1-1-2009

Unsupervised learning for biomedical applications

Nasim Shams
Ryerson University

UNSUPERVISED LEARNING FOR BIOMEDICAL APPLICATIONS

Nasim Shams, B.Sc
Isfahan University of Technology, Isfahan, Iran, 2007

A thesis
presented to Ryerson University
in partial fulfillment of the
requirements for the degree of
Master of Applied Science
in the program of
Electrical and Computer Engineering

Ryerson University
Toronto, Ontario, Canada, 2009

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.

Abstract

UNSUPERVISED LEARNING FOR BIOMEDICAL APPLICATIONS

©Nasim Shams, 2009

Master of Applied Science
Electrical and Computer Engineering
Ryerson University

With the growth of application of computers in the generation and analysis of biomedical data, a variety of computerized methods and algorithms have been proposed to optimize the process of acquisition and analysis of the data. Although advanced computerized techniques have provided the means for more precise diagnosis, the interpretation of the recorded data in some cases is an issue due to the large amount of the data or complexity of it.

While most of the existing work in the literature consider supervised techniques for analysis of the collected data, the use of unsupervised techniques in the area of analysis and classification of biomedical signals is relatively unexplored compared to supervised approaches. In general, the investigation of application of unsupervised techniques for analysis of biomedical signals can be worthwhile from different view points. In some cases, biomedical databases tend to contain a large amount of data. Genomic databases or pathological speech databases are examples of this kind. The development of any supervised method for analysis of such databases requires precise manual labeling of the data, which can be extremely costly. However, the use of an unsupervised classifier can be beneficial to accelerate the process and to acquire information about the structure of the dataset. In addition, the characteristics of the collected biomedical data can be affected by the recording process.

In this work application of unsupervised learning in two biomedical signal processing problems is investigated. In the first problem, fuzzy C-means clustering has been used in design of a computer aided diagnosis method for detection of abnormalities in small bowel capsule endoscope images. The performance of the system shows an accuracy of 76which is an acceptable rate for an unsupervised method. In the second case, self organizing tree maps (SOTM) has been applied to audio signal classification for hearing aids. An accuracy of 96% was achieved for discrimination of human voice from the environmental noise, which is one the major classification scenarios for hearing aid applications.

Acknowledgment

It is a pleasure to thank those who made this thesis possible. First, I would like to sincerely thank my supervisor Dr.Sridhar Krishnan for providing me the opportunity to pursue my Master's degree and for his continuous guidance and support throughout the two years of my Masters. I am forever grateful to him.

This thesis would not have been possible without the help of my friends and colleagues Behnaz Ghoraani and Clair Winter who provided me with the features used in Chapter 4 and Chapter 3 of this thesis and Dr.Karthi Umapathy for his invaluable assistance on many occasions.

Also, I would like to show my gratitude to Dr.Matthew Kyan for his enlightening and informative discussions he had with me about the SOTM.

Last but not the least, I would like to thank my family, whom without their constant encouragement I would not be able to accomplish any of this, and my unforgettable friends at Ryerson University, Elnaz shokrollahi, Mehrnaz shokrollahi, Payman Shokrollahi and Sina Zarei for their priceless support and for making the two years of my Master's at Ryerson so memorable.

Thank you...

Dedication

To my parents, Nasrin and Behrouz, for their love, support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

AS the application of computers in the acquisition and generation of medical data is growing, the use of computerized analysis methods in processing the medical data is increasing. Although the use of advanced imaging and recording techniques has provided the physicians with more precise diagnosis, the interpretation of the data is sometimes an issue due to the large amount of data or complexity of it. As a result, a variety of computer based machine learning methods have emerged to assist the doctors to interpret the data and extract more information from the recorded signals. In general, machine learning techniques can be divided into three groups; Supervised learning, unsupervised learning and reinforcement learning [1].

Supervised learning: In supervised learning, a teacher provides a category label or cost for each pattern in a training set, the goal is to reduce the sum of the costs for these patterns.

Unsupervised learning: In unsupervised learning or clustering the category or label of the data is not known beforehand. There is no explicit teacher, and the system forms clusters or natural groupings of the input patterns.

Reinforcement learning The reinforcement learning method is analogous to learning with a critic. In this case no desired category is given for a datum; critic instead, only gives a binary feedback that states whether the tentative category is right or wrong but does

not say specifically how it is wrong.

Unsupervised classification is a natural way to proceed towards computer-aided diagnostic systems and the main motivation of using such scheme is to provide the automatic clustering of the image features in the same way human visual system does. It helps to get an insight about the structures and patterns that already exists in the data and hence enables us to find more robust features, which correspond to the natural characteristics of the data. The use of unsupervised methods might seem unpromising at first. One might even ask the question whether or not it is possible to learn anything of value from unlabeled samples. However, there are many cases where unsupervised classification could be very beneficial. For example, collecting labeled data is not always an easy task. In fact, sometimes labeling a large dataset can be surprisingly costly and not feasible. Unsupervised classification can be used to discover the natural groupings that exist in the dataset and then use supervision only to label the clusters found. Furthermore, in some cases the characteristics of the features change with time. Hence an unsupervised classifier can be used to track the changes and make the necessary corrections. Another application of unsupervised learning is to get some insight about the structure of the dataset. The knowledge about the intrinsic characteristics of the dataset and the patterns that might exist in the dataset, can help us to come up with more efficient feature extraction and classification strategies.

There is a considerable amount of work in the literature on the use of unsupervised techniques for analysis and classification of biomedical signals. Here we discuss the application of some of the popular unsupervised techniques for biomedical signals.

- **Independent component analysis (ICA):** ICA is an emerging field in biomedical signal processing. The wide usage of ICA is motivated by the common practical problem in biomedical signal processing. Recording biomedical signals usually involves several source signals and several sensors. Each sensor receives a mixture of source signals. The problem consists of recovering the source signals from the mixture. In [2] and [3] a combination of wavelet transform and ICA has been used to separate fetal ECG from [4] mother ECG. In [5] Bigan adopts ICA to detect chaothic cardia

arrhythmia in ECG signals. Gao et al. [4] use a combination of ICA and Single value decomposition (SVD) to extract fetal ECG from the mixture signal. In [6] and [7], Joyce et al. and Zhou et al. have used ICA to remove eye blink artifact and power line artifacts from EEG signal. The works done by Navarro et al. [8] and Joshua et al. [9] are more examples of adaptation of ICA for EEG signals.

- **Principle component analysis (PCA):** PCA is a widely used dimensionality reduction technique in data analysis and its popularity comes from three important properties: First, it is the optimal (in terms of mean squared error) linear scheme for compressing a set of high dimension vectors into a set of lower dimension vectors and then reconstructing. Second, the model parameters can be computed directly from the data - for example by diagonalizing the sample covariance. Third, compression and decompression are easy to perform given the model parameters, and require only matrix multiplications. In [10] a PCA based method for ECG-QRS detection has been proposed. Once the QRS complex has been identified, a more detailed examination of ECG signal can be performed. A combination of wavelets and PCA is proposed in [11] for decomposing EMG signals. In [12] and [13] PCA has been used along with neural network and self organizing maps (SOM) for pattern recognition in EMG signals. In [14] original PCA has been applied to the data for classification of cardiac arrhythmias. In [15] a method for clustering analysis of QRS complexes has been proposed that integrates PCA and SOM. Another example of integrating methods for ECG can be found in [16] where PCA and SVM have been used. The main goal is to classify normal from abnormal signals and then specify the kind of abnormality for abnormal signals.

- **K-means clustering:** K-means clustering is one the simplest and most basic clustering techniques, which will be described in Chapter 2. In [17] a k-means clustering technique has been adopted to classify all discrete points forming a heart model with respect to their position vectors or source-to measurement transfer matrices. [18] also

uses k-means clustering for EEG arousal detection.

- **Fuzzy C-means clustering:** Fuzzy C-means clustering is another popular clustering technique that is used widely in pattern recognition problems. This method is very close to K-means clustering and will be described in more details in Chapter 2. In [19] a fuzzy clustering method has been used to classify three types of abnormality. Average period and the pulse width are the features used for classification, and then fuzzy clustering was performed for these two features. The work by Geva and Kerem [20] also utilizes wavelet transform for feature extraction and unsupervised fuzzy clustering for classifying brain-states. In the work by Ajiboye and Weir [21] also fuzzy clustering is used for EMG Pattern Recognition for Multifunctional Prosthesis Control. Finally, in [22] Ajiboye and Weir use fuzzy C-means clustering to classify six major grasping patterns of the human hand.

The use of unsupervised techniques in the area of biomedical signal analysis has been the topic of many research works. In general, the investigation of application of unsupervised techniques for analysis of biomedical signals can be worthwhile from different view points. In some cases, biomedical databases tend to contain a large amount of data. Genomic databases or pathological speech databases are examples of this kind. The development of any supervised method for analysis of such databases requires precise manual labeling of the data, which can be extremely costly. However, the use of an unsupervised classifier can be beneficial to accelerate the process and to acquire information about the structure of the dataset. In addition, the characteristics of the collected biomedical data can be affected by many factors during the recording process. For instance, the recorded EEG signal can be affected by the stress level of the patient or movement artifacts. In the process of recording biomedical data, some patients might need special medications (e.g sedative drugs) or the recording procedure needs to be performed in a modified way due to the special conditions of the patient. Another example is the capsule endoscopy where preparation of the bowel for the experiment is one of the factors that affects the characteristics of the captured images. Hence, images captured during different experiments could posses more or less different

characteristics and this could deteriorate the performance of a supervised classifier. Finally, unsupervised learning methods can be used to get some insight about the structure of the dataset and intrinsic characteristics if the data or can be combined as a preprocessing with a supervised approach to build a robust classifier.

In this work the application of fuzzy C-means and self organized tree maps (SOTM) for the biomedical signals will be examined. These two algorithms will be explained fully in Chapter 2. Fuzzy C-means has been previously applied to biomedical signals such as EEG, ECG etc. It has also been used for segmentation of medical images. However, its application for classification of biomedical images is unexplored. In Chapter 3, fuzzy C-means has been used for classification of abnormalities in the small bowel images. Chapter 4 covers the adaptation of SOTM for classification of audio signals for hearing aid application. SOTM has been used for segmentation of biological images but in this work the application of this algorithm for analysis of biomedical signals will be investigated for the first time. In Section 1.1 and Section 1.2 some of the background information required for Chapters 3 and Chapter 4 are provided respectively.

