iteration.

2. In RBF2 method, we use both the positive and negative samples for training the system. The RBF2 model uses the query modification approach given in Equation 4.23. The query is moved towards the center of the relevant class and away from the non-relevant class with each iteration.

3. In RBF3 method, we use both the positive and negative samples for training

the system. The RBF3 model uses the query modification approach given in Equation 4.24. The query is moved towards the center of the relevant class and away from the non-relevant class with each iteration.

Most of the improvement is achieved after the 1st and 2nd iterations. We observe a slight improvement after the 3rd iteration. However it is observed that the improvement is not very significant for the 3rd iteration. A significant improvement in the retrieval efficiency is observed by employing non-linear RBF similarity measure. A 7.5 per cent increase is obtained by using RBF3 similarity measure compared to city-block distance. The retrieval accuracy of The MARS system is also given for evaluation purposes. The similarity measure in MARS case is cosine distance. The parameters values chosen in the experiment are $[\alpha = 1, \beta = 4, \gamma = 0.8]$. These values were determined empirically. The value of η is adjusted at 2.5. RBF3 method out-

Method	Iteration 0	Iteration 1	Iteration 2	Iteration 3
RBF1	67.40	79.20	81.50	81.91
RBF2	67.40	79.40	81.62	82.25
RBF3	67.40	78.60	81.70	83.00
MARS	67.10	75.12	76.42	76.95
Wavelet Moments	44.10	72.4	77.42	78.36

Table 4.2: Average recall rate (in percentage) for 1856 query images using RBF method (20 features)

performs the RBF1 and RBF2 methods. The best performance was achieved with the parameter values $[\alpha_R = 2, \alpha_N = 0.2, \eta = 2]$, that were determined empirically. The performance comparison of the three RBF methods with MARS and Wavelet moments (WM) method is shown in the Figure 4.9. The Wavelet moments method has been implemented with RBF1 query modification approach [96]. It is clear that the proposed scheme performs better than MARS and Wavelet moments method. Figures 4.10, 4.11, 4.12, 4.13 and 4.14 show the performance of the system for a few example queries:

