Chapter 1: Introduction

This chapter presents an introduction to the problems discussed in this thesis, motivation behind the present work and objectives of the thesis. The chapter concludes with a list of contributions to this thesis in the field of content-based image retrieval and its application for the diagnosis of breast cancer using mammograms.

1.1 Background

In this computer and internet age, with the development multimedia technologies and the availability of image capturing devices, virtually all spheres of human life including education, commerce, government, academics, medicine, security, surveillance, engineering, architecture, journalism, entertainment, historical research, and graphic design etc., use images for their services [1]. Due to this, the size of digital image collection is increasing rapidly and consequently searching for relevant images is getting difficult [2]. This has created an on-going demand for systems that can store and retrieve multimedia data in an effective way.

As data mining, image mining broadly deals with extraction of the valuable information embedded in the large image and video databases [3]. Retrieval of a query image from a large database of images is an important task in the area of computer vision and image processing. The advent of large multimedia collection and digital libraries has led to an important requirement for development of search tools for indexing and retrieving information from them. The development of search engine is nice evidence of this. However, most search engines are based on text retrieval, where images are indexed and retrieved based on these rudimentary descriptions such as size,

keywords, type, date and time of capture, the identity of the owner or some text description of the image. As a result, this is often called description based or text-based image retrieval process. This traditional manual keyboard-based approach to searching and indexing images are labor intensive, expensive, less accurate and also trivial for untagged image search [4]. Further, description of images is very much subjective, and a picture holds many different backgrounds. Due to this, surrounding text may not be able to describe the image accurately. Also, digital images are rapidly expanding in quantity and heterogeneity, and this textual-based information retrieval technique does not meet the user's demands [5]. Therefore, there is a need to develop an efficient system. To overcome these disadvantages in text-based retrieval system, content-based image retrieval (CBIR) was introduced in the early 1990s [6].

In CBIR systems, the images are searched and retrieved based on the visual content of the images. Ideally, a CBIR system should automatically extract the low-level features about the images for a specific application area. To a very large extent, the low-level image features such as colour, texture, and shape are widely used for CBIR [7-9].

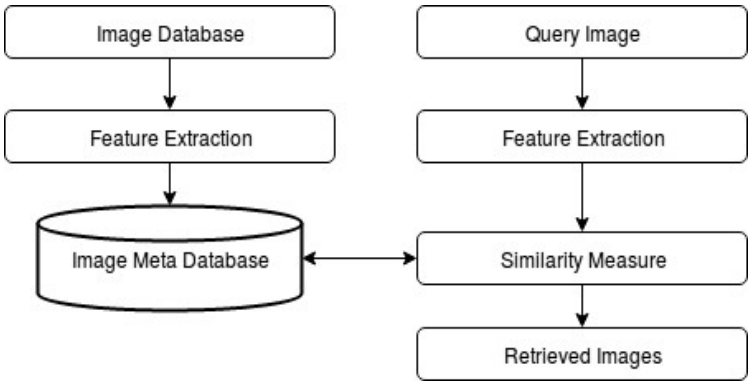

Fig. 1.1: Architecture of a general CBIR system

The framework of a general CBIR system is shown in Fig. 1.1. The CBIR systems architecture is divided into two sections. In the first section, feature vector is

extracted from each image in the database to form the metadata information and these features are used to index the image, and they are stored into the database along with the images. In the second section, at a query time, a feature vector is extracted from the query image, and it is matched against the feature vectors in the database to retrieve closest possible similar images available in the database [10].

In last decade content-based image retrieval has been applied in various domains such as economy, industry, army, sports, publishing and advertising, historical research, fashion and graphic design, architectural and engineering design, forensic and crime detection, medical applications, geographical information and remote sensing systems , etc. [11]. Among these, medical image retrieval has become more attractive and has been proposed for various applications especially in health-care database management, medical diagnosis aid, and medical research [12, 13]. The goals of medical information systems have often been defined to deliver the needed information at the right time, the right place and to the right persons for improving the quality and efficiency of care processes. In the last several years, developing computer-aided detection and/or diagnosis (CAD) schemes that use CBIR to search for the clinically relevant and visually similar medical images (or regions) depicting suspicious lesions has also been attracting research interest.