1.1 Computer methods in medical imaging

Medical imaging is one of the most explosive developments that has taken place in the last two decades. The new findings in this area not only provide a better diagnosis, but also offer new hopes for treatment of many critical diseases. Different imaging techniques such as X-ray, computed tomography (CT) and magnetic resonance imaging (MRI), provide the physicians with a more precise and non-invasive diagnostic tool. For example, for cancer or epilepsy, the precise identification of the lesion already facilitates the use of surgery, the only therapeutic option for some patients. Also, they can provide more accurate diagnosis for some parts of the body which are not easy to evaluate using conventional methods. The small intestine is one of the parts that has been always difficult to evaluate because of its shape and size. Traditional endoscopy used to be the only way for the gastroenterologists to get an insight form the small bowel and detect abnormalities. The procedure is extremely

inconvenient for the patients. During the operation the endoscopic tube, which is rather stiff, is inserted from the mouth and moves around to navigate patient's gastrointestinal tract. The procedure causes a lot of discomfort and the patients are given anesthetics before the operation.

In 2000, a new product was introduced by the Given imaging Ltd that attracted a large amount of interest from the gastroenterologists; the PillCam. PillCam is a tiny capsule endoscope with nearly the size of an ordinary capsule, which has a built in camera. The capsule is ingested from the mouth and as it goes down through the gastrointestinal tract (by the natural movements of the tract) it captures images and sends them wirelessly to a receiver that the patients wears around his/her chest. The capsule is exerted naturally and the patient lives normally during the procedure. Fig 1.1 shows images of the PillCam.

Figure 1.1: Images of the PillCam

1.1.1 Computer aided diagnosis (CAD)

The benefits of the imaging techniques to achieve reliable diagnosis however depends on the quality of image interpretation as well as image acquisition. Computer technology, has made a significant contribution in the quality of interpretation of medical images in the recent years. The use computer-aided diagnosis (CAD) in the area of medical imaging was initiated in the 1960*s* and has increasingly grown since then [23]. Nowadays, CAD is being widely used in detection and diagnosis of many different kinds of abnormalities in medical images. For instance, CAD has become a part of the routine clinical procedure for detection of breast cancer from the mammograms in many hospitals [23].

CAD is a diagnosis made by a radiologist who uses the output from a computer. The computerized analysis of medical images is provided to the radiologist as a second opinion in detecting lesions, assessing extent of disease, and making diagnostic decisions. While the final diagnosis is made by the radiologist, the use of CAD is expected to improve the interpretation component of medical imaging [24]. These are some of the reasons that the use of CAD in the area of medical imaging is growing rapidly.

In addition, interpretation of images by humans can be affected by the presence of structure noise in the image and the presentation of complex disease states requiring the integration of vast amounts of image data and clinical information.

Another benefit of using CAD in the analysis of medical images is to deal with the large amount of data. The interpretation of screening images is a repetitive and tedious task, which involves visual scanning of mostly healthy subjects for a specific abnormality. Screening of mammograms for early detection of breast cancer, the use of CT for detection of lung cancer in high risk individuals and the use of colonography for detection of polyps that may lead to colon cancer are examples of this kind.

It might be useful to emphasize some of the differences between computer-aided diagnosis and another similar concept in this area, automated computer diagnosis [23]. In both approaches, medical images are analyzed by computer algorithms. However there are major differences between the two methods. In CAD, radiologists use the computer output as a second opinion, and make the final decisions. The computer output may be accepted or rejected by the radiologists based on their level of confidence. Furthermore, in this approach even if the performance of the computer is not equal to or higher than that of radiologists, it can be still combined with the radiologist's skills to achieve better diagnosis. With automated computer diagnosis, however, the decision is made by the computer. Thus the efficiency of the processing technique is required to be very high and comparable to that of radiologist's.

1.1.2 CAD for small bowel images

The method developed in this work for the analysis of the small bowel images is designed as a CAD method. Although images captured by the PillCam provide the gastroenterologists with more information about the inside of small intestine, one major drawback of this technology is the large amount of data that is generated in each experiment. During each examination, an average of 50000 images or an equivalent of 8 hours of video is captured. Manual evaluation of such a large number of images is a very time consuming and laborious task and important clues might be missed due to fatigue or repetitive nature of the task. Hence, a CAD method can be developed and used as a second opinion to point out the suspicious regions to the gastroenterologists. The first work on a CAD method for detecting abnormalities in the small bowel images captured by the PillCam was published in 2006 by Khademi et al. [25]. In this work multiresolutional analysis is performed on the gray scale images to extract the texture information and linear discriminant analysis is used for classification of the images. In [26], Li and Meng use color information to detect bleeding in the small intestine. In [27] Bonnel et al. propose a feature extraction method based on wavelet analysis and cross co-occurrence matrices, where the extracted features contain both color and texture information. Canonical discriminant analysis is then applied to the features for classification. In the work by Barbosa et al. [28], the features are extracted from wavelet coefficients and multi layer perceptron (MLP) is used as the classifier. All of the mentioned papers use supervised classification for detecting abnormalities in the images. In this work however, the application of unsupervised classification will be investigated. Although the existing methods with supervised classification typically report higher accuracy rates, the use of unsupervised classification can be advantageous in many ways. The performance of a supervised classifiers depends on the train data. Hence, a wrongly labeled datum, which is not rare in biomedical databases, can affect the overall performance of the classifier. Besides, in order to obtain sufficient reliability, the dataset needs to be large enough to overcome problems such as overfitting and the curse of dimensionality [29]. In addition, characteristics of the images captured by the PillCam, are affected by the bowel preparation procedure.

Colors of intra luminal material may be significantly different between examinations. This leads to different image characteristics and consequently different features for each experiment. A supervised classifier could be biased by the characteristics of the images in the train set. Whereas, under such circumstances, an unsupervised classifier does not suffer from the change of the image characteristics like a supervised classifier does. Finally, the application of unsupervised techniques could be useful to discover clusters that might naturally exist in the data and the features that are related to these groupings.

In this work, a feature extraction scheme similar to the method used in [27] is used, which extracts both color and texture information. A fuzzy C-means classifier is applied to the dataset to find two clusters in the dataset, representing normal (healthy) and abnormal (diseased) images. The results of the unsupervised classification not only can be used as CAD, but also can be used to get more insight about the structure of the data and help find the features that best represent the characteristics of the data.

1.2 Audio signal classification

Audio classification for hearing aids is one the growing areas of application of signal processing and machine learning methods in biomedicine. Although there are a wide variety of hearing aids available, studies show that hearing aid users are not very satisfied with the performance of their hearing aid in the noisy outdoor environments such as restaurants, workplace, street etc [30]. In fact, in a survey performed in [31], low performance in the noisy environments is one of the major reasons that hearing impaired people are reluctant to use their hearing aid devices. Similar studies show that better performance of the hearing aids in the appearance of the noise, is one the most desirable improvements among the hearing aid owners. In order to overcome these problems, several audio processing and classification algorithms have been proposed for the hearing aids to discriminate different auditory classes and detection of the audio environment. In a survey obtained by Kochkin [32] it was observed that a hearing aid that can operate efficiently under different listening conditions is very desirable. From 223 hearing aid users that took this survey, less than one third were satisfied

with their hearing aid if the device worked properly in only three or fewer environments. However, over 91% of the users were satisfied if the hearing aid could be adjusted according to the audio environment. Thus, there is growing evidence that substantially better user satisfaction can be expected of the performance of the hearing aid can be improved.

Audio classification is a one of the research areas that has attracted many researchers in the recent years. Discrimination of different audio classes is one of the tasks that humans do effortlessly everyday. However, implementing such capability in machines is a demanding job and takes a large amount of effort. A large number of papers in the literature is dedicated to various techniques for classification of audio signals for different applications. There is a wide range of applications for the classification of audio signals. Speech processing for security applications and human computer interaction, multimedia data management and distribution, security, biometrics and bioacoustics are some of the applications of audio signals classification [33]. Furthermore, with the growth of application of computerized processing techniques in the area of biomedical signals, the use of audio processing and classification algorithms for biomedical applications such as hearing aids and pathological voice recognition is rapidly increasing.

Various methods have been proposed for discrimination of different audio classes. However, most of the existing works use supervised classification schemes. The proposed solutions include hidden Markov model [34], k-means clustering, histogram driven Bayes classifiers, multilayer perceptrons [35], Gaussian mixture models [36], k nearest neighborhood (K-NN)[37], support vector machine (SVM)[38] and linear discriminant analysis (LDA) [33]. The application of unsupervised methods, on the other hand, is relatively unexplored.

At this point, the results of the unsupervised method can be either presented to the user or can be followed by a supervised approach for further processing.

The works proposed by Shao et al. [39] and Rauber et al. [40] are two examples of application of clustering methods for the music databases. Using a clustering method has the advantage of avoiding the constraints of a fixed taxonomy, which may suffer from ambiguities and inconsistencies. Considering the variety of the audio signals, some of the signals may

simply not fit within a given category [41].

The classification method used in this work is a fusion of supervised and unsupervised classifier. The proposed method in this work is based on the self organizing tree maps followed by a fuzzy labeling of the data. Another important issue in the classification of audio signals is the extracted features. There is a large amount of work in the literature on various feature extraction methods for audio signal classification. The feature extraction strategy depends on the classification scenario and characteristics of the signals. In this work, however, the main focus is on the classification part rather than feature extraction. A brief overview of the existing techniques for audio feature extraction is provided in Chapter 4.

1.3 Organization

In this thesis, the suitability of two unsupervised techniques for biomedical data will be explored. Chapter 3 is dedicated to the application of an unsupervised technique (fuzzy C-means clustering) for detection of abnormalities in the capsule endoscopy images while Chapter 4 describes an unsupervised method (SOTM) for classification of audio signals for hearing aid application. The organization of this thesis is described here;

1. **Introduction:** In the first Chapter, background information on CAD in medical imaging and audio classification for hearing aid application is provided. An overview of the existing works in the literature on the application of unsupervised methods for biomedical signals is also given in this Chapter.

2. **Unsupervised learning and clustering:** An overview of the clustering techniques is provided in this chapter. In addition the two clustering method used in this work is explained in more details.

3. **Unsupervised Learning in Medical Image Classification:** In this chapter the application of fuzzy C-means clustering method for detection of abnormalities in the small intestine images will be described. A feature extraction method based on wavelet

coefficients and cross co-occurrence matrices is used to extract color and texture information of the images. Then fuzzy C-means is applied to the extracted features.

4. **Unsupervised Learning in Hearing Aids Signal Analysis:** In Chapter 4 a classification method based on the SOTM clustering algorithm is used for discrimination of audio signals for hearing aid. The feature extraction technique used in this work is based on time-frequency decomposition of the audio signal, which is more suitable to handle the non-stationary audio signals. A classification technique, which is a fusion of supervised and unsupervised classification is applied to the extracted feature and tested in different scenarios such as discrimination of human/non-human, natrual/artificial and human/music.