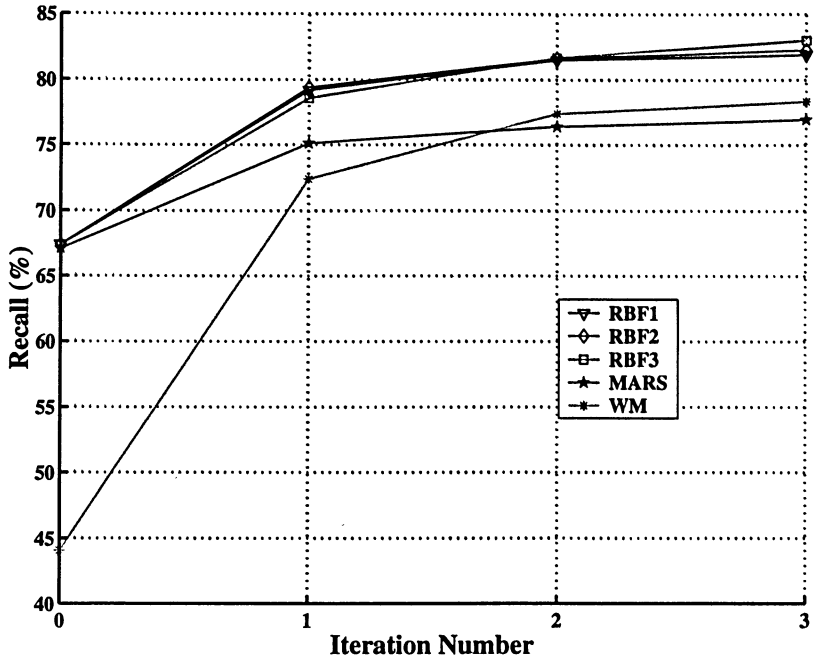

Figure 4.9: Performance Comparison

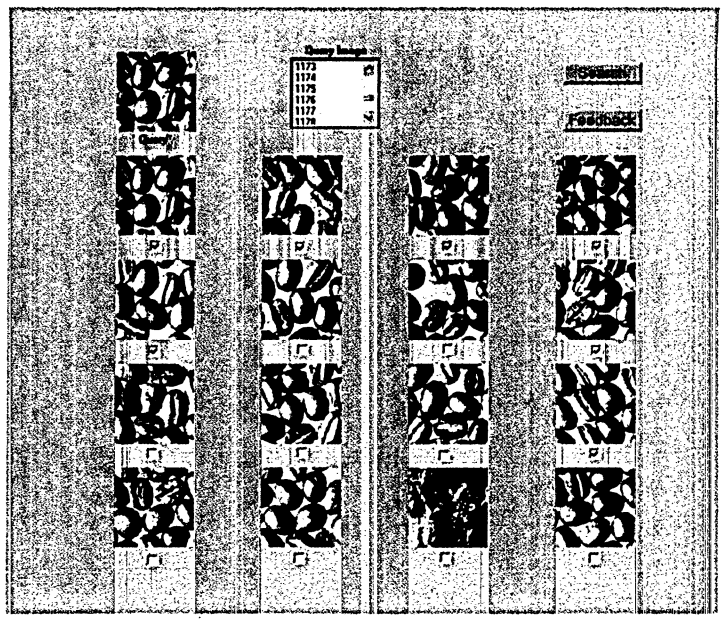

Figure 4.10: Retrieval Results for Query 1

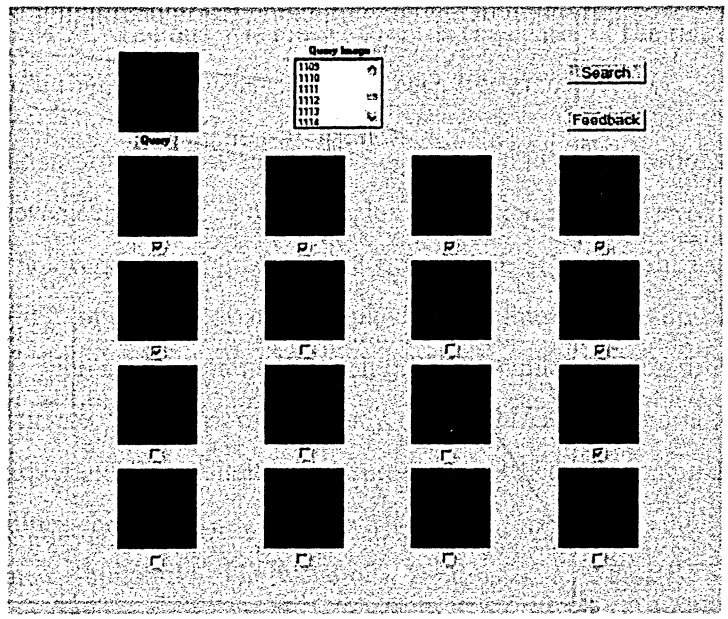

Figure 4.11: Retrieval Results for Query 2

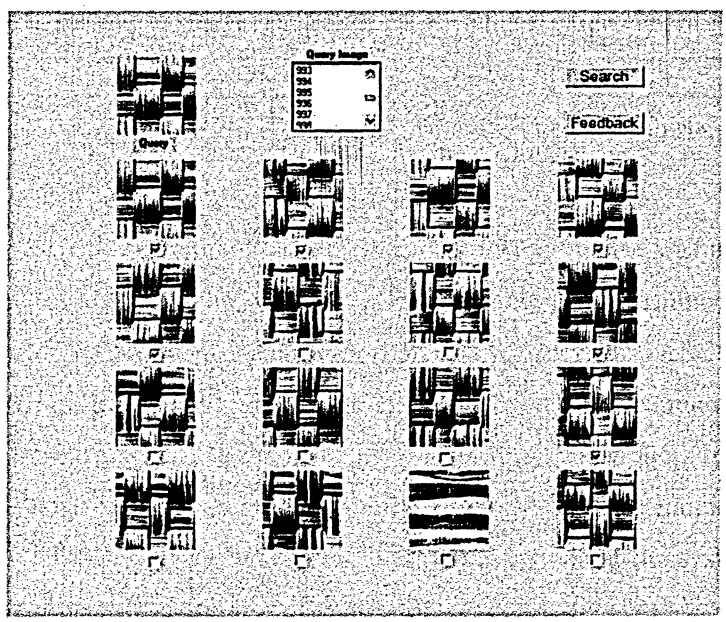

Figure 4.12: Retrieval Results for Query 3

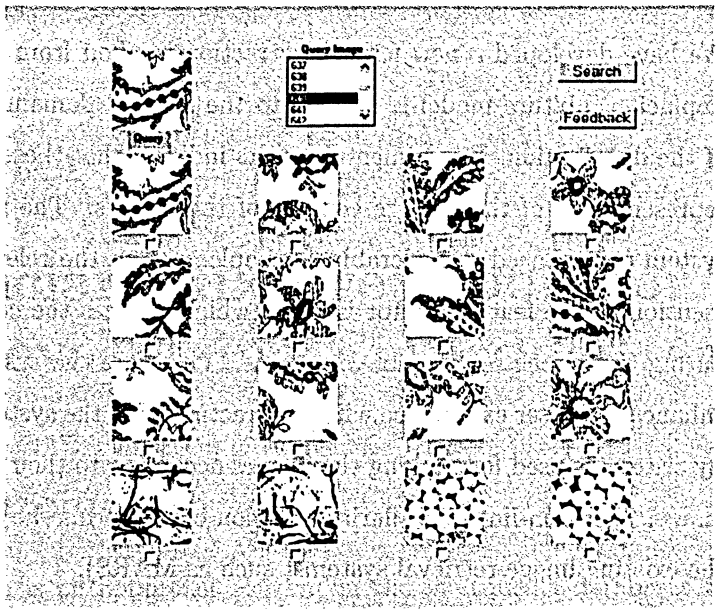

Figure 4.13: Retrieval Results for Query 4

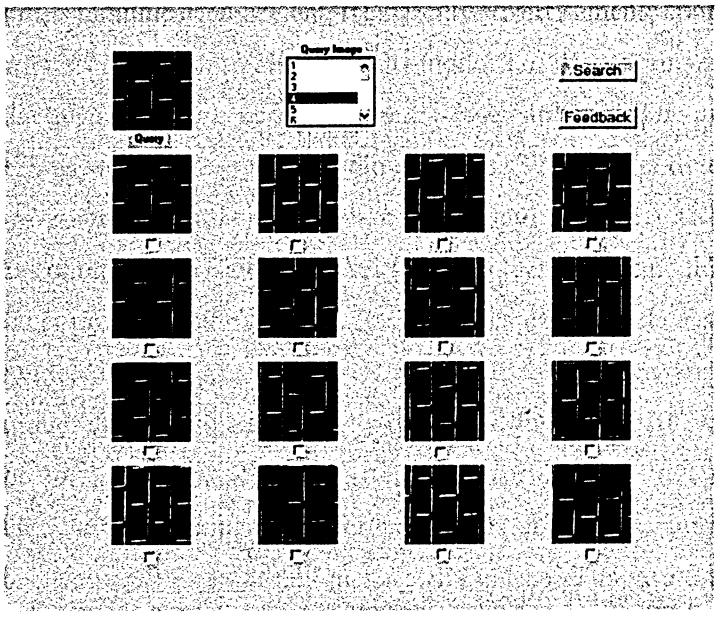

Figure 4.14: Retrieval Results for Query 5

4.7 Conclusions

We have developed a new feature extraction method from the texture images. The Laplacian mixture model is applied to the wavelet domain to catch the peaky-ness of the distribution. Experimental results indicate that these new features are a good representation of the texture content of the images. The retrieval efficiency of the system is enhanced considerably by implementing the relevance feedback. The dimension of the feature vector is small which reduces the computational complexity during the search and feedback process. Therefore the system response is fast and enhances the user experience while interacting with the system. A non-linear similarity criteria is used for ranking the images according to their relevancy with the query image. This non-linear similarity criterion outperforms the other methods as well as the existing image retrieval systems(such as MARS).

Chapter 5

Laplacian Mixture Model for Video Retrieval Using Embedded Audio

5.1 Introduction

Digital video is an important element in the multimedia databases that continues to witness phenomenal growth in recent years. Due to the exponential growth in the computing power and storage capacities, the use of the digital video is gaining popularity in various areas. The most common application areas where video is an effective way of communication are entertainment, advertisement, sports, education, surveillance and security etc. Video is multi-modal i.e. it contains information in different media such as visual, audio and text (close caption). The audio in videos contains speech as well as non-speech audio contents. The textual description of the video data by human analyst is even more complex and subjective than the images. Therefore, the content-based video retrieval (CBVR) techniques should be developed to automate this process. Efficient CBVR techniques can provide easy and flexible access to the video data.

The techniques and principles primarily developed for the CBIR can be extended for application to the videos. But this extension of principles is not a straight forward task. There are some inherent differences between the images and videos that require the development of new techniques for the videos. The visual information in the digital videos is a sequence of frames. These frames may be taken as images for application of CBIR methods. But the video contains a huge amount of data. Index-

ing every frame is not feasible in practice. Moreover, video is a structured sequence of frames. The arrangement of these frames in time conveys high level concepts. Video data also contains audio and text information which should also be taken into account while building meaningful indexes. The indexing of digital video usually consist of three main steps; video parsing, abstraction and content analysis [94]. A brief description of these three steps is given below:

5.1.1 Video Parsing

Video contains huge volume of data that is very difficult to handle and process in an efficient way. Therefore, the large video files are divided into smaller units as a first step to the indexing process. There are two approaches towards modelling the video content. One approach focusses on the physical division of longer video files into smaller structural units. This is known as the segmentation process. In the other approach; known as the stratification, the contextual information is segmented into chunks of data.

Segmentation

Successful video segmentation is necessary for most multimedia applications. Video segmentation is the process of dividing a sequence of frames into smaller meaningful units called shots. A video shot is defined as a sequence of frames recorded contiguously and representing a continuous action in space and time. A video scene on the other hand carry some high level concept and is a collection of two or more shots. Figure 5.1 illustrates the segmentation process. Shots are the building block of a video. Shot boundary detection has been approached by several studies from a variety of perspectives including techniques that are pixel based, statistics based, transform based, feature based and histogram based. It is widely recognized that histogram based and feature based approaches offer the best solution to the problem.

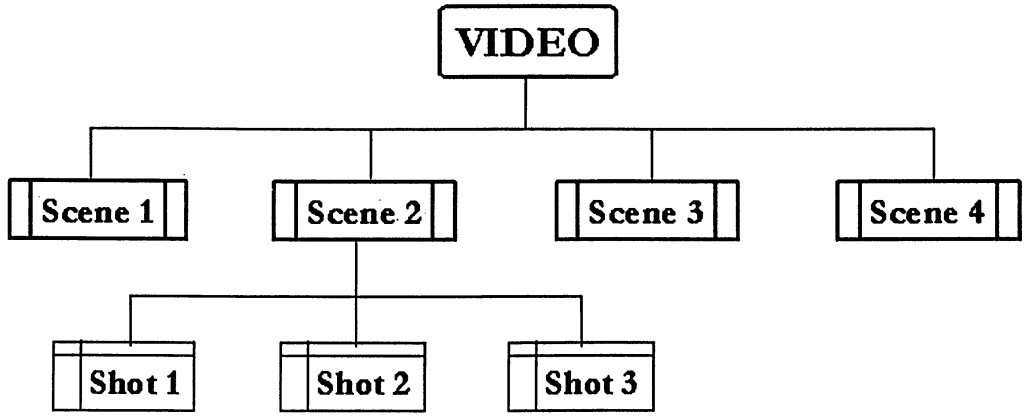

Figure 5.1: Video Segmentation

Stratification

Stratification is a context-based approach to model the video contents as strata. Strata may overlap and thus any frame can have a variable number of strata associated with it. Multi-layered descriptions are attached to the video data [102]. A comparison between the traditional segmentation model and the stratification model is given below.

- The basic atomic unit in the segmentation model is a shot. Therefore the granularity of the video data is limited to the shots. It is not possible for the users to access the data within the shot boundaries. However, in the stratification approach the granularity of information is a frame. The access and presentation of the video contents is thus very flexible from the user point of view.

- The segmentation model imposes authors' intentions early during the shot creation stage. Thus, once the video is segmented, it is difficult to support other users who may need to use the same material for a different purpose. The stratification model views video contents as layers of overlapping strata. Thus users may combine strata to flexibly retrieve video, or easily build additional strata to cater to their specific needs.

Figure 5.2 shows the Stratification modelling of a news video.

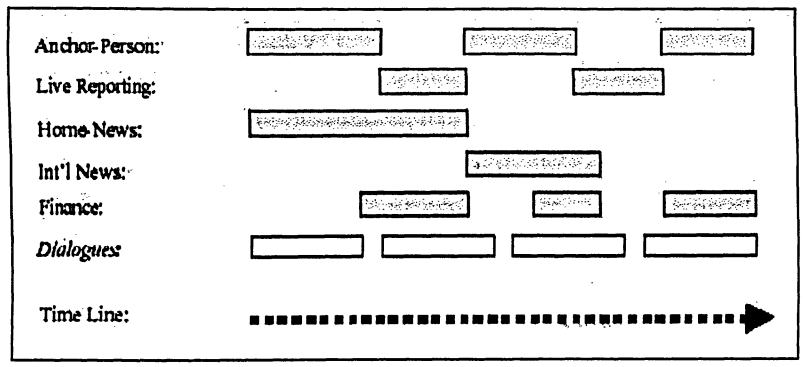

Figure 5.2: Video Stratification

5.1.2 Abstraction

The video abstraction is the process of extracting an economical representation of the visual information. This facilitates the browsing of the video content. The most common ways of abstraction are key-frame and highlight sequence.

Key-Frame

The most representative frame in a video shot/scene is called the key-frame. The simplest method is to take the first, last or the middle frame as the key-frame. More complex techniques are being evolved to extract the key-frames based on the motion analysis, shot activity etc.

Highlight Sequences

The production of a shorter sequence of frames representing a larger video clip is known as highlight generation. This approach is also known as video summarization. The utilization of multiple sources such as shot boundaries, human faces and text is essential for building an efficient highlight sequence. Video summaries enable the effective browsing of video data. The process of generating highlights for the sports, movie or news video can be done automatically by the summarization algorithms.

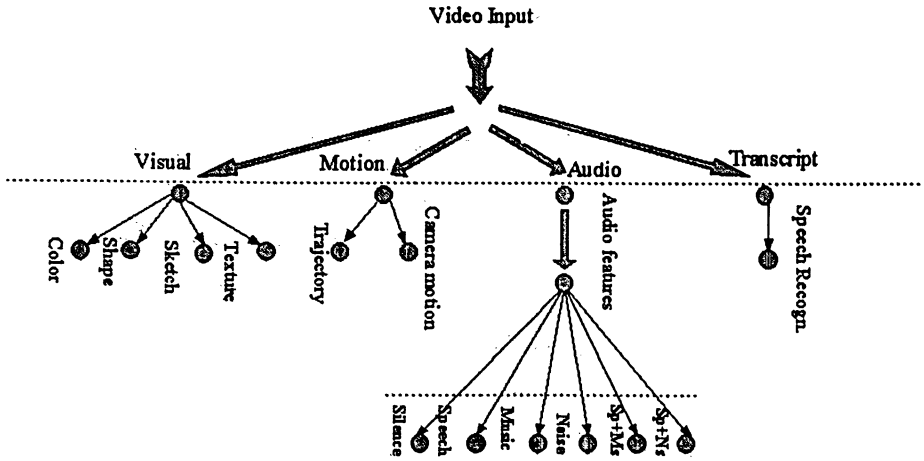

Figure 5.3: Multi-modality of Video Data

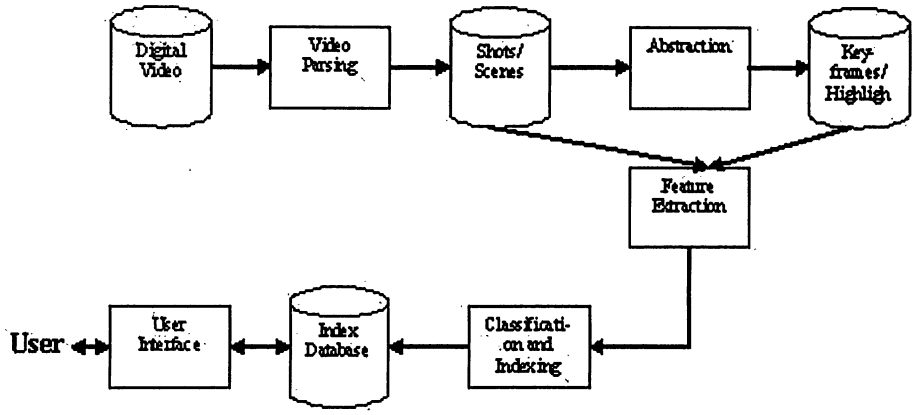

Figure 5.4: Block Diagram of a Content-Based Video Retrieval System [94]

5.1.3 Content Analysis

This step analyzes the video contents to generate indexes for retrieval and content-based access to the video data. CBIR techniques can be applied to the representative frames. In order to produce effective indices all possible content information such as visual, audio and text should be employed. A general overview of the contents of video data is shown in Figure 5.3. The block diagram of a content-based video system is given in Figure 5.4.

5.2 Video Indexing using Embedded Audio

Traditionally, visual information is used for video indexing. However, the users of the video data are often interested in certain action sequences that are easier to identify in the audio domain. While visual information may not yield useful indexes in this scenario, the audio information often reflects what is happening in the scenes. The audio is a very rich source of information that can distinguish these actions. Although visual information is very popular for indexing the video data, yet a few researchers have used audio information for this purpose. The existing techniques to classify audio or the embedded audio may not be very useful for application to the news video. In this type of application, music, speech, noise and crowd voice may be found together in the same video clip. Therefore, we need features that represent the global similarity of the audio content. A statistical approach has been adopted here to analyze the audio data and extract the features for video indexing.

5.2.1 Wavelet Decomposition

The proposed indexing scheme does not depend on the segmentation method. The video can be segmented into shots using any algorithm. Here, we use manual segmentation of the video data. The next step to manual segmentation is separating the embedded audio by demultiplexing video shots. The audio signals are then resampled to a uniform sampling rate. Each audio segment is decomposed using a one-dimensional DWT. We have used different wavelet kernels for decomposition of the audio signals. This is done to compare the performance of different wavelets in describing the audio contents. The one-dimensional DWT decomposes the signal into 2 subbands at each wavelet scale; a low frequency subband and a high frequency subband. Audio signal is much different from gray or color image signal. In images, the values of adjacent pixels usually don't change sharply. On the other hand digital audio signal is a form of oscillating waveform, which includes a variety of frequency components varying with time. The LL subband of the image is an icon of the original image. However, most audio signals consist of a wide variety of frequencies that

produce a complex waveform. The wavelet coefficients of audio signal have many large values in detail levels, and the LL subband coefficients don't always provide good approximation of the original signal.