In this thesis, we have designed and developed some new CBIR methods for the general images and x-ray mammograms, where for a mammogram; we applied the different variants of computer vision techniques on the database of x-ray images for the diagnosis of breast cancer.

Breast cancer remains the leading cause of death in women worldwide, and after the age of 60 years the risk of developing it in women is very high. Although the most accurate detection method in the medical environment is biopsy; it is an aggressive

invasive procedure that involves some risks, patient discomfort, and high cost. Moreover, there is a high percentage of negative cases (70% - 90%) where breast biopsies were performed unnecessarily [14]. Currently, mammography is the dominant method for early detection of breast cancer, which attempts to reduce the negative biopsy ratio and the low-cost to society by improving feature analysis. The acquisition of a mammogram (i.e., a low dose X-ray of the breast region) is done by compressing the breast of the patient between two acrylic plates for a few seconds when the x-ray is emitted [15]. A typical mammogram is an intensity image with gray levels, showing the levels of contrast inside the breast which characterize normal tissue, vessels, different masses or calcification and noise. From the image a trained physician screens it searching for abnormalities of various forms and if found that these artefacts on the image could be signed for the presence of a benign or a malignant tumour [16]. Sample mammogram for the screening of breast cancer is shown in Fig. 1.2.

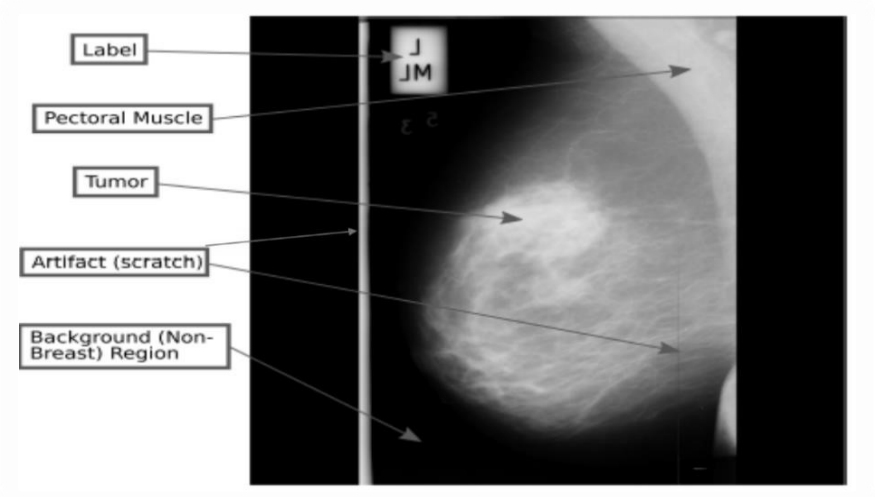

Fig. 1.2: Sample mammogram

From this sample mammogram, we can see that various distorting artefacts like labels, scratches and pectoral muscles are presented. These artefacts have high- intensity

gray values, and visible appearances are much closer to abnormal mammograms [17]. The presence of these artefacts in the X-ray mammogram misguided the existing segmentation algorithms, and they are unable to segment accurate pathology-bearing regions. So suppression of these is an essential pre-processing step [18]. Therefore, for analysis of mammogram, the region of interest (ROI) is manually cropped from the breast area for avoiding unwanted tags, labels and pectoral muscles. But this manual pre-processing requires too much time and is expensive to implement.

Nowadays, a huge number of mammograms is generated in hospitals creating a need to develop an automated tool which may help radiologists to retrieve and analyse current images with past stored images [18-20]. CBIR system assisted with screening mammography has been well suited and effective method for the diagnosis of abnormalities in the breast. Its goal is to provide the radiologist with a set of images from past cases that are relevant to the one being evaluated, along with the known pathology of these past cases.