5. **Conclusion:** The conclusion for this thesis and the discussion of future works is given in the last chapter.

Chapter 2

Unsupervised Learning and Clustering

2.1 Introduction and Motivation

UNSUPERVISED classification is a pattern recognition technique that aims to construct decision boundaries based on unlabeled dataset. That is, we are interested in exploring the dataset and see what can be done when all we have is a collection of unlabeled samples. Unsupervised classification is also known as data clustering which is a generic label for a variety of procedures designed to find natural groupings, or clusters, in multidimensional data, based on measured or perceived similarities among the patterns [42]. One example of clustering is the detection of a region containing a high density of a specific pattern compared to the rest of the background. Some of the functional definitions proposed for a cluster are:

- Patterns within a cluster are more similar to each other than those belonging to different clusters.
- A cluster, which consists of an area with relatively high density points, is separated from other clusters by an area of relatively low density.

Figure 2.1 shows examples of clusters with different sizes and shapes [43]. The problem of unsupervised classification or clustering is very challenging because data can contain

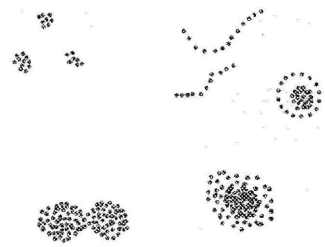

Figure 2.1: Clusters in different shapes and sizes

clusters with different shapes and sizes. Even the number of clusters in the data depends on the resolution with which we view the data. The question that might come to the mind is that why anyone is interested in using unlabeled samples and whether or not it is possible even in principle to learn anything valuable from an unlabeled dataset. There are at least five main reasons for using unsupervised classification [1].

- First, in some cases labeling a large dataset can be surprisingly costly. One example could be the application of land-use classification in remote sensing. In this case obtaining the "ground truth" information for the samples, which is the category for each pixel in an image, requires one to visit the specific site associated with the pixel. Another example is speech classification. Recorded speech is free but accurately labeling it (which is marking the word or phoneme uttered at each time) is extremely time consuming. If a classifier can be crudely designed on a small labeled dataset and then run without supervision on a large unlabeled dataset, much time and trouble can be saved.

- The second advantage of using unsupervised learning is that it makes it possible to proceed in the reverse direction; train with large amounts of inexpensive unlabeled data, and then used supervision only to label the groupings found. This is the case for

large data mining applications where we are dealing with a large dataset with no prior knowledge about the contents of the data.

- Third, in many applications the characteristics of the data can change over time. For example, in an automated food classification problem, the extracted features may change as the season changes. In such case, the performance of the system can be improved by running a classifier in unsupervised mode to track the changes.

- Fourth, unsupervised methods can be used prior to a supervised classifier to improve feature selection. We can use these methods to find more meaningful and discriminatory features that will be used for classification. There are unsupervised methods that represent a form of"smart preprocessing" or"smart feature extraction".

- Lastly, in the early stages of an investigation we can use unsupervised methods to get some insight into the nature or structure of the data. The discovery of distinct subclasses or similarities among patterns or of major departures from expected characteristics may suggest we significantly alter our approach to design the classifier.

2.2 Steps of a Clustering Task

A typical clustering task usually consists of following steps [44]:

1. **Pattern representation (including feature extraction and/or feature selection):** This phase refers to representing the data to the clustering algorithm. The information regarding the number of classes, type and scale of the features are considered in this phase. In this step one can use either the original dataset or use a set of features extracted from the dataset to represent the data. Feature extraction is the process of applying different transformations, decompositions and analysis on the dataset to obtain salient features. In many cases, feature extraction is followed by a feature selection step to identify and choose the most effective feature subset from the original feature set.

2. **Defining (or selecting) a proximity measure:** There are a variety of distance measures defined for measuring the proximity of the points in the dataset e.g Euclidean distance, Mahalanabis distance, Minkowski distance etc. The distance measures will be described in more details in Section 2.4.

3. **Clustering:** Grouping the samples in the dataset can be done in a number of ways. The result of the clustering depends on the type of clustering method used to group the data. The output can be hard (each point belongs to only one cluster) or fuzzy (where each point has a membership value in different clusters) or a nested series of partitions when a hierarchical clustering approach is used. Various clustering techniques will be discussed in Section 2.5.

4. **Data abstraction (optional):** Typically data abstraction is a compact representation of each cluster, usually by using cluster prototypes or cluster centroid.

5. **Cluster validation (optional):** Cluster validation is the assessment of the output of the clustering algorithm. It determines how "good" the clustering results are. All clustering algorithms, when represented with a dataset, produce clusters regardless of whether or not the data actually contains clusters. In those cases where the dataset actually contains clusters, some clustering methods return better results. In order to determine if the groupings found by a clustering algorithm are actually meaningful and evaluate how good or how poor the clusters are, different quantitative measures are developed.

Figure 2.2 shows the block diagram of the first three steps, including a feedback loop where the feature extraction and selection methods can be adjusted based on the grouping results [45].

2.3 Clustering techniques

Cluster analysis is a very useful technique in different areas of pattern recognition. The speed, reliability and consistency with which a clustering algorithm can organize a large dataset has

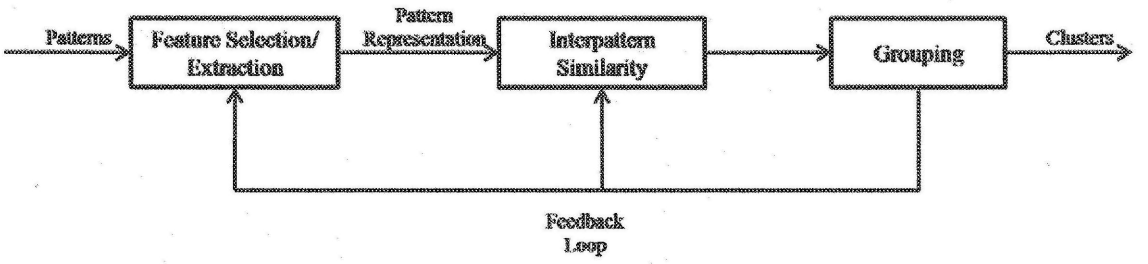

Figure 2.2: Clustering block diagram

led to widespread use of clustering techniques in areas such as data mining [45], information retrieval [46][47] image segmentation [48], signal compression and coding[49] and machine learning. Consequently, numerous clustering algorithms have been proposed in the literature and new ones continue to appear. Classification of the existing clustering methods can be done based on different points of view. Figure 2.3 shows a hierarchical representation of the clustering algorithms [43] [50]. Based on this taxonomy, the algorithms can be divided into two major classes, parametric and non-parametric. Non-parametric approaches, in turn, fall within two groups: Partitional clustering and Hierarchical clustering. The techniques in the first category are mainly based on the popular iterative square-error partitional clustering. These algorithms aim to obtain the partition which minimizes the within class scatter or maximizes the within class scatter. Hierarchical algorithms in the second category are mostly based on the agglomerative hierarchical clustering. These algorithms attempt to organize data in a nested sequence of groups which can be displayed in the form of a dendrogram or a tree.

2.3.1 Partitional Algorithms

Partitional clustering algorithms attempt to obtain a single partition of the data. These methods have the advantage in applications where a large amount of data is to be processed. In such cases, the use of a dendrogram is not computationally feasible. The partitional techniques usually generate clusters by optimizing a criterion function which is defined either

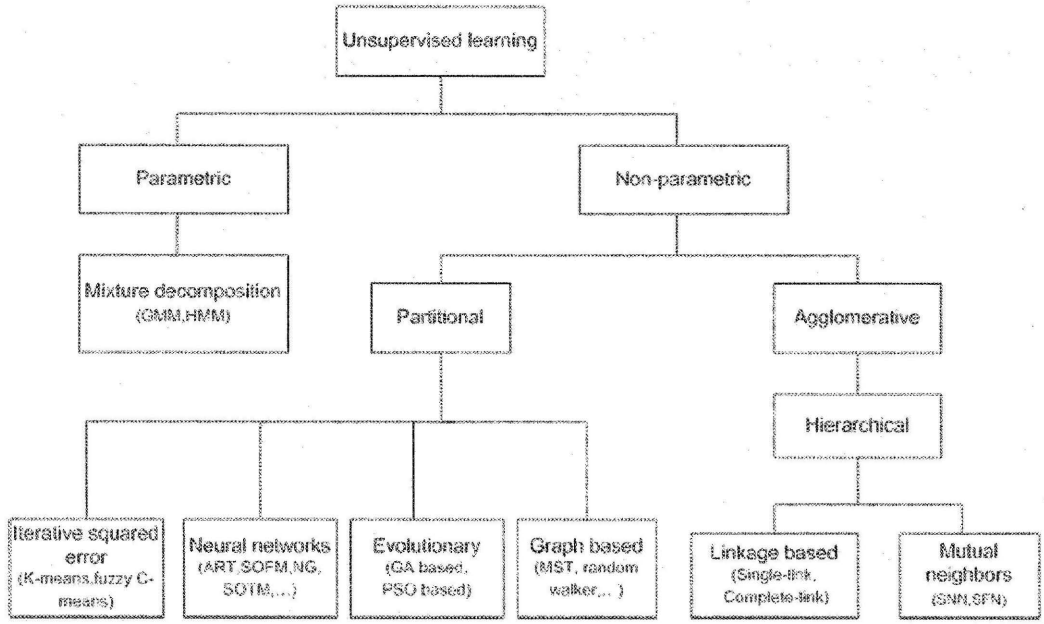

Figure 2.3: Hierarchy of unsupervised approaches

locally or globally. The algorithm is run multiple times with different starting points and the best configuration is then selected as the result of clustering. One of the most popular partitional clustering algorithms is square-error clustering algorithm. The general objective is to find the cluster configuration within the dataset, for which the squared-error is minimum for a fixed number of clusters. The squared-error for cluster C_k is defined as the sum of Euclidean distances between each pattern in C_k and its cluster center m^k. This distance is also called the within-cluster variation.

$$e_k^2 = \sum_{i=1}^{n_k} \left\| x_i^k - m^k \right\| \tag{2.1}$$

Where x_i^k is the ith pattern belonging to cluster C_k, n_k is the number of patterns in the cluster C_k and m^k is the mean, or center of the Kth cluster defined as

$$m^k = \left(\frac{1}{n_k}\right) \sum_{i=1}^{n_k} {x_i}^k \tag{2.2}$$

The overall squared-error for a configuration is the sum of the square-error for all clusters described as:

$$E_k^2 = \sum_{k=1}^{K} e_k^2 \tag{2.3}$$

The objective of the squared-error algorithm is to find the cluster configuration that minimizes the total square-error for a fixed number of clusters K. The resulting partition has also been referred to as the minimum variance partition.

The k-means clustering is one the simplest and the most popular square-error algorithms. The algorithm is computationally efficient and gives good results on a dataset that consists of compact and well separated clusters with a hyperspherical shape [43]. The algorithm is even able to detect hyperellipsoidal clusters if the Mahalanobis distance is used in 2.3 in defining the squared-error. The following briefly explains the algorithm steps [1]:

1. Begin with K initial cluster centroid.
2. Classify the n samples according to the nearest distance.
3. Recompute the cluster center for each cluster. If the new cluster centers are the same as previous ones, there is no need to recalculate the centers again. The current cluster centers are the final ones. Otherwise, go back to step 2 and classify the points with the new cluster centers.

A big drawback of the algorithm, however, is the lack of a guideline to select the critical parameters such as the number of clusters and the initial cluster centers [51]. Several variations have been proposed to improve the performance of the basic k-means algorithm. One of the possible modifications is to introduce a fuzzy criterion function. This results in fuzzy c-means algorithm, which will be described in the next subsection.

2.3.2 Fuzzy C-means Clustering

In the traditional clustering approaches, each pattern belongs to one and only one cluster. This type of clustering is called *hard* clustering. In contrast to hard clustering methods,

fuzzy clustering methods assign a degree of membership in each cluster to each pattern. A fuzzy clustering algorithm can be converted to a hard algorithm by assigning a pattern to the cluster with the largest degree of membership. The steps involved in performing a fuzzy c-means algorithm is very close to that of k-means, except for the objective function, which is defined as

$$J_m = \sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij})^{m}\|x_j - c_i\|^2, 1 \leq m \leq \infty \tag{2.4}$$

Where m is the fuzziness index, μ_{ij} is the degree of membership of observation x_j in the cluster i, x_j $(j = 1, 2, ..., N)$ is the jth d-dimensional data point and C_i is the d-dimensional center of the cluster.

The fuzzy set Theory was initially applied to data clustering by Ruspini [44]. Although the results of the algorithm is better than the hard k-means algorithm, FCM can still converge to the local minima of the squared-error criterion function.

2.4 Neural Network Approaches

Artificial Neural Networks (ANN) has been widely used in pattern recognition applications in both supervised and unsupervised ways. ANN approaches typically fall into two groups:

- The first group are those based on competitive learning or learning vector quantization [50]. In competitive learning similar patterns are grouped together by the network represented by a *neuron*. This grouping is done based on correlation among the data. In unsupervised context, well-known example of ANN are the Kohonen's self-organizing map (SOM) and adaptive resonance theory proposed by Carpenter and Grossberg in 1990 [50]. The architecture of these networks are single-layered. Patterns are represented to the input layer and associate to the output layer. The weights between the input and output layers are updated iteratively until a termination criterion is fulfilled. This group of algorithms will be discussed in more details shortly.

- The second group are techniques derived from the principle component analysis (PCA), factor analysis and independent component anlysis (ICA)[52].

2.4.1 Self-Organizing Architectures

Self-Organizing methods are closely related to unsupervised learning. A number of self-organizing architectures are: the Kohonen self-organizing feature map, neural gas approaches, hierarchical feature map, dynamic hierarchical architectures, non-stationary architectures and hybrid architectures [50]. The self-organizing technique used in this work is self-organized tree mapping, which is a derivation of the Kohonen self-organizing map and will be the focus of this Section. Figure 2.4 shows the hierarchy of different self-organizing methods.

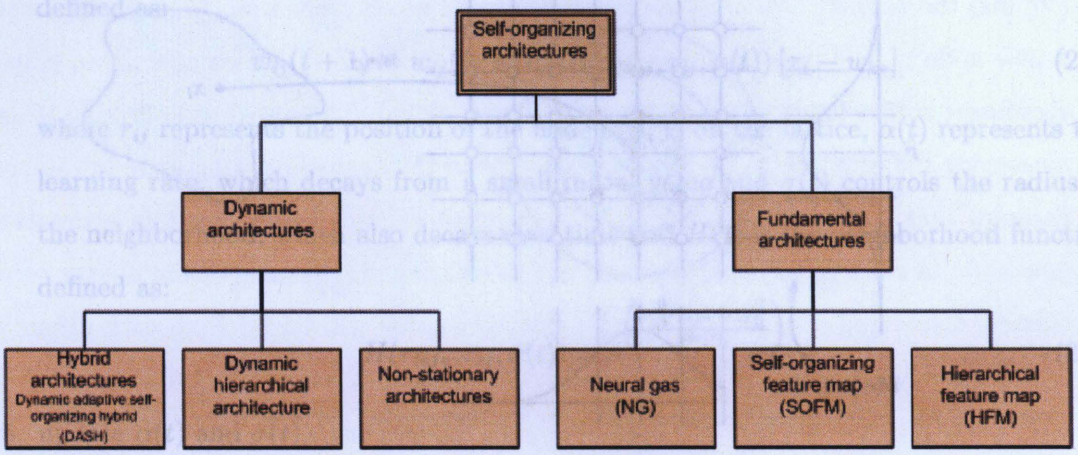

Figure 2.4: hierarchical representation of self-organizing approaches

2.4.2 The Kohonen Self-Organizing Feature map (SOFM)

In the basic SOFM algorithm, input samples from a d dimensional feature space, are mapped onto a grid with lower dimensions (usually two or tree dimensional) [50]. Each node on the grid acts like a memory element; it stores the prototype vector that describes commonly occurring vector patterns from the input space. The points that are close to each other in the input space are mapped onto the neurons that are nearby in the grid. Whenever a node is updated, the nearby nodes are also updated based on their distance from the original winning node. Figure 2.5 shows the mapping of samples onto an SOFM lattice. The steps

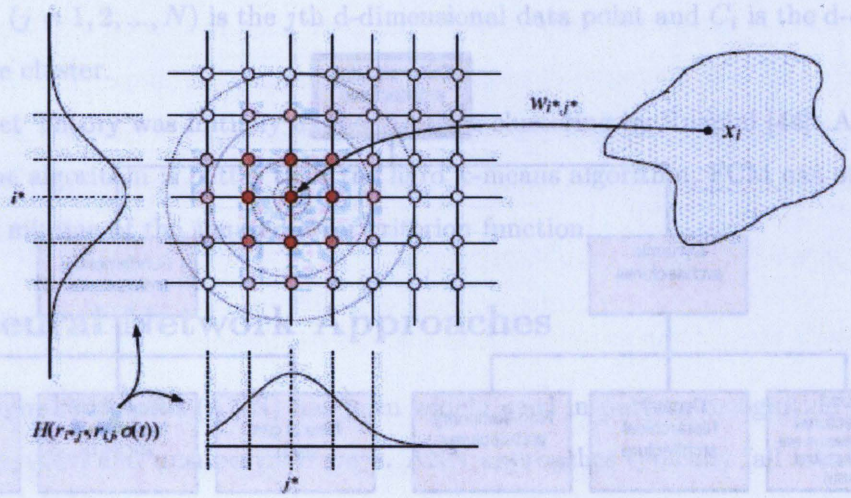

Figure 2.5: Mapping samples from the input space onto the SOFM lattice: The input x_i is assigned to the winning node. The neighbors that are connected to the winning node in the lattice are updated according to the gaussian neighborhood function (courtesy of M.Kyan)

involved in the SOFM algorithm are:

1. Initialize the weight vector w_{ij} of each neuron in the lattice using a random value. This random value can be a sample randomly selected from the dataset X.

2. randomly select an input vector x_i from the dataset and present to the network.

3. choose a winning node w_{i*j*} based on the minimum Euclidean distance.

4. update the neurons on the lattice according to a Gaussian neighborhood function defined as:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)H(r_{i*j*}, r_{ij}, \sigma(t))\,[x_i - w_{k*}] \tag{2.5}$$

where r_{ij} represents the position of the node at(i, j) on the lattice, $\alpha(t)$ represents the learning rate, which decays from a small initial value and $\sigma(t)$ controls the radius of the neighborhood, which also decays over time and $H(t)$ is the neighborhood function defined as:

$$H(r_{i*j*}, r_{ij}, \sigma(t)) = e^{\left(\frac{-\left\|r_{i*j*} - r_{ij}\right\|}{2\sigma(t)}\right)} \tag{2.6}$$

5. update $\alpha(t)$ and $\sigma(t)$

6. repeat iteration from step 2 until there is no significant change in w_{ij}

Association between the nodes is an important advantage in SOFM that helps the evolution of the network can be useful for extracting inter-clusters relationships. This property is useful for visualization of multivariate data, where data with high dimension is mapped onto a two dimensional lattice. Since the mapping preserves the topology, neighbor nodes in the lattice represent the samples with related properties in the original data [50].

2.4.3 Self-Organizing Tree Map (SOTM)

SOTM was originally introduced in [53] to remove impulse noise from images. The algorithm is a hybrid of the traditional SOFM ,which was explained in the previous chapter and the Adaptive Resonance Theory (ART) [54]. Like ART, the growth of the network is controlled

by a vigilance test, which essentially watches for an input that is in contrary to the current knowledge about the input feature space. If such an input is found the resonance occurs and results in refinement of the winning node or generation of a new node. On the other hand, like the SOFM the generated network is more topologically aware and the refinement of the existing prototypes is guided by a Kohenen style learning rule. Like its counterpart SOFM, the SOTM algorithm uses competitive learning approach to find clusters within the data while maintaining the general topology of the feature space. However, unlike the SOFM, SOTM does not suffer from the disadvantage of nodes being trapped in the low density areas [50] and the network has a dynamic structure and grows from a single node. Generation of a new node is guided by a hierarchical control function $H(t)$, which acts as an ellipsoid of significant similarity. $H(t)$ can be assumed as a global vigilance threshold that is used for measuring the proximity of a new input sample to the nearest existing node in the network. Samples that fall outside the scope of the nearest existing node, result in generation of a new node as child of the winning node. By initializing $H(t)$ to start from a large value, the clusters discovered at the early stages of the clustering will be far from each other. Decay of $H(t)$ over time results in partitioning the data space in low resolution at the early stages of the clustering, while favoring partitioning at higher resolutions later. Figure 2.6 depicts the clustering process in SOTM and SOFM.

The SOTM Algorithm

The steps involved in the basic SOTM algorithm are:

1. Initialization: randomly select a training vector from the feature space X. Initialize the network parameters $H(0)$ and $\alpha(0)$

2. randomly select an input x from the feature space and calculate the distance d_j from x to all currently existing neurons $w_j (j = 1, ..., N_c)$ when N_c is the total number of currently existing neurons.

3. select the node with the minimum distance as the winning node w_k such that $d_j(x, w_k) = min_j d_j(x, w_j)$

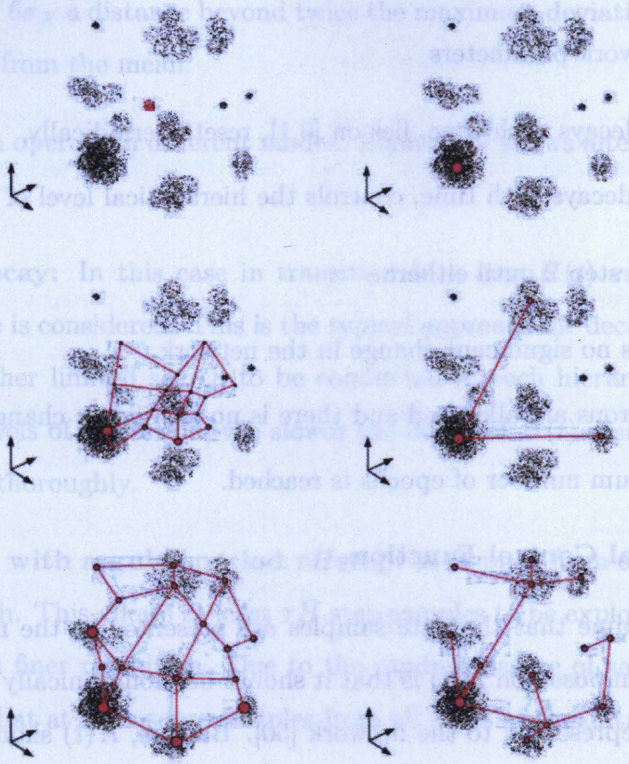

Figure 2.6: Clustering procedure in the SOTM(left) vs the SOFM(right). The SOFM uses a predefined lattice to span the input steps and assigns the samples to the closest node, or the *winning node*. The input is used to update the winning node and its immediate neighbors in the lattice. The SOTM (right), on the contrary, explores the input space by a growing structure in a top-down manner. As it can be seen in the figure, unlike the SOFM, the SOTM does not suffer from the nodes begin trapped in low density areas. (courtesy of M.Kyan)

4. if $d_j(x, w_k) \leq H(t)$, then: update the weight vector of the winning node using the reinforced learning rule: $w_k(t+1) = w_k(t) + \alpha(t)[x - w_j]$ where $\alpha(t)$ is the learning rate and $H(t)$ is the hierarchical control function.

5. alternatively if $d_j(x, w_k) > H(t)$, then spawn a new node at the position x from the winning node w_k.

6. update network parameters

   - $\alpha(t)$: decays with time, lies on [0,1], resets periodically.
   - $H(t)$: decays with time, controls the hierarchical level of the tree.

7. repeat from step 2 until either:

   - there is no significant change in the network.
   - all neurons are allocated and there is no significant change in the network.
   - maximum number of epochs is reached.

The Hierarchical Control Function

In general we assume that the date samples are presented to the network randomly. The only constrained imposed on $H(t)$ is that it should be monotonically decreasing over time as the samples are represented to the network [50]. Besides, $H(t)$ should be ideally initialized to a large enough value to cover the span of the data. There are two standard hierarchical control function proposed for the original SOTM algorithm: linear and exponential decay.

$$H(t) = H(0) - [(1 - e^{-\xi/\tau H})H(0)/\xi]t \tag{2.7}$$

$$H(t) = H(0)e^{-t/\tau H} \tag{2.8}$$

where τH is a time constant, which is bound to the projected size of the input data X, $H(0)$ is the initial value, t is the number of iterations (or sample presentation) and ξ is the number of iterations over which the linear version of $H(t)$ would decay to the same level

as the exponential version. One benefit of initializing $H(t)$ to a large value, possibly larger than the maximum variation within the data, is that all levels of resolution across the data can be explored. There are two natural choice proposed for $H(0)$ [50]:

Range-based H(0): Σ (all ranges across the dimensions of X)

SD-based H(0): $> 6\sigma_X$ a distance beyond twice the maximum deviation ($r = 3\sigma_X$) of 99% of the samples from the mean.

In addition, $H(t)$ can operate in different modes. Figure 2.7 shows alternative strategies for decay of $H(t)$ [50].

1. **Pure $H(t)$ decay:** In this case in transition to a new $H(t)$ value, only one single random sample is considered. This is the typical approach for decay of $H(t)$. However, it results a rather limited search to be conducted in each hierarchical level. In fact, in the lower levels of resolution, the slower the decay of $H(t)$ means the data is being assessed more thoroughly.

2. **Stepped $H(t)$ with regular period τHstep:** A stepped form of decay is introduced in this approach. This allows at least τH step samples to be explored before narrowing the search to a finer resolution. Due to the random nature of sample representation, it is assumed that at least some samples from all parts of the data are explored in this period.

3. **Stepped $H(t)$ with irregular period:** This mode is in fact an extension of mode 2. In this mode the counter is begin reset every time a new node is generated. This guarantees that the search will continue for at least another $\tau H step$ samples after a new node is generated. This allows a new node to have a chance to adjust itself.

4. **Stepped $H(t)$ with irregular period and node inhibition:** This mode adds an additional constrain to mode 2 by forcing the network adaptations only for a period of $\tau H step$ before inserting a new node. This allows nodes to organize and have sufficient

time to adjust themselves before new (and possibly unnecessary) nodes are allocated. This process repeats every time a new node is spawned and gives the network a period of time to settle before generating a new node.

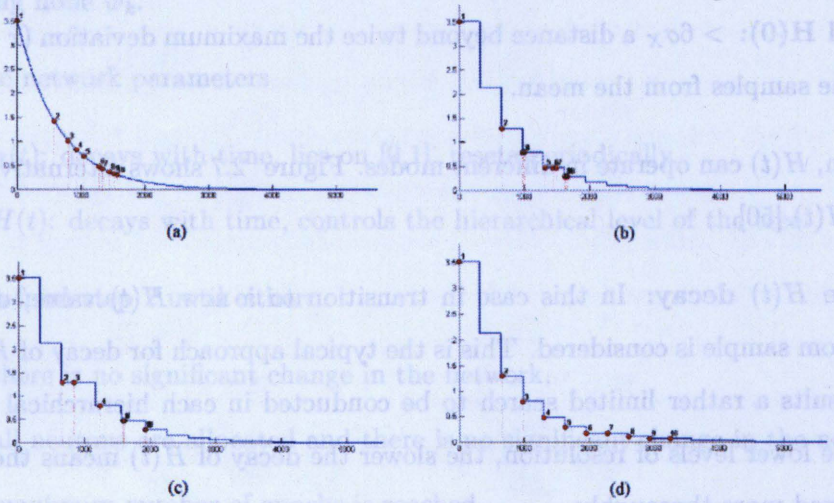

Figure 2.7: Different H(t) decay strategies illustrated for period of generation of 10 neurons. (a) Pure H(t) decay; (b) Stepped $H(t)$ with regular period; (c) Stepped $H(t)$ with irregular period; (d) Stepped $H(t)$ with irregular period and node inhibition. (courtesy of M.Kyan)

Learning Rate

The learning rate $\alpha(t)$ is an important factor in organizing the network. Like the hierarchical control function, $H(t)$, $\alpha(t)$ can also operate in number of different global or local modes. In global modes a single learning rate is applied to all node, whereas in local modes an individual rate operates for each node a set of nodes. There are a few modalities proposed for the operation of the learning rate. Some of these modes are discussed below. The first mode is the original periodic reset strategy proposed for the SOTM. Modes 2-4 are the new approaches suggested in [50]. However, it has been mentioned in [50] that Modes 1 and 2 are noticed to have better results for an SOTM process.

- **Global periodic reset:** In this traditional approach the network memory is refreshed with regard to the underlying density.

- **Global reset upon node generatioin:** This approach is a modification of the first mode based on the idea that a network needs to reorganize its memory only when a new node is generated.

- **Local rest of winner and child upon node generation:** This modification restricts the plasticity only to the region of the map which is recently grown. based on the assumption that the adjustment of the nodes that are distant from the growing region is not necessary.

- **Local reset of winner, child and siblings upon node generation:** This mode is very similar to mode 3, with the exception that children of the winning node are also considered to be plastic within the updating region. As mentioned in [50], the global reset modes (1,2) tend to outperform the local reset modes. In addition, it is suggested that mode 2 is preferred because the reset is justified when new information is to be induced to the network after node generation.

Chapter 3

Unsupervised Learning in Medical Image Classification

3.1 Small Intestine Images

MEDICAL imaging is certainly one of the most explosive developments that has taken place in the last two decades. The new findings in this area not only provide a better diagnoses, but also offer new hopes for treatment of many critical diseases. Different imaging techniques such as MRI and x-ray provide the physicians with a more precise non-invasive diagnostic tool. Different medical imaging techniques are complementary and their progress has immediate repercussion on the development of treatments as they provide a much less invasive diagnosis compared to previous methods. For example, for cancer or epilepsy, the precise identification of the lesion already facilitates the use of surgery; the only therapeutic option for some patients. Also, imaging techniques can provide more accurate diagnosis for some parts of the body which are not easy to evaluate using conventional methods. The small bowel for example has always been difficult to evaluate because of its shape and size. Traditional endoscopy used to be the only way to gather actual images from inside the patients intestine. The operation needs to be performed by highly skilled doctors and is inconvenient for the patients. The endoscopy's tube, which is inserted from the mouth, is rather stiff and causes some discomfort as the doctor navigates the patient's gastrointestinal tract. In addition, since the camera cannot reach all parts of the small

intestine, diagnosing diseases of the small intestine was a major problem for doctors[55]. The appearance of capsule endoscopy in 2000 has generated a large amount of interest among gastroenterologists. PillCam is a tiny capsule (10mm × 27mm)[56], which was introduced by Given Imaging Ltd. The capsule is digested from the mouth and moves slowly through the gastrointestinal tract (including the small intestine) by a dint of natural contractions. As the capsule moves through the gastrointestinal tract, it captures color images and transmits them wirelessly to a receiver that the patient wears around his or her waist [25]. The capsule is exerted naturally with the natural bowel movements [25]. The data collected through the examination is an 8-hour-long video that provides visualization of the 21 foot long small bowel, which used to be a "black-box" to doctors [25]. The procedure is ambulatory and enables the patient to live normally during the endoscopic examination. Clinical results show that PillCam is a superior diagnostic method for detecting the diseases in the small intestine [55]. Four main types of cancer, which are usually found in the small intestine are listed and described below [25].

Adenocarcinoma: This type of cancer originates in the epithelial lining of the mucosa and is mainly found in the duodenum.

Sarcoma: This cancer originates in the muscle wall of the small intestine and is mostly found in the ileum.

Carcinoid: This type of cancer originates in the specialized neuroendocrine cells are found in the small intestine, the ileum and sometimes in the appendix.

Lymphoma: This type of malignancy is usually formed within the lymphoid tissue of the small bowel. They are commonly found in the jejunum or ileum.

The PillCam provides gastroenterologists with a new method for detection of the small bowel diseases through a live video representation, which was not available with the traditional endoscopy. However, the drawback of this technology is the large amount of data which is collected in each experiment. An average of 50000 images or 8 hours of video is

recorded during an examination. Manual evaluation of these images is a an extremely laborious and time consuming task and important clues might be missed due to fatigue or repetitive nature of the job [57]. Therefore, a computer aided diagnostic method can be developed and used as a secondary opinion, that views and points out the suspicious areas to the doctor. Figure 3.1 shows sample images taken by the PillCam, which includes three normal and three abnormal images.

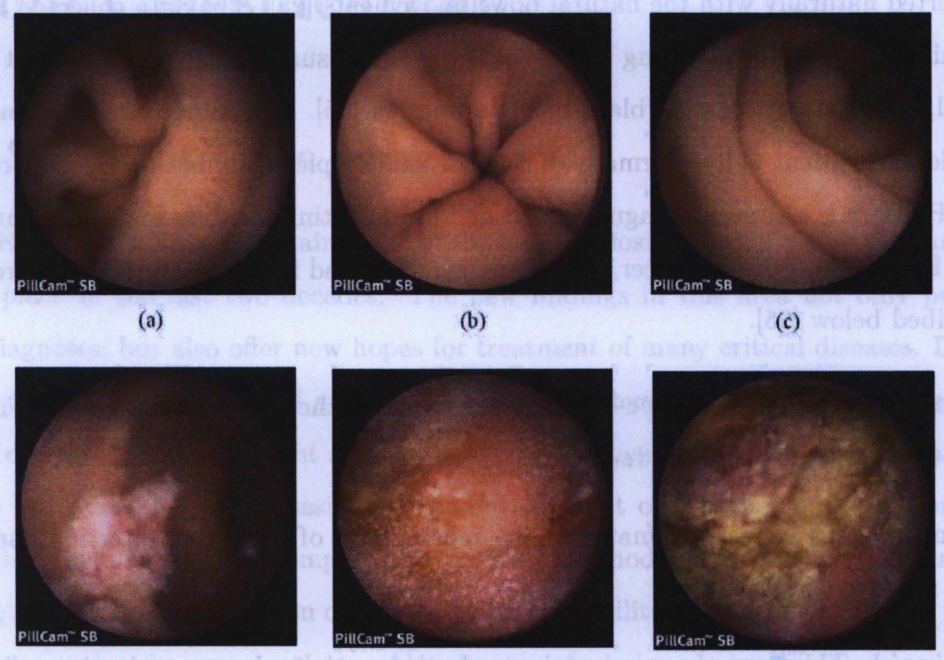

Figure 3.1: Sample small bowel images collected by the PillCam obtained from the Image Atlas of Given Imaging Ltd. (a) Healthy small bowel, (b) normal pyloric region, (c) normal jejunum, (d) small bowel polyp, (e) small bowel lymphoma, (e) small bowel lymphoma

In addition, a computer aided system can be used to confirm and compliment the doctor's diagnosis. It can help to decrease the number of required biopsies, detect cancer in an early stage, and in general improve the quality of diagnosis [25]. The first work on the automatic

detection of abnormalities in the capsule endoscope images was proposed by Khademi et al. [25], where linear discriminant analysis has been uses for classification. Other classification techniques used in this area include canonical discriminant analysis [27] and MLP [28][26]. While all the previous works have used supervised techniques, in this work the application of an unsupervised method will be explored.

The investigation of application of unsupervised approaches for medical images can be useful from different viewpoints. One of the reasons for considering an unsupervised method is that image characteristics might vary in different experiments. One of the most important features of the capsule endoscopy procedure is the bowel preparation. Colors of intra luminal material may be significantly different between examinations. Therefore, the characteristics of the images used for training the supervised classifier might be different from those of images captured in the test experiment and this could affect the performance of a supervised classifier. However, the performance of an unsupervised classifier does not depend on the characteristics of the training and test data. Another important factor to be considered is the size of dataset. Although the ground truth for the image dataset is given in this case, in order to build a robust supervised classifier the dataset has to be large enough to guarantee good generalization. In addition, unsupervised techniques can be used to get some insight about the structure of the data and existence of the natural patterns, discovery of distinct subclasses or similarities among patterns and to find meaningful and discriminatory features that best represent natural groupings in the data.

3.2 Feature Extraction

Like almost any other classification problem, the first step in the classification of small bowel images is extracting a set of descriptors from the images that can efficiently represent characteristics of the images and have high discriminatory power. The extracted features are then fed to the classifier, which is unsupervised in this case, to make the decision. The outcome of the classifier is related to the diagnosis of the images, which can be either a normal (healthy) or an abnormal (diseased) image. In addition, since the input space consists of images the

input data is expected to have a very high dimensionality (256×256 in this case). Performing any classification method on data with such high dimensions would be extremely costly and computationally intensive. Hence, the need for a feature extraction scheme becomes more significant. Figure 3.2 shows the feature extraction procedure performed in this work. The Images are first converted into CIE lab color space, then shift invariant wavelet transform is performed on the images, and then cross co-occurrence matrices are calculated on the wavelet coefficients. Each of the blocks will be described in more details shortly.

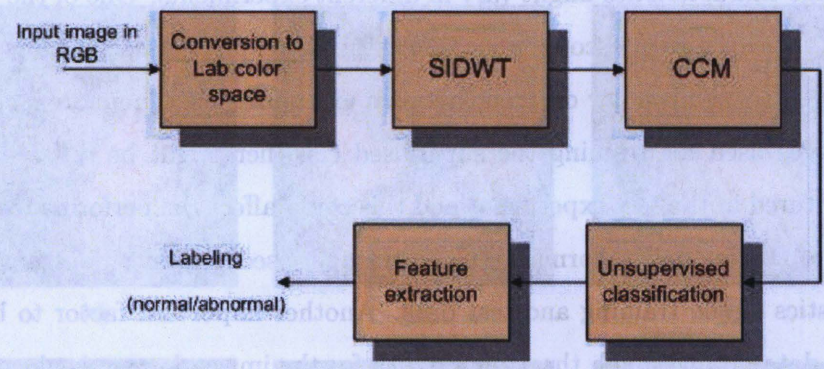

Figure 3.2: Block diagram of the feature extraction procedure

The two main features used in this work are color and texture. These features are directly related to the clues used by the doctors to evaluate the images. Texture is one of the important clues in analyzing both color and gray scale medical images. The human visual system can discriminate different textured areas in an image effortlessly. However, implementing this task on computers has been the subject of research in the area of machine vision for a long time.

Texture: The images captured by the PillCam are from different organs, structures and anatomical objects along the gastrointestinal tract. It can be noticed from the experimental dataset that normal images contain mostly smooth and homogeneous texture with very little disruptions in uniformity except for folds and crevices [25]. On the other hand, abnormal images tend to contain different textures at the same time and

more heterogenous textured areas. This can be seen in Figure 3.3.

Color: Color contents of an image also provides discriminatory information about the region or objects in the image. Normal regions usually exhibit pinkish colors, whereas abnormal regions show some difference in color compared to the surrounding area. Malignant tumors are usually inflated, more reddish and severe in color compared to normal areas while benign tumors show less intense hues. Redness may specify bleeding, blackness could be treated as deposits due to laxative, green may be the presence of fecal materials and yellow relates to pus information of the image [58].

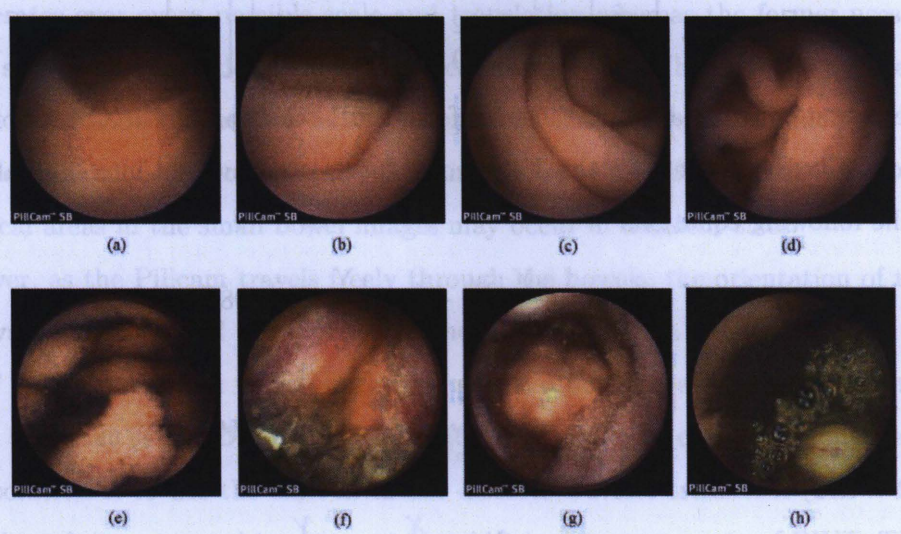

Figure 3.3: Top row: normal small bowel images. (a) normal small bowel, (b) normal jejunum, (c) normal jejunum, (d) normal small bowel. Bottom row: abnormal images. (e) small bowel polyp, (f) small bowel lymphoma, (g) polypoid mass, (h) GIST tumor.

3.2.1 CIE Lab Color Space

As explained in [59], abnormal regions are observed to show more or less differences in color compared to the surrounding regions. In fact, malignant tumors are usually inflamed, reddish and more severe in color. Hence, color information plays an important role in the

detection of abnormalities in small bowel images. The images taken by the PillCam are compressed in JPEG and coded in RGB color space. However, in this work the feature extraction is performed in the CIE lab color space. Unlike the RGB, the Lab color space is designed to approximate the human vision. The main advantage of using lab color space is that this color space is perceptually uniform, which means a change in the color value results in a change of about the same visual importance. In addition, Euclidean distance measure has a better performance in this color space. The L component defines the luminance, a is red/blue chrominance, and b is yellow/blue chrominance. The equations for converting the RGB color space to the Lab color space are given below [59]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412 & 0.357 & 0.180 \\ 0.212 & 0.715 & 0.072 \\ 0.019 & 0.119 & 0.950 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3.1}$$

The first step is to transform the color form RGB color space into XYZ color space using Equation 3.1. Next, values in the XYZ color space are converted into the Lab color space using the following equations

$$L = 116(\frac{Y}{Y_n})^{1/3} - 16 \text{for} \frac{\text{Y}}{\text{Y}_\text{n}} > 0.008856 \tag{3.2}$$

$$L = 903.3(\frac{Y}{Y_n}) \text{ for} \frac{Y}{Y_n} \leq 0.008856 \tag{3.3}$$

$$a = 500(f(\frac{X}{X_n}) - f(\frac{Y}{Y_n})) \tag{3.4}$$

$$b = 500(f(\frac{Y}{Y_n}) - f(\frac{Z}{Z_n})) \tag{3.5}$$

where

$$f(t) = \sqrt[3]{t} \text{ for } t > 0.008856$$

$$f(t) = 7.7787t + \frac{16}{116} \text{ for } t \leq 0.008856$$

where X_n, Y_n, and Z_n correspond to the white color in the XYZ color space and L, a and b are the luminance and chrominance in the Lab color space respectively.

3.2.2 Shift Invariant Discrete Wavelet Transform

In the previous works on the classification of the small bowel images [27][25], the application of the shift invariant discrete wavelet transform has been investigated and has been proven to be efficient for extracting the texture information in particular. Multiresolutional analysis of the images is a natural way to highlight the features of interest, such as texture, in an image. It provides a representation of the image in which the textural information can be retrieved easily. This method is basically a projection of the images onto a set of finite-length and fast-decaying oscillating functions known as wavelets. Wavelet transforms can be classified into discrete wavelet transforms (DWT) and continuous wavelet transforms (CWT). The latter operates over every possible scale and translation whereas the former uses a specific subset of scale and translation values or representation grid. The DWT is a scale-invariant transform since a decomposition of the image contains all the basic functions needed to decompose different scales of the image. This feature of the DWT is of importance since pathological areas in the small bowel images may occur in different sizes.

However, as the Pillcam travels freely through the bowels, the orientation of the images is not always the same and the location of the suspicious areas is unpredictable. However, the DWT is a shif-invariant transform, which means different translations of an input image results in different set of DWT coefficients [60]. In order to extract a consistent feature set, one solution is to use the shift-invariant discrete wavelet transform (SIDWT). Several solutions have been proposed to overcome the shift-invariant property of DWT. The method suggested by Mallat et al is based on selecting the local exterma from the filtered and fully sampled version of the image. These local extermas are used to detect and translate the shift since a shift in the signal results in a shift in the local extermas. However, due to lack of decimation there is a large amount of redundancy and each level of decomposition has as many samples as the original input image, which makes the algorithm to be costly overall. One solution for the cases where the dictionary contains many redundant wavelet basis functions is the Matching Pursuit (MP) algorithm. However, this algorithm is computationally intensive itself and can slow down the the system. Bradley proposes a method

which is a trade off between the sparsity of representation and time invariance where critical sampling is performed for certain subbands only and the rest are fully sampled. The result of this method is an approximate SIDWT. The mentioned algorithms either suffer form high computational cost or achieve only an approximation of SIDWT. The SIDWT algorithm proposed by Beylkin does not have the discussed shortcomings. It calculates the DWT for all circular shifts in a computationally efficient way. In addition, since this transform uses orthogonal basis, it results in less redundancy. An extension of Beylkin's algorithm to 2-D signals is developed by Lian et al. The application of this algorithm to the biomedical images is shown to give promising results in the previous works [25][27].

The algorithm proposed by Liang and Parks in [61] is used in this work to decompose the images in the wavelet domain. In fact, this algorithm is an easy and fast implementation of multiresolutional analysis using filterbanks. It makes for a good localization for high frequencies and a good frequency precision for low frequencies.

The $2-D$ filterbank scheme used for an $N \times N$ image applies a high pass filter on the image followed by a low pass filter. Applying the low pass filter $H_0(z)$ and then the high pass filter $H_1(z)$ to each row of the image X creates two images: one containing the low frequencies of X, $X(L)$ and the other one containing the high frequencies $X(H)$. The rows, and $X(L)$ and $X(H)$ are subsampled by a factor of 2, then the same filters H_0 and H_1 are applied to the columns of each image. Finally another subsampling by 2 is performed on the columns. The result, as depicted in Figure 3.5, is four images LL, HL, LH and HH for two levels of decomposition. The same procedure is repeated for further decomposition. The high pass filters applied in the horizontal and vertical directions in this scheme emphasize the high frequency contents of the image and give oriented: The $HH, HL,$ andLH subbands represent the diagonal, horizontal and vertical edges respectively. The 5/3 Gull wavelet has been used in this work as used in [62] because the filter lengths are small and can warrant an efficient implementation. In order to be invariant to translations, the algorithm should look at all translations of the input image and select the best set of wavelet coefficients. The procedure consists of two parts, first, an efficient algorithm for computing the wavelet

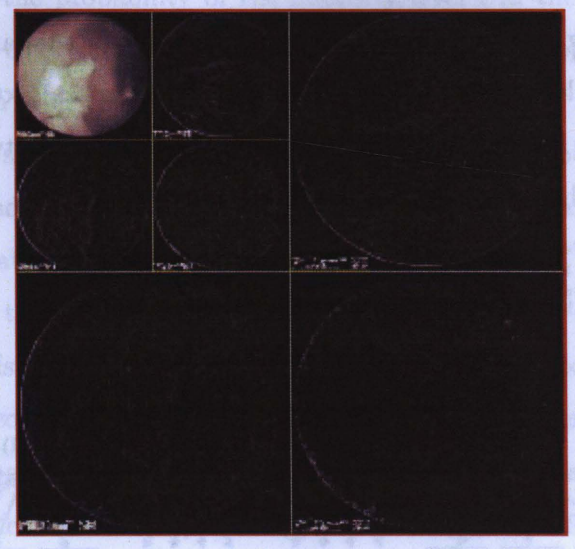

Figure 3.4: Wavelet coefficients for two level decomposition of a small bowel image

transform for all the translations and second a fast quadtree search algorithm. The wavelet decomposition is performed for different shift values. There are four elementary shifts in this algorithm: $(0,0),(0,1),(1,0)$ and $(1,1)$ where the first index corresponds to the row and second index corresponds to the column. Every shift can be represented as a combination of these elementary shifts. So the j_{th} level of decomposition for the input shift (a,b) can be obtained by [25]

$$LL^j(a,b) = \sum_m \sum_n h_0(m-2a)h_0(n-2b)LL^{j-1}(m,n) \tag{3.6}$$

$$HL^j(a,b) = \sum_m \sum_n h_0(m-2a)h_1(n-2b)HL^{j-1}(m,n) \tag{3.7}$$

$$LH^j(a,b) = \sum_m \sum_n h_1(m-2a)h_0(n-2b)LH^{j-1}(m,n) \tag{3.8}$$

$$HH^j(a,b) = \sum_m \sum_n h_1(m-2a)h_1(n-2b)HH^{j-1}(m,n) \tag{3.9}$$

The result of this decomposition is a tree shown in Figure 3.5 [27], which contains all the DWT coefficients for N^2 translates of the image X, where the size of the image is $N \times N$. In this work, since the images are represented in the lab color space, three trees are

generated, each corresponding to a color channel. The total entropy of the color image is found by calculating the entropy of the corresponding color subbands and then adding them up. Since the middle wavelet detailed channels contain most textural information [63] and in order to limit the number of coefficients being generated, only two levels of decomposition was performed on the images in this work.

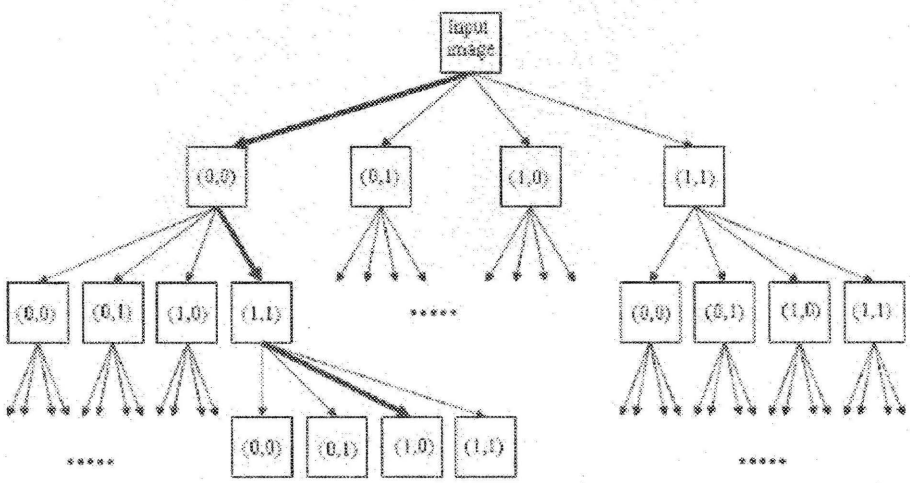

Figure 3.5: SIDWT decomposition tree for three levels of decomposition with the best selection corresponding to the minimum cost path

3.2.3 Cross Co-occurrence matrices

The principle of cross co-occurrence matrices is based on the gray scale co-occurrence matrices(GCM). The GCM for a gray scale image shows the distribution of co-occurring values at a given offset. Each entry in a GCM, $M(i,j)$, indicates how often a pixel with gray-level value i occurs at the distance d to a pixel with the value j, where d is the given offset vector. A cross co-occurrence matrix (CCM) is the counterpart of GCM for color images. Let I be an $N \times N$ small bowel image and $b1$, $b2$, and $b3$ the three color subbands. $C_d^{b1,b2}$ is the CCM matrix for the color subbands $b1$ and $b2$ for the offset d. Hence, each entry of the matrix,

$C_d^{b1,b2}(i,j)$, represents the probability of the intensity level i in the color subband $b1$ and intensity level j in the color subband $b2$ to occur at two locations seperated by the distance vector d. As shown in [27], since the subbands are oriented, only some particular CCMs are calculated on each subband. The displacement vectors are grouped according to the orientation of the subbands: vertical, diagonal and horizontal. Six matrices are generated for each subband or 18 matrices in total for an image. Finally, since $C_d^{b1,b2}$ and $C_d^{b2,b1}$ represent the same information, the average of these two matrices $M_d^{b1,b2} = \frac{C_d^{b1,b2}+C_d^{b2,b1}}{2}$ is used in this work. The use of CCMs has the advantage of extracting color and texture information at the same time. As proposed in [27][64], four principal features can be derived from each matrix: contrast, energy, homogeneity and entropy. In this work however, based on the efficiency of the features only two features are kept: energy and homogeneity. The former is calculated as the sum of the squared elements. If M is a cross co-occurrence matrix, the energy for the matrix is calculated as

$$N = \sum_{i,j} M(i,j)^2 \tag{3.10}$$

Homogeneity is another feature used to describe the textural characteristics in the image. This feature measures the closeness of the distribution of elements in the co-occurrence matrix to the matrix diagonal and is defined by

$$H = \sum_{i,j} \frac{M(i,j)}{1+|i-j|} \tag{3.11}$$

Two sets of features are extracted from each image based on the energy and homogeneity measures. As mentioned earlier the CCMs are calculated for three groups of offsets, vertical, horizontal and diagonal. Hence, there are 6 matrices for each subband or 18 matrices per image. Finally, two sets of features are extracted from each CCM based on energy and homogeneity measures, which makes for a total number of 36 features for each image.

3.3 Classification and Results

To evaluate the performance of an unsupervised classification scheme on the small bowel dataset, two sets of experiments were conducted using k-means and fuzzy C-means clustering

algorithms. The algorithms were applied to the extracted features. The database contains 75 images, including 41 healthy (normal) images and 34 diseased (abnormal) images. using the feature extraction techniques in the previous sections, each image in the database is represented with a feature vector of 36 features. Since there is a considerable difference in the range of the values for different features, the features are normalized prior to further analysis. In both classification scenarios (using k-means and fuzzy C-means) the number of clusters is needed to be known beforehand. Since in this work we aim to detect the existence of abnormalities in the images, and not determine the type of abnormalities, the number of clusters is defined to be 2 to represent normal and abnormal images. The Fuzzy C-means algorithm calculates, for each image X, the degree of membership for the healthy cluster and the diseased cluster. Then the images are separated into two clusters based on the criterion of maximum membership. For the k-means algorithm, it is the same method; the same matrix of extracted features F is used. The algorithm calculates the squared Euclidian distance between each row of F (which represents one small bowel image) and the centroid. The centroids are then recalculated and these steps are repeated until the algorithm converges. The result of the two algorithms is a 75×1 matrix. Each row of the matrix corresponds to one image in the dataset and indicates whether the image belongs to group one or group two. Finally it is the physician who labels one group as the healthy bowels and the other as the diseased bowels.

The efficiency of the algorithm is provided in the confusion matrix (or the matching matrix) given in Table 3.2. Table 3.1 shows the definition of the confusion matrix where the specificity and sensitivity are defined as:

$$Sensitivity = \frac{Number\ of\ correct\ positive\ predictions}{Total\ number\ of\ abnormal\ cases} = \frac{TP}{TP + FN} \tag{3.12}$$

$$Specificity = \frac{Number\ of\ correct\ negative\ predictions}{Total\ number\ of\ normal\ cases} = \frac{TN}{TN + FP} \tag{3.13}$$

$$efficiency = \frac{Number\ of\ correctly\ classified\ images}{Total\ number\ of\ images} \tag{3.14}$$

As it can be seen in Table 3.2 that an accuracy rate of 76% is achieved which is a rather satisfactory result for an unsupervised classifier.The results of other classification methods

(k-means and SOTM) using energy and homogeneity features are provided in Tables 3.3 and 3.4 for comparison.

Another measure to evaluate the performance of the classification method is the Receiver Operating Characteristic (ROC) curve. The ROC curve represents the fraction of true positive (TP) vs the false positive (FP). The TP corresponds to the sensitivity and is the proportion of diseased bowel images classified as abnormal while the FP represents the portion of normal images classified as abnormal. An ideal classifier would yield a point in the upper left corner or coordinate (0,1) of the ROC space, where all images have been correctly classified. This point represents 100% sensitivity (no false negatives) and 100% specificity (no false positives). The classification accuracy is also measured by calculating the area under the ROC curve. An area of 1 corresponds to perfect classification, whereas an inefficient classification is represented by a horizontal straight line going from the point $(0, 0)$ to the point $(1, 1)$. In order to have an efficient classifier, the curve has to be above this line. The ROC curve for the unsupervised classification techniques used in this work is given in Figure 3.6, where the area under the ROC curve was calculated to be 0.76. Table 3.5 shows the results of using different feature sets along with supervised and unsupervised classification methods. In the supervised classification, LDA has been used in conjunction with leave one out method (LOOM) to combat the problem of small sample size. In the unsupervised column, the results of applying fuzzy C-means is provided. Both techniques are used on the same database of 75 images (including 41 normal and 34 abnormal images). As it can be seen from the table, the extracted feature for a supervised classifier are not necessarily optimal for an unsupervised classifier. However, a feature set that yields a good results with an unsupervised classifier may naturally lead to better results if a supervised classifier is used. This shows how an unsupervised classification can be used as a first step in classification to select the naturally most discriminant features. From Table 3.5 it can be observed that using the SIDWT along with cross co-occurrence matrices in the RGB color space returns an accuracy rate of 52% for the k-means or fuzzy C-means clustering while a relatively high accuracy rate is achieved using a supervised classifier. Nevertheless, using

the feature set that is extracted in the Lab color space for unsupervised classification results in an accuracy of 76%. In an attempt to test the methods with more images, all the images were rotated by 180 degrees to obtain a database of 150 images. The classification accuracy for the enlarged database is 70.7% which shows the method could be applicable to larger databases.

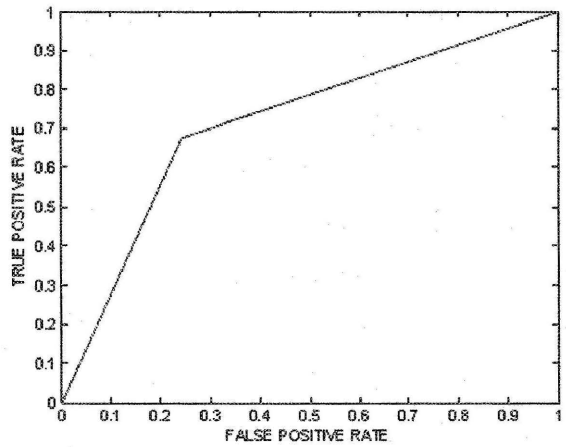

Figure 3.6: The Receiver Operating Characteristics curve with an area of 0.76

3.3.1 Future work

Although wavelets are shown to be effective as texture feature extraction tools, the adaptation of other texture descriptors for the medical images is growing. Among the new textural features, textons have shown promising results in extracting texture features for classification. Textons are used to describe the fundamental micro-structure elements in natural images. The appearance of the textons has a root in the psychological study of the texture recognition process in human. The theory of textons was first proposed by Jluesz [65] to explain the "preattentive discrimination" of the texture pairs. To discuss Julesz pioneering

	Predictive positive	Predicted negative
Actually positive	TP(true positive)	FN(false negative)
Actually negative	FP(false positive)	TN(true negative)

Table 3.1: The definition of confusion matrix

	Normal	Abnormal
Normal	32	10
Abnormal	8	25

Table 3.2: Classification results for the fuzzy C-means classifier

work on textons, we need to describe these two concepts:[66]

First order statistics refers to the probability of occurrence of a gray value at a random location in an image. These statistical measures can be calculated from the histogram of gray level intensity of the image. First order statistics depend only on individual pixel values and not on the co-occurrence of the neighbor pixels. The mean gray level value in an image is an example of first order characteristics.

Second order statistics measure the likelihood of gray level intensities occurring separated with a displacement vector d where the length and orientation of the vector d is random.

	Normal	Abnormal
Normal	29	10
Abnormal	11	23

Table 3.3: Classification results for the k-means classifier

	Normal	Abnormal
Normal	25	16
Abnormal	8	26

Table 3.4: Classification results for the SOTM classifier

These two attributes were used by Julesz to determine whether two textures are preattentively discriminable. The theory of textons was proposed to address this problem. Textons can be considered as visual events in an image such as collinearity, termination and closure. Using the theory of textons, the two different textures in Figure 3.7 can be described as follows. The two regions in Figure 3.7(a) have identical second order statistics and the number of terminations (i.e texton information) in both the upper and lower regions is the same, therefore the human visual system is not able to discriminate the two textures preattentively. On the other hand, in Figure 3.7(b), the number of terminations in the upper and lower region is different (three in the upper half and four in the lower half). Because of the difference in this texton, the two textures are discriminable.

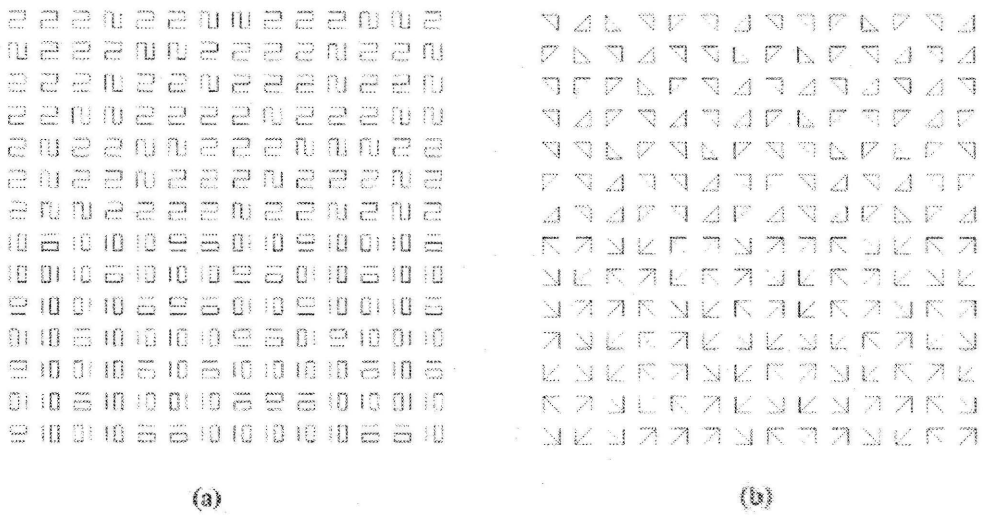

Figure 3.7: Texture pairs with identical second-order statistics. (a) The upper half and lower half contain the same textons. The visual system can not discriminate the different textures without careful scrutiny. (b) The upper region contains textons different from the lower region. Humans can differentiate the two textures effortlessly.

The application of textons in the area of medical image processing is growing recently. in [67] Harms et al. extract texture micro-edges and textons between these micro edges to diagnose leukemic malignancy in samples of stained blood cells.

In [68] a texture feature extraction based on textons is used to classify the breast density pattern to determine the breast cancer risk.

In [69] Tuzel et al. use texton histograms to distinguish among hematology cases directly from microscopic specimens. The images contain normal images and for groups of four different hematologic malignancies. Initially, the basic texture elements (textons) for the nuclei and cytoplasm are learned, the cells are represented through texton histograms and finally a SVM classifier is applied to the extracted features. The work proposed by Adjeroh et al. in [70] is one example of using the textons for segmentation of retinal images.

The application of textons in the area of medical image analysis for extracting texture information appