CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

The literature related to feature extraction, feature selection and similarity measurement part of this research work is discussed in this chapter. In the case of feature extraction, the various texture based methods related to the mammogram retrieval work are reviewed. For similarity measurement, the literatures using geometric, statistical and information-theoretic similarity measures are reviewed. The mammogram retrieval related literatures using feature selection techniques are the one referred pertaining to feature selection stage of this research work. The summary of this chapter is given as the concluding part.

2.2 FEATURE EXTRACTION WORKS

Wei et al. [2005] generated gray level co-occurrence matrices (GLCMs) of pixel distance one, three and five in order to estimate the Haralick’s texture features for the retrieval of abnormal mammograms from the MIAS database. Eisa et al [2009] investigated the retrieval of mass and calcification mammograms from the MIAS database using texture and moment-based features [Chang et al, 2006]. In [Felipe et al, 2006], the textural features of a medical database consisting brain, spine, heart, lung, breast, adiposity, muscle, liver and bone images of 11 each are extracted from gray level co-occurrence matrices. The descriptor combining gradient, entropy, and homogeneity performs better than the remaining features. In [Choraś, 2008], for the classification and retrieval of benign and malignant type mammograms in the MIAS database, the Gabor and GLCM based texture features are used in addition to shape features. Sun et al. [2008] proposed texture features based on the combination of distortion constraint and weighted moments for the retrieval of abnormal mammograms from the MIAS database and the
result show that their performance is better than region features. In [Wiesmuller and Chandy, 2010], the gray level aura matrix (GLAM) is used to extract texture information for the retrieval of four categories of mammograms from the DDSM database. In [Quellec et al, 2010], the biorthogonal and orthogonal wavelets within the adaptive lifting scheme framework have been evaluated for the retrieval of mammograms from the DDSM database and is compared with the retrieval methods based on Daubechies 4-tap wavelet (Db4) [Do and Vetterli, 2002] and Zernike moments [Khotanzad and Hong, 1990]. From the literatures it is found that the texture features based on the gray level statistical matrices, Gabor filters and wavelets were generally used for texture based image retrieval. Therefore, in this work these methods are considered as the competitors for comparison with the proposed gray level statistical matrix.

### 2.3 Feature Selection Works

The previous works dealing with the feature selection for content-based mammogram retrieval problem are given in the following. In [Wei, 2005], for the retrieval of six types of abnormal mammograms from mammographic image analysis society (MIAS) database, the t-test was used to select the features out of 11 statistical texture features computed from the gray level co-occurrence matrix (GLCM). There is no mention about the existence of more than one near-optimal feature subset and their selection procedure. In [Felipe et al., 2006], the Statistical Association Rule Miner (StARMiner) algorithm together with the Attribute Significance Estimator based on Fractal Dimension (FD-ASE) was used to select Zernike moments (ZMs) for shape-based retrieval of mammogram masses. Mammograms were pre-processed so that the ZMs could be extracted without prior segmentation, and the same strategy is followed in this research work. However, in [Felipe et al., 2006], the number of images used for the experiment is very less and only one feature set was selected as the near-optimal one. In [Wei et al., 2007], the t-test under relevance feedback was used to select the optimal set
of features from the extracted Gabor-based statistical features for mammogram retrieval and the results were compared with that of the GLCM and Gabor features without feature selection [Manjunath and Ma, 1996]; but, only the macro calcification and micro calcification class mammograms were considered for the experiment. In [Lamard et al., 2007], the Cohen-Daubechies-Feauveau (CDF) 9/7 wavelet features within the lifting scheme framework were used to retrieve images from diabetic retinopathy, mammogram and face databases. The feature descriptor consists of the entire maximum likelihood estimators (MLEs) of the wavelet transform coefficient distribution in each sub-band. Since the feature selection strategy towards finding the optimal MLEs in the context of mammogram retrieval was not considered before, we have introduced this in this research work. In [Sun and Zhang, 2008], texture features based on distortion constraint and weighted moments were employed without the feature selection process for mammogram retrieval, and their performance was compared with region features [Lu and Bottema, 2003] and Gabor features. Five mammograms from each of the six abnormal mammogram classes in the MIAS database were considered in [Sun and Zhang, 2008] as queries for the experiment, which is a small dataset for retrieval problem. In [Quellec et al., 2010], the genetic algorithm (GA) and Powell’s direction set method were used to tune the wavelet basis in the lifting scheme frame framework to retrieve relevant images from face and medical databases. The results were compared with that of the histogram, pyramidal wavelet transform [Do and Vetterli, 2002] and ZMs [Khotanzad and Hong, 1990] methods. The complexity of their work is the search for the optimal wavelet transform to maximize the precision of the system resulted in a large feature descriptor of fixed dimension. From the above discussions, it is clear that there are opportunities to find optimal features for mammogram retrieval.

2.4 SIMILARITY MEASUREMENT WORKS

The literature related to similarity measures and the methods to reduce the computation of similarity measurement are given in the following. Kokare et al., (2003)
have compared nine image similarity measures such as Manhattan (L1), Weighted-Mean–Variance (WMV), Euclidean (L2), Chebychev (L), Mahalanobis, Canberra, Bray-Curtis, Squared Chord, and Squared Chi-Squared distances for texture image retrieval. Chen and Chu (2005) studied the effect of similarity measurement between texture features derived from Gabor and wavelet transforms Db4 and Haar. The evaluation of information-theoretic similarity measures for content-based retrieval and detection of mass type mammograms in the DDSM database is the main focus of [Tourassi et al., 2007].

In [Wang et al., 2009] the k-NN classifier is used in a two stage (semantic and visual) hierarchical framework to test the retrieval of mammograms from DDSM database for 80 queries. In [Quellec et al., 2010a] the performance analysis of image attributes based on the histogram and GGD, and, wavelet basis of orthogonal, biorthogonal and adaptive types for the retrieval of images from digital database for screening mammogram (DDSM) and diabetic retinopathy (DR) are reported. In [Quellec et al, 2010b] the medical cases from the heterogeneous datasets (DDSM and DR) are retrieved using a set of decision trees (DTs) generated based on the randomized c4.5 algorithm. The DTs are used to define a similarity measure between two cases rather than as classifiers. In [Wei et al., 2011], a machine learning approach based on support vector machines and user’s relevance feedback is used to retrieve mass and calcification type mammograms from DDSM database. Image classification, one of the learning techniques is used as a pre-processing step for speeding-up image retrieval in large databases and improving the retrieval accuracy by restricting the similarity measurement to a subset of feature dataset that belong to the same class as predicted for the query. The decision trees, k-NN classifier etc., are examples for image classification techniques [Cover and Hart, 1967]. Even though the decision trees have shown a significant performance in the case of mammogram retrieval, their use is so far limited [Quellec et al, 2007], [Quellec et al., 2010b].
From the above discussion it is clear that there is no fixed distance measure for a retrieval approach. Hence, the retrieval accuracy depends on the selection of distance measure. The computation of similarity measure has significant impact on the retrieval time and this could be solved by considering a smaller subset of feature dataset, which is focussed in this research work. The hybrid features, decision tree and classifier are considered as useful tool towards solving this issue. Therefore, in this research work the aspects like finding suitable similarity measures and computation reduction are considered related to similarity measurement involved in mammogram retrieval.

2.5 CONCLUSION

The texture features based on GLCMs, GLAM, Gabor filter, Daubechies and biorthogonal wavelets (CDF9/7) as well as shape features based on Zernike moments have been used for retrieving mammograms. The above texture feature methods are considered as the competitors in this research work. The gray level statistical matrix based methods are found to be interesting and motivation for the development of a higher order statistical matrix presented in this thesis. Euclidean, weighted mean variance and Kullback-Leibler divergence are the measures commonly used in mammogram retrieval work. A list of distance measures under geometric, statistical and information-theoretic categories are selected for testing its suitability with different feature types. Hybrid features, i.e., image and contextual attributes with decision tree algorithm were used for selecting a subset of the feature dataset for similarity measurement. The idea for selecting decision tree in the case of multiple decision trees is tested for reducing computation for similarity measurement in this research work. The student’s t-test, GA and StARminer with FD-ASE are considered as the competing feature selection methods. A new neighbourhood search method is proposed for selecting the near-optimal feature subsets under various feature types and the details of the same are given in the next chapter. The mammograms available in MIAS and DDSM databases are identified for using in the experimental work.