1.2 Motivation

CBIR system is applied in various areas for the ease of image management and knowledge discovery from a huge database of images. Finding an image on this database is an often complicated task due to the excessive amount of irrelevant records. As the amount of collection of digital images has grown continuously, it is necessary to develop an appropriate system for effective management of these collections and also for retrieving the most relevant images in less searching time. As it is a well-known fact that image before entering to a retrieval process has to be captured by a device like a camera, and after capturing of an image, some information of the real world is automatically lost. This information loss may be due to several factors such as low-

resolution, bad illumination, and viewing angles or a deficiency in the lenses. In retrieval procedure, the major lack is the difference between the extracted low-level features and the objects. In general, there is no direct link between the high-level semantics and the low-level features [21]. Therefore, most CBIR approaches are facing problem to bridge the semantic gap between low-level visual features and high-level understanding of objects. Also, an issue of searching time affects the retrieval process because in conventional CBIR system, query image is compared with all images of the database, in turn, this exhaustive search slows down the retrieval process.

Further, in the diagnosis of breast cancer, CBIR system may help radiologist by providing relevant supporting information from prior known cases potentially leading to improvement in their diagnostic accuracy. It can contribute more reliable diagnosis by classifying the query mammograms and retrieving similar types of mammograms already annotated by diagnostic descriptions and treatment results. But it has been reported that the mammograms are much affected by artefacts and noises. So, there are big challenges for the removing of artefacts, finding the ROIs, and capturing the accurate representation of regions. There is big research gap in pre-processing, and feature extraction level. Therefore, these challenges, gaps and crucial application of CBIR for the diagnosis of breast cancer motivated us to design abstractions for effective browsing and retrieving.

1.3 Problem Statement and Thesis Objectives

This thesis is related to the development of a CBIR system to be used exclusively in general application and medical context. As previously discussed, the main problem behind an efficient CBIR is a semantic gap between low-level features and high-level image analysis. Therefore, low-level visual features cannot detect an object reliably. For

example, colour histogram of an image can be easily extracted but, the presence of particular objects such as elephants, dinosaurs, or horses cannot be truly detected [22]. Still, this problem is not adequately addressed in the current CBIR systems. The performance of CBIR systems depends upon fast and efficient indexing of the images. So it is very important to extract fast and informative features for holding the accurate representation of an image. Further, one of the most important aspects is how long queries take to execute for the desired outcome because the size of image databases has been increasing rapidly and making a linear search in the whole database is time-consuming process [23]. A system that always gives perfect results but takes longer time for the output is not useful. Therefore, CBIR systems should be interactive, so that users can easily access the answer of a query in less amount of time. Thus, for ease of use and interaction, graphical user interface (GUI) must be provided to the user. Such GUI based CBIR system shall increase the retrieval accuracy and decrease response time of query.

The objective of the proposed works in this thesis for general CBIR is to apply machine learning approaches with different variants of fast features for relevant browsing in less searching space. Further, for mammogram retrieval, the main objective is to improve the retrieval and diagnosis information in less search time through the methods related to image pre-processing, segmentation, wavelet-based local neighbourhood feature extraction, clustering, and classification.

The brief contribution and objectives of this thesis are as follows:

(i) To study and compare the performances of the various conventional as well as state-of-the-art methods for CBIR systems.

(ii) Design, development, and implementation of efficient algorithms for CBIR system by using a fusion of fast features, varying weighted similarity measures

and machine learning approaches which are fast in indexing and reduce the effect of the semantic gap and search space.

(iii) Design, development, and implementation of an effective classification cum retrieval system for mammograms by using relevant wavelet-based centre symmetric-local binary patterns features (WCS-LBP) features with random forests, which can assist as a CAD system for the diagnosis of breast cancer.

(iv) Design, development, and implementation of a new automated algorithm for the removal of artefacts, scratches, labels, and pectoral muscles. Further, selection of a more reliable region of interest using region growing algorithm with gray level co-occurrence matrix (GLCM) contrast based selective thresholds.

(v) Design, development, and implementation of the automated algorithms using wavelet-based CS-LBP features and self-organizing map (SOM), which retrieves most relevant mammograms to a query in less searching time.

All above mentioned proposed methods have been implemented using MATLAB R 2014 software on a standard Intel Core i7 PC with 8 GB RAM and tested for standard datasets. Their performance were evaluated using various performance measures and also the performance of each of the proposed methods is compared against state-of-the-art methods available in literature. The obtained results and their performance analyses justify the applicability of the proposed approaches.

1.4 Outline of the Thesis

This thesis consists of six chapters. Outline of the thesis is as follows:

Chapter 1 presents a brief introduction of the problems addressed in this thesis followed by the application for the diagnosis of breast cancer using mammogram retrieval, motivation, and objectives of the thesis.

Chapter 2 discusses the theoretical background for CBIR system. In this chapter, we have also given an overview of feature extraction methods and similarity measures. Further, in this chapter, comprehensive literature review and comparative study of various traditional as well as state-of-art methods for CBIR system are given.

Chapter 3 is organized in two-folds. Firstly, it presents improved CBIR system using a fusion of fast features with varying weighted similarity measure and random forests. This contribution is based on the combination of computationally light weighted colour, and texture features viz. chromaticity moments, colour percentile, and local binary patterns. For searching, inverse variance based varying weighted similarity measure is proposed. A query image classification and retrieval model by filtering out irrelevant class images using random forests is also proposed. In second fold, it presents fast and effective image retrieval based on supervised learning framework with a combination of orthogonal-LBP and statistical moments. This contribution includes supervised learning based image management and retrieval techniques. It utilizes machine learning approaches as a prior step for speeding up image retrieval in the large database with enhanced accuracy. In both works, we have incorporated fast feature indexing and reduced the semantic gap and searching space. Finally, the quantitative performance values for different benchmark databases have been computed for the proposed methods and then compared with state-of-art methods using various evaluation metrics.

Chapter 4 presents a method based on center symmetric-LBP features in wavelet domain using random forests for the classification and retrieval of mammograms. In this chapter, firstly, mammogram clasification framework is introduced. For the classification of mammogram, most relevant multi-resolution CS-LBP texture

characteristics from non-overlapping regions of the mammograms are taken. Using these features with random forests, the test images are classified into different categories having the maximum posterior probability. Finally, using some percentage of known ground truth mammograms we have trained the random forests, and retrieve most relevant images in less searching time. Performance of the proposed method has been analyzed in terms of confusion matrix, classification accuracy, precision, recall, Matthews's correlation coefficient, F-measure, FP-rate, and speedup evaluation metrics.

Chapter 5 presents an automated and effective method for content-based mammogram retrieval using wavelet- based CS-LBP feature and self-organizing map. In this chapter, we have proposed automatic methods for the removal of artifacts and pectoral muscle. For segmentation, selective thresholds based seeded region growing algorithm is introduced. Further, WCS-LBP features are extracted from the segmented region. Then, extracted features are fed to self-organizing map (SOM) which generates clusters of images based on the visual similarities. SOM produces different clusters with their centres and query image features are matched with all cluster representatives to find the closest cluster. Finally, images are retrieved from this closest cluster using Euclidean distance similarity measure. So, at the retrieval time, a query image is searched only in a small cluster, reflects a superior response time and good retrieval performances as compared to conventional exhaustive search method. Performance of the proposed method has been analysed in terms of precision, recall, speedup and searching time evaluation metrics.

Chapter 6 presents conclusions of the thesis and summarizes main findings of this thesis work. This chapter also proposes possible future perspectives of this thesis.