The range of audible frequencies, or the sound frequency spectrum, is divided into sections, each having a unique vital quality. The divisions are usually referred to as octaves. An octave is a logarithmic relation in frequency and is defined as the interval between any two frequencies that have a relation of 2 : 1. The range of human hearing covers about 10 octaves. Starting with 20 Hz, the first octave is 20 Hz - 40 Hz; the second, 40 Hz- 80 Hz; the third, 80 - 160 Hz; and so on, up to 20480 Hz. Wavelet decomposed audio signals highly resemble to the octave-band decomposition of audio signals. The ratio of the size of the wavelet sub-bands is also 2 : 1, i.e. the number of coefficients of a subband at decomposition level i is exactly half the number of coefficients of a subband at level *i-1*. The wavelet decomposition scheme matches the models of sound octave-division for perceptual scales. Wavelet transform also provides a multi-scale representation of sound information, so that we can build indexing structure based on this scale property. These properties of wavelet transform for sound signal decomposition is the foundation of audio retrieval and indexing system developed in this chapter. The decomposition is taken up to 7 levels. We experiment with different levels of decomposition. An increase in the level of decomposition also increases the number of features extracted for indexing. This improves the retrieval performance at the expense of more computational overhead.

5.2.2 Feature Extraction

The statistical model based on the Laplacian mixture of two components developed for the texture retrieval in Section 4.2.2 and 4.2.3 is applied for feature extraction from the embedded audio. It has been noted that parameters of the model are a good representation of the audio segments and define the global characteristics of the audio. The shape of the Laplacian distribution is determined by the single parameter b. The value of this parameter is a good representation of the contents of the embedded audio clip. The parameters for this statistical model are estimated using EM algorithm. The

Following components form the feature vector used for indexing the video clips:

- Mean of the wavelet coefficients in the Low frequency subband
- Variance of the wavelet coefficients in the Low frequency subband
- Model parameters $[P_l, b_l, bs]$ calculated for each of the high frequency subband

The feature vector is 17-dimensional in case of 5-level wavelet decomposition of the audio clips and 23-dimensional in case of 7-level decomposition. The normalization of the feature vector is required to put equal emphasis to each of the component of the feature vector. The components of the feature vector represent different physical quantities so their values have different dynamic range. Gaussian normalization procedure discussed in Section 4.3 is employed to convert the dynamic range of the component feature to $[-1, 1]$.

5.3 Similarity Measure

After the normalization of the component features in the feature vector, the Euclidean distance measure or L2-norm is used in the initial search.

$$d(\mathbf{x}, \mathbf{q}) = \sum_{i=1}^{N} B_i \sqrt{(x_i - q_i)^2} \tag{5.1}$$

where $\mathbf{x}$ and $\mathbf{q}$ are the feature vectors of the query image and the test image respectively. The weighting factor B_i is used to put different emphasis on the component of the feature vector depending upon its perceptual importance. These values are updated in the relevance feedback process provided by the users of the system. This is discussed in the next section.

5.4 Relevance Feedback

The initial search cycle is performed by assigning equal weight to each of the components in the feature vector. However, some features are more important to the human auditory perception than others. The performance of the system can be enhanced significantly by putting more emphasis on perceptually relevant features. This is

achieved by updating the weights of the feature components B_i during the feedback cycle as defined by equation 4.21.

5.5 Experimental Results

5.5.1 Database Description

The database used in these experiments consists of 302 video clips from a Cable News Network (CNN) video. The duration of the clips is around 3 second. These clips are produced by the manual segmentation of the news video. The database contains a heterogeneous mixture of clips with regard to the embedded audio. The video clips are classified into five different classes based on the audio contents. These five categories are representative of a typical news broadcast. The database contains 47 shots in *Male Anchor* class, 79 in *Male Reporter* class, 63 in *Female Reporter* class, 14 in *Noise* class and 97 in *Commercials* class.

5.5.2 Performance Metrics

For the Performance evaluation of the system, the retrieval ratio is calculated for each of the class as well as the overall ratio for the whole database. The graphical user interface is provided for display of the retrieval results and obtaining feedback from the user. A set of 16 most relevant video clips is presented to the user after each search cycle. The video clips are represented by their respective key-frames. The first frame of the video clips is chosen as the key frame. The recall for each class is calculated as:

$$Recall = (R/M)100 \tag{5.2}$$

where R is the number of relevant video clips in the audio domain and M are the total number of clips output by the system. The overall retrieval ratio of the system is the average of the individual recalls in each of the 5 classes.

$$Overall\ Recall = \frac{1}{C}\sum_{i=1}^{C}[(R/M)_i 100] \tag{5.3}$$

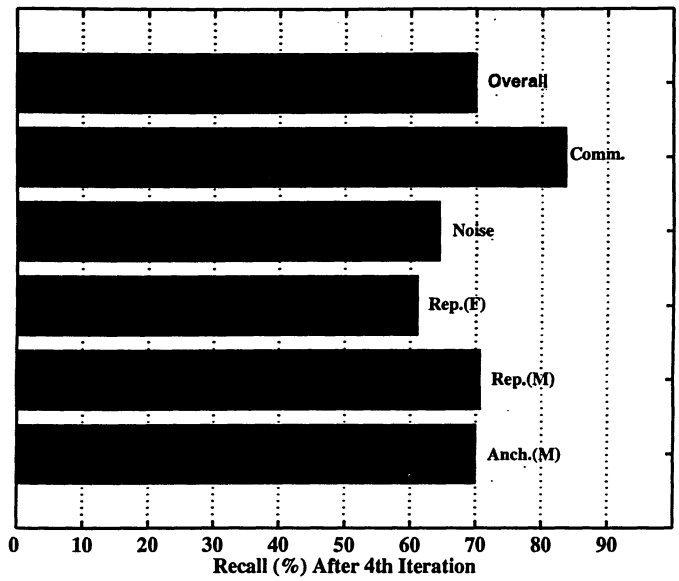

Figure 5.5: Retrieval Performance in 5 Classes (5-level Decomposition using db2)

5.5.3 Summary of Results

The results obtained by performing a 5-level decomposition of the embedded audio clips using Daubechies-2 (db2) wavelet kernel are summarized in Table 5.1.

Category/Iteration	Initial	Iteration 1	Iteration 2	Iteration 3	Iteration 4
Anchor(Male)	52.2	67.2	69.7	70.0	70.0
Reporter(Male)	60.2	68.5	70.5	70.6	70.7
Reporter(Female)	50.4	59.1	60.4	61.1	61.1
Noise	61.5	64.4	64.4	64.4	64.4
Commercials	71.2	81.7	83.3	83.6	83.7
Overall	59.1	68.7	69.7	69.9	70.0

Table 5.1: Average recall rate (in percentage) for top 16 video clips retrieved (5-level decomposition using db2)

The highest performance is achieved in the commercials class. The overall recall of 70 per cent is obtained after the 4th iteration. The greatest performance improvement is observed after the first iteration. Figure 5.5 depicts the comparison of retrieval performance for the 5 classes. When the audio clips are decomposed up to 7 levels, more features become available for indexing. The dimension of the feature

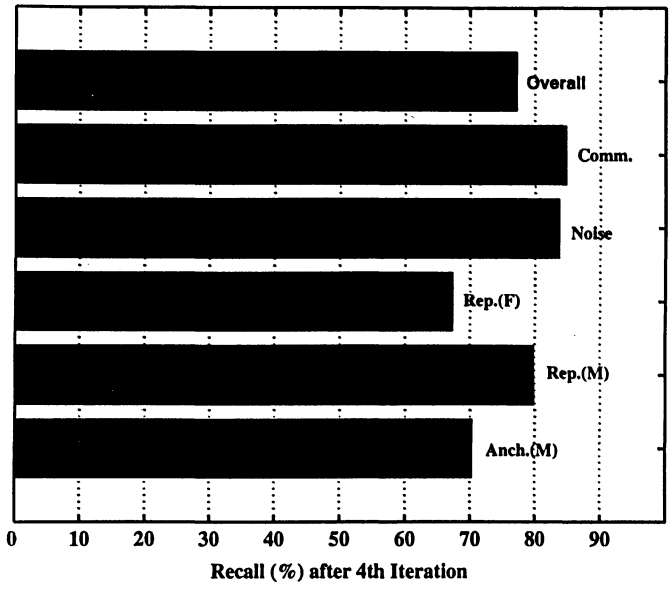

Figure 5.6: Retrieval Performance in 5 Classes (7-level Decomposition using db2)

vector becomes 23. This increases the performance of the system. These results are summarized in Table 5.2. The overall recall is increased from 70 percent to 77.2 per cent by taking more features. The improvement is very significant, especially in the case of Noise class where the recall is increased from 64.4 percent to 83.7 per cent. The results also emphasize the importance of the relevance feedback in improving the accuracy of the system. The comparison of retrieval performance in the 5 classes is

Category/Iteration	Initial	Iteration 1	Iteration 2	Iteration 3	Iteration 4
Anchor(Male)	53.1	69.1	70.0	70.3	70.3
Reporter(Male)	60.8	78.0	79.7	79.8	79.8
Reporter(Female)	46.1	64.7	67.1	67.3	67.3
Noise	79.8	83.7	83.7	83.7	83.7
Commercials	70.6	83.2	84.7	84.8	84.8
Overall	62.1	75.7	77.0	77.2	77.2

Table 5.2: Average recall rate (in percentage) for top 16 video clips retrieved (7-level decomposition using db2)

depicted in Figure 5.6.

In Tables 5.3 and 5.4, we have presented the results using the db4 wavelet kernel

for decomposition of embedded audio clips. It is observed that db4 wavelet kernel performs better than the db2 wavelets. The overall recall after 4th iteration is 79.6 per cent in case of 7-level decomposition using db4 wavelets. We observe an overall improvement of 1.1 per cent in case of 5-level decomposition. An overall improvement of 2.4 per cent in results is attained with 7-level decomposition using db4 wavelet kernel in comparison with db2 wavelets. However, the performance of db2 wavelets is better than db4 in noise category. Figures 5.7 and 5.8 illustrate the class wise

Category/Iteration	Initial	Iteration 1	Iteration 2	Iteration 3	Iteration 4
Anchor(Male)	51.8	70.3	72.5	73.1	73.1
Reporter(Male)	63.5	79.4	81.6	82.2	82.2
Reporter(Female)	45.5	54.5	56.9	57.4	57.4
Noise	57.2	58.2	58.2	58.2	58.2
Commercials	69.3	82.9	83.9	84.5	84.6
Overall	57.5	69.1	70.6	71.1	71.1

Table 5.3: Average recall rate (in percentage) for top 16 video clips retrieved (5-level decomposition using db4)

Category/Iteration	Initial	Iteration 1	Iteration 2	Iteration 3	Iteration 4
Anchor(Male)	56.6	71.9	75.0	75.3	75.3
Reporter(Male)	66.4	84.9	88.0	88.4	88.4
Reporter(Female)	45.0	64.7	69.5	70.1	70.2
Noise	79.8	80.3	80.3	80.3	80.3
Commercials	70.1	82.0	83.5	83.9	84.0
Overall	63.6	76.7	79.3	79.6	79.6

Table 5.4: Average recall rate (in percentage) for top 16 video clips retrieved (7-level decomposition using db4)

performance of the system with 5-level and 7-level wavelet decomposition respectively. The best results were attained in the commercials category while the performance in the female reporter class was the lowest.

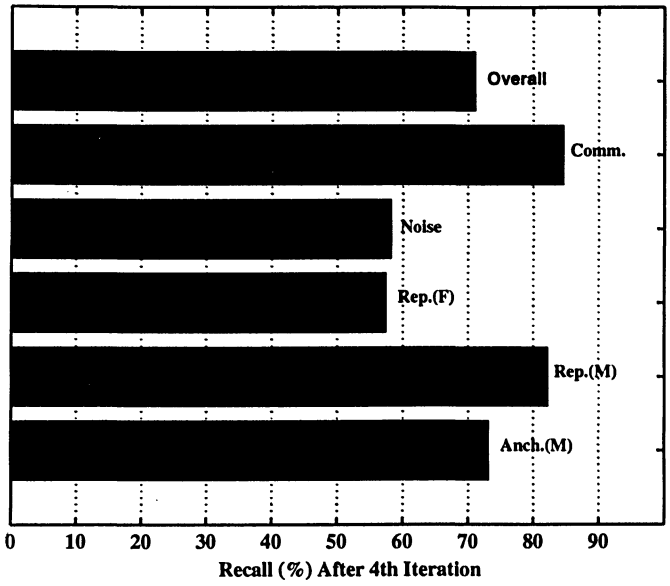

Figure 5.7: Retrieval Performance in 5 Classes (5-level Decomposition using db4)

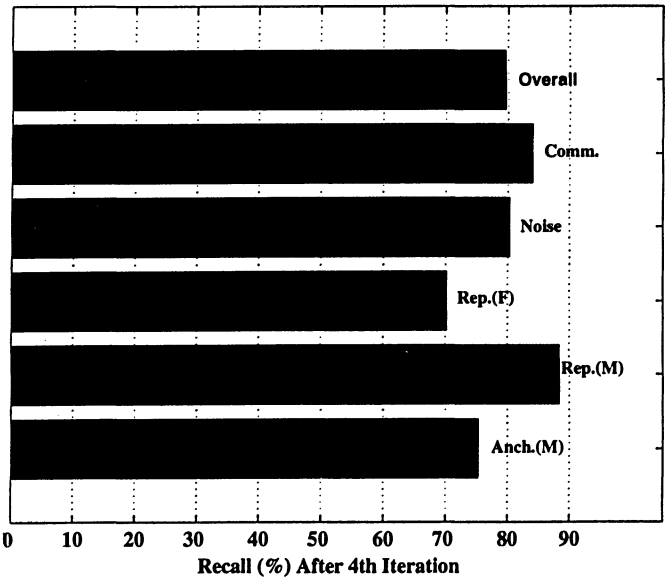

Figure 5.8: Retrieval Performance in 5 Classes (7-level Decomposition using db4)

5.6 Conclusions

We presented a new feature extraction method for video retrieval based on the embedded audio content. The video clips are indexed using a low dimensional feature vector that is a good representation of the global similarity of the audio contents. We demonstrate the ability of the embedded audio content of the digital video for searching the databases based on the auditory information. This may be particularly useful for finding the action sequences in the videos. A comprehensive experimental evaluation of the system is presented using a news video database. It has been observed that the recall rate is improved significantly by increasing the decomposition level. This is due to the increased number of features becoming available. This increase in performance is, however, achieved at the expense of higher computational cost. The experimental results also demonstrate the better performance of db2 wavelet filter over db1 for analysis of embedded audio.

Chapter 6

Conclusions

In this work, we have proposed a new feature extraction technique based on the statistical analysis of the wavelet coefficients. The shape of the wavelet coefficients distributions is modelled by a mixture of Laplacians. It has been observed that the proposed model is highly suitable for modelling the peaky distributions of the wavelet coefficients. The proposed approach is applied to the image and audio-based video retrieval. The parameters of the model are used for indexing the texture images. Experimental results indicate the validity of the adopted modelling method. It is observed that the extracted features possess a high discriminatory power for the texture description. The dimensionality curse is the main drawback in any feature based indexing scheme. It is a very important factor that should be kept in mind for an efficient retrieval strategy. The computational complexity of the system during retrieval and feedback cycle depends on the dimension of the feature space. The proposed technique generates a low dimensional feature space which reduces the computational complexity during the retrieval process. The time taken by the retrieval process is very important. The users are interested in the relevancy of the results as well as in the quick system response. This has been achieved by keeping the dimension of the feature vector low.

Audio is an important component of the multi-modal video data. The proposed technique is based on the shape of the wavelet coefficient distributions. Therefore, it is also applicable to the audio analysis. In this study, we have successfully applied this technique to find the global attributes of the embedded audio content. It is noted

that the proposed features are good for describing the global characteristics of the audio. A comprehensive experimental evaluation has been performed on a news video database. The unique characteristics of embedded audio contents in the video data requires the development of new techniques for its analysis. The traditional methods found in literature focus on the audio classification into certain number of classes such as speech, silence, music. The news videos contain mixed types of audio content. The music, speech and noise are often present in the same clip. It is observed that the proposed features are effective for this type of mixed audio sources.

6.1 Learning from User Feedback

In this work, we clearly demonstrated the power of relevance feedback in improving the retrieval efficiency of the systems. In the image retrieval case, both negative and positive examples has been used for learning from the user input. However, the two class learning strategy is not very flexible. For example, in the relevant set of images, the users might consider some images as more relevant than others. The relevance feedback can be extended to obtain multi-class input from the users. The users will label the images after the initial search. This approach can allow a better understanding of the users' notion of image similarity. Video retrieval is more complex than the image retrieval. The video data contains significantly more information than images. Presently, the proposed system implements a simple relevance feedback learning scheme. The weights associated with the feature vector components are updated in each iterative cycle. The negative examples are not used for learning about what the user is looking for. We are considering to implement the multi-class approach for the relevance feedback in which both negative and positive examples will be used for tuning the system parameters.

6.2 Future Research Extension

The feature set proposed in this study is a very good representation of the texture contents of the images. The approach has been tested using the standard Brodatz

image database which contains only grey-scale images. Application of this technique to the color image database is one of the possible research extensions. The color images contain much more information than that of grey-scale images. The proposed features can be combined with other features describing color and shape contents of the images.

6.2.1 Fusion of Multi-modality

The text based retrieval of the audio visual data is still the most popular way of searching the databases. Most of the web based search engines have extended their capabilities to the search of multimedia documents. The purpose of the content-based search is to complement the existing systems rather than replace them entirely. The text indices combined with the content-based features can provide a flexible interaction with the system. Users can search the database based on the text, by providing an example,by sketching a diagram or a combination of different clues.

In video analysis, we have explored the embedded audio contents for building suitable indices. Some of the events are easy to identify in audio domain while the others are easier in the visual domain. Using different media will result in better understanding of the video data. The semantic retrieval of the video clips needs an integration between the features. The fusion of the multi-modality in the video data is a future direction of research.

6.2.2 Flexible Queries

Although QBE retrieval paradigm is the most popular, it may not suit the user requirements in certain cases. To find suitable content samples may be hard in certain situations. Hence, the system should support other query formats such as text, sketch or vocal input. This is particularly useful when the users do not have a clear idea about what they are looking for. If at the beginning they only have a rough sketch in mind, they can just start the retrieval process by drawing a that sketch and may refine their query at later stages.

Bibliography

[1] H.P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", *IBM Journal of Research and Development*, pp. 309–317, October 1957.

[2] Google, *http://www.google.com*

[3] Lycos, *http://www.lycos.com*

[4] Altavista, *http://www.altavista.com*

[5] S.F. Chang, T.and A. Puri, "Overview of the MPEG-7 Standard", *IEEE Trans. Circuits and Systems for Video Technology*, vo. 1, no. 6, pp. 688–702, June 2001.

[6] Alberto Del Bimbo, *Visual Information Retrieval*, San Francisco, CA, Morgan Kaufmann Publishers Inc., 1999.

[7] M.J Swain and D.H. Ballard, "Indexing Via Color Histograms", *Proc. Third International Conference on Computer Vision*, pp. 11–32, April 1991.

[8] J.R. Smith and S.F. Chang, "Single Color Extraction and Image Query", *Proc. ICIP*, 1995.

[9] Y. Gong, et. al., "Image indexing and retrieval using color histograms", *Proc. Multimedia Tools and Applications*, Vol. 2, pp. 133–156, 1996.

[10] W. Hsu, et. al., "An integrated color-spatial approach to content-based image retrieval", *Proc. 3rd ACM Multimedia Conference*, Nov 1995.

[11] J.R. Smith and S.F. Chang, "Tools and techniques for color image retrieval", *Proc. SPIE Proceeding*, 1996.

[12] J.R. Smith and S.F. Chang, "VisualSeek: A fully Automated Content-based Image Query System", *Proc. Proceedings of ACM Multimedia Conference*, pp. 87–98, 1996.

[13] Stricker and Dimai, "Color indexing with weak spatial constraints", *Proc. SPIE Proceeding*, 1996.

[14] Stricker and Dimai, "Spectral Covariance and fuzzy regions for image indexing", *Journal Machine Vision and Applications*, Vol. 10, pp. 66–73, 1997.

[15] Pass and Zabih, "Histogram refinement for content-based image retrievalg", *IEEE Workshop on Applications of Computer Vision*, 1996.

[16] Pass, Zabih and Miller, "Comparing Images using Color Coherence vectors", *Proc. Fourth ACM Multimedia Conference*, 1996.

[17] Cinque, Levialdi and A. Pellicano, "Color-based image retrieval using Spatial-Chromatic histograms", *Proc. IEEE Multimedia sustems*, Vol. II, pp. 969–973, 1999.

[18] M. Mitra, T.J. Huang and S.R. Kumar, "Combining supervised learning with color crrelograms for content-based image retrieval", *Proc. I5th ACM Multimedia Conference*, 1997.

[19] M. Stricker and M. Orengo, "Similarity of color images", *Proc. SPIE Storage and Retrieval for image and video databases*, 1995.

[20] C.E Jacobs, A. Finkelstein and D.H. Salesin "Fast multiresolution image querying", *Proc. Computer Graphics Conference*, 1995.

[21] J. Vellaikal and C.C.J. Kuo, "Content-based Retrieval using multiresolution histogram representation", *Digital Image storage and Archiving Systems*, pp. 312–323, 1995.

[22] K.C. Liang and C.C.J. Kuo, "Progressive Image Indexing and Retrieval based on embedded wavelet coding", *Proc. International Conference on Image Processing*, vol. 1, pp. 572–575, 1997.

[23] R. Haralick, K. Shanmugam and I. Dinstein, "Texture features for image classification", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, 1973.

[24] C.C. Gotliab and H. Kreyszig, "Texture descriptors based on co-occurrence matrices", *Proc. Computer Vision, Graphics and Image*, pp. 70-80, 1990.

[25] H. Tamura, S. Mori and T. Yamawaki, "Texture features corresponding to visual perception", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, no. 6, pp. 460-473, 1978.

[26] J.R. Smith and S.F. Chang, "Automated binary texture feature sets for image retrieval", *Proc. IEEE International Conference on Accoustics, speech and Signal Processing*, Atlanta, GA, 1996.

[27] T. Chang and C.C.J. Kuo, "Texture analysis and classification with tree-structured wavelet transform", *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 429–441, October 1993.

[28] M.H. Gross, R. Koch, L. Lippert and A. Dreger, "Multiscale image texture analysis in wavelet spaces", *Proc. IEEE International Conference on Image Processing*, 1994.

[29] K.S. Thyagarajan, T. Nguyen and C. Persons, "A maximum likelihood approach to texture classification using wavelet transform", *Proc. IEEE International Conference on Image Processing*, 1994.

[30] A. Kundu and J.L. Chen, "Texture classification using qmf bank-based sub-band decomposition", *Journal Computer Vision, Graphics and Image Processing*, vo. 54, no. 5, pp. 369–384, September 1992.

[31] W.Y. Ma and B.S. Manjunath, "A comparison of wavelet transform features for texture image annotation", *Proc. IEEE International Conference on Image Processing*, 1995.

[32] Y. Rui, A.C. She and T.S. Huang, "Modified fourier descriptors for shape representation - a practical approach", *Proc. First International Workshop on Image Databases and Multimedia Search*, 1996.

[33] E. Persoon and K.S. Fu, "Shape discrimination using fourier descriptors", *IEEE Transactions on Systems, Man and Cybernetics*, 1977.

[34] C.T. Zahn and R.Z. Roskies, "Fourier descriptors for plane closed curves", *IEEE Transactions on Computers*, 1972.

[35] Y. Rui, A.C. She and T.S. Huang, "Fourier descriptors for plane closed curves", *IEEE Transactions on Computers*, 1972.

[36] M.K. Hu, "Visual pattern recognition by moment invariants", *Proc. IEEE conference on Computer Methods in Image Analysis*, Los Angeles, 1977.

[37] E.M. Arkin, L. Chew, D. Huttenlocher, K. Kedem and J. Mitchell, "An efficiently computable metric for comparing polygon shapes", *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vo. 13, March 1991.

[38] G.C.H. Chuang and C.C.J. Kuo, "Wavelet descriptor of planer curves: Theory and applications", *IEEE Transactions on Image Processing*, vo. 5, pp. 56–70, January 1996.

[39] J. Lay and L. Guan, "Concept-based retrieval of art documents", *Proc. Int. Conf. on Image and Video Retrieval*, Champaign, 2003.

[40] J. Lay and L. Guan, "Retrieval of color artistry concepts", *To appear in IEEE Trans. on Image Processing.*,

[41] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval", *IEEE Transactions on Circuits and Systems for Video Technology*, vo. 8, pp. 644–655, January 1998.

[42] Y. Rui, T.S. Huang and S. Mehrotra, "Content-based image retrieval with relevance feedback in MARS", *Proc. of IEEE International Conference on Image Processing*, pp. 815–818, 1997.

[43] Cox, Miller, Omohundro and Yianilos, "PicHunter: Bayesian relevance feedback for image retrieval", *Proc. 13-th International Conference on Pattern Recognition*, Vol. 3, pp. 25–29, August 1996.

[44] T.P. Minka and R.W. Picard, "Interactive learning with a Society of Models", *Proc. IEEE Computer Vision and Pattern Recognition Conference*, pp. 447–452, June 1996.

[45] J. Peng, "A multi-class relevance feedback approach to image retrieval", *Proc. IEEE Int. Conf. on Image Processing*, Thessaloniki, Greece, vol. 1, pp. 46–49, October 2001.

[46] T.V. Ashwin, N. Jain and S. Ghosal, "Improving image retrieval performance with negative relevance feedback", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, vol. 3, pp. 1637–1640, 2001.

[47] N. Vasconcelos and A. Lippman, "Bayesian relevance feedback for content-based image retrieval", *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, Hilton Head, SC, pp. 63–67, 2000.

[48] L. Zhang, F. Lin and B. Zhang, "Support vector machine learning for image retrieval", *Proc. IEEE Int. Conf. on Image Processing*, Thessaloniki, Greece, vol. 2, pp. 721–724, October 2001.

[49] Y. Chen, X.S. Zhou and T.S. Huang, "One-class SVM for learning in image retrieval", *Proc. IEEE Int. Conf. on Image Processing*, Thessaloniki, Greece, vol. 1, pp. 34–37, October 2001.

[50] Y. Wu, Q. Tian and T.S. Huang, "Integrating unlabeled images for image retrieval based on relevance feedback", *Proc. IEEE Int. conf. on Pattern Recognition*, Barcelona, Spain, vol. 1, pp. 21–24, 2000.

[51] T. Wang, Y. Rui and S.M. Hu, "Optimal adaptive learning for image retrieval", *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Hawaii, USA, vol. 1, pp. 1140–1147, 2001.

[52] Y. Zhuang, X. Liu, and Y. Pan, "Apply semantic template to support content-based image retrieval", *Proc. SPIE Storage and Retrieval for Multimedia Database*, USA, pp. 442–449, Jan 2000.

[53] M. La Cascia, S. Sethi and S. Sclaroff, "Combining textual and visual cues for content-based image", *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, Los Angeles, USA, pp. 24–28, June 1998.

[54] X.S. Zhou and T. Huang, "Unifying keywords and visual contents in image retrieval", *IEEE Multimedia*, Vol. 9, no. 2, pp. 23-33, 2002.

[55] J. Laaksonen, M. Koskela and E. Oja, "PicSom-self-organizing image retrieval with MPEG-7 content descriptions", *IEEE Trans. on Neural Network*, vol. 13, no. 4, pp. 841–853, July 2002.

[56] P. Muneesawang and L. Guan, "Interactive CBIR using RBF-based relevance feedback for WT/VQ coded images", *Proc. IEEE int. Conf. Accoust., Speech, Signal Processing*, pp. 1641–1644, May 2001.

[57] Y. Rui and T.S. Huang, "Optimizing Learning In Image Retrieval", *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Hilton Head, SC, vol. 1, pp. 236–243, June 2000.

[58] F. Arman, R. Depommier, A. Hsu and M.Y. Chiu, "Content-based Browsing of Video Sequences", *Proc. Second ACM international conference on Multimedia*, pp. 97–103, 1994.

[59] H.J. Zhang, A. Kankanhalli and S.W. Smoliar, "Automatic Partitioning of Full-Motion Video", *Multimedia Systems*, vol. 1, no. 1, pp. 10–28, June 1993.

[60] H.J. Zhang, S.Y. Tan, S.W. Smoliar and G. Yihong, "Automatic Parsing and Indexing of News Video", *Multimedia Systems*, vol. 2, no. 6, pp. 256–266, Jan. 1995.

[61] M.J. Pickering, D. Heesch and S. Rger, "Retrieval Using Global Features in Keyframes", *Proc. The Eleventh Text Retrieval Conference (TREC)*, 2002.

[62] A. Nagasaka and Y. Tanaka "Automatic Video Indexing and Full-Video Search for Object Appearances", *Proc. Second Working Conference on Visual Database Systems*, Budapest, Hungary, pp. 113–127, 1991.

[63] J.D. Courtney, "Automatic video indexing via object motion analysis", *IEEE Pattern Recognition*, vol. 30, no. 4, pp. 607–625, April 1997.

[64] S. Dagtas, W. Al-Khatib, A. Ghafoor and R.L. Kashyap, "Models for motion-based video indexing and retrieval", *IEEE Trans. on Image Processing*, vol. 9, no. 1, pp. 88–101, 2000.

[65] E. Ardizzone and M. La Cascia, "Automatic video database indexing and retrieval", *Multimedia Tools and Applications*, vol. 4, pp. 29–56, 1997.

[66] R. Nelson, R. Polana, "Qualitative recognition of motion using temporal texture", *Proc. Computer Vision, Graphics, and Image*, vol. 56, no. 1, pp. 78–99, July 1992.

[67] K. Otsuka, T. Horikoshi, S. Suzuki and M. Fujii, "Feature extraction of temporal texture based on spatio-temporal motion trajectory", *Proc. 14th Int. Conf. on Pattern Recognition*, pp. 1047–1051, August 1998.

[68] M. Szummer and R.W. Picard, "Temporal texture modeling", *Proc. 3rd IEEE Int. Conf. on Image Processing*, pp. 823–826, September 1996.

[69] R. Fablet and P. Bouthemy, "Motion-based feature extraction and ascendant hierarchical classification for video indexing and retrieval", *Proc. 3rd Int. Conf. on Visual Information and Information Systems*, June 1999.

[70] R. Fablet, P. Bouthemy and P. Prez, "Statistical Motion-Based Video Indexing and Retrieval", *Proc. 6th Int. on Content-Based Multimedia Information Access*, April 2000.

[71] J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music", *Proc. IEEE ICASSP 1996*, vol. 2, pp. 993–996, May 1996.

[72] E. Wold, T. Blum, D. Keislar and J. Wheaton, "Content-Based Classification, search and Retrieval of Audio", *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.

[73] Compaq Corporate Res. Lab, *Available: http://speechbot.research.compaq.com.*

[74] J. Nam, A.E. Cetin and A.H. Tewfik, "Speaker identification and video analysis for hierarchical video shot classification", *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA, October 1997.

[75] M. Akutsu, A. Hamada and Y. Tonomura, "Video handling with music and speech detection", *IEEE Multimedia*, vol. 5, no. 3, pp. 17-25, 1998.

[76] P. Jang and A. Hauptmann, "Learning to recognize speech by watching television", *IEEE Intell. Syst. Mag.*, vol. 14, no. 5, pp. 51-58, 1999.

[77] X. Tang, X. Gao, J. Liu and H. Zhang, "A spatial-temporal approach for video caption detection and recognition", *IEEE Trans. on Neural Network*, Vol. 13, Issue 4, pp. 961–971, July 2002.

[78] R. Wang, M.R. Naphade and T.S. Huang, "Video retrieval and relevance feedback in the context of a post-integration model", *IEEE Int. Workshop on Multimedia Signal Processing*, Cannes, France, pp. 33–38, 2001.

[79] M.R. Naphade, T. Kristjansson, B. Frey and T.S. Huang: "Probabilistic Multimedia Objects (Multijects): A Novel Approach to Video Indexing and Retrieval in Multimedia Systems", *Proc. International Conference on Image Processing*, vo. 3, pp. 536–540, 1998.

[80] N. Haering, R.J. Qian and M.I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 6, pp. 857–868, September 2000.

[81] S.H. Jeong, J.H. Choi and J.D. Yang, "A concept-based video retrieval model with spatio-temporal fuzzy triples", *Proc. IEEE Region 10 Int. Conf. on Electrical and Electronic Technology*, vol. 1, pp. 424–429, 2001.

[82] M.A. Smith and T. Kanade "Video Skimming and Characterization through the Combination of Image and Language Understanding", *Proc. IEEE CS Conference on Computer Vision and Pattern Recognition*, pp. 775–781, 1997.

[83] Y. Nakamura and T. Kanade "Semantic Analysis for Video Contents Extraction-Spotting by Association in News Video", *Proc. Fifth ACM International Conference on Multimedia*, pp. 393–401, 1997.

[84] R. Polikar, "The Wavelet Tutorial", *http://www.engineering.rowan.edu/ polikar/WAVELETS*, Jan. 2003.

[85] P.M. Bentley and J.T.E. Mcdonnel, "Wavelet transforms: an introduction", *IEEE Electronics and Communication Engineering Journal*, pp. 175–186, August 1994.

[86] Agostino Abbate, Casimer M. DeCusatis and Pankaj K. Das, *Wavelets and Subbands: Fundamentals and Applications*, York, PA, Birkhauser Boston, 2002.

[87] M. Figueiredo and A.K. Jain, "Unsupervised selection and estimation of finite mixture models", *Proc. International Conference on Pattern Recognition*, Barcelona, 2000.

[88] A.K. Jain, R.P.W. Duin and Jianchang Mao, "Statistical pattern recognition: a review", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, Pp. 4–37, January 2000.

[89] P.M. Bentley and J.T.E. Mcdonnel, "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov model", *Technical Report ICSI-TR-97-021, University of Berkeley* , Berkely, CA, April 1998.

[90] A.A. D'Souza, "Using EM To Estimate A Probablity Density With A Mixture Of Gaussians", *http://www-clmc.usc.edu/ adsouza/notes/mix$_g$auss.pdf*

[91] H. Yuan, X. Zhang and L. Guan, "Content-based image retrieval using a Gaussian mixture model in the wavelet domain", *Visual Communications and Image Processing* , Lugano, Switzerland, July 2003.

[92] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra, "Relevance Feedback: a power tool for interactive content-based image retrieval", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 8, pp. 644–655, April 1998.

[93] M.J. Swain and D.H. Ballard, "Color Indexing", *Journal of Computer Vision*, vol. 7(1), pp. 11–32, May 1997.

[94] Oge Marques and Borko Furht, *Content-Based Image and Video Retrieval*, Norwell, MA, Kluwer Academic Publishers, 2002.

[95] T.P. Minka and R. Picard, "Interactive learning using a society of models", *Technical Report 349, MIT Media Labs*, 1995.

[96] P. Muneesawnag and L. Guan, "A non-linear RBF model for interactive content-based image retrieval", *The First IEEE Pacific-Rim Conference on Multimedia*, pp. 188–191, December 2000.

[97] P. Muneesawnag and L. Guan, "An interactive approach for CBIR using a network of radial basis functions", *To appear in IEEE Transactions on Multimedia*, 2004.

[98] Sigitani et. al., "Image Interpolation for progressive transmission by using radial basis function networks", *Neural Networks*, vol. 10, no. 2, pp. 381–390, December 1999.

[99] T.S. Huang, S. Mehrotra and K. Ramchandram, "Multimedia analysis and retrieval system (MARS) project", *Proc. 33rd Annual Clinic on Library Application of Data Processing*, Digital Image Access and Retrieval, 1996.

[100] G. Salton and M.J. McGill, *Intoduction to Modern Information Retrieval*, NY, McGraw- Hill Book Company, 1983.

[101] J. Peng, B. Bhanu and S. Qing, "Probabilistic feature relevance learning for content-based image retrieval", *Computer Vision and Image Understanding* , vol. 75, no. 1/2, pp. 150–164, July/August 1999.

[102] L. Guan, S.Y. Kung and J. Larsen, *Multimedia Image and Video Processing*, Florida, CRC Press LLC, 2001.

Appendix A

List of Publications

In this section, we list the publications resulted from our research work for the thesis.

- L. Guan, P. Muneesawang, J. Lay, I. Lee and T. Amin, "Recent Advancement in Indexing and Retrieval of Visual Documents"; Proceedings of the 9th International Conference on Distributed Multimedia Systems, Miami, Florida, USA, September 24-26, 2003.

- Tahir Amin and Ling Guan, "Interactive Content-Based Image Retrieval Using Laplacian Mixture Model in the Wavelet Domain"; Accepted for presentation at IEEE International Symposium on Circuits and Systems, Vancouver, May 23-26, 2004.

- Tahir Amin, Mehmet Zeytinoglu, Ling Guan and Qin Zhang, "Interactive Video Retrieval Using Embedded Audio Content"; Accepted for presentation at IEEE International Conference on Acoustics, Speech and Signal Processing, Montreal, May 17-21, 2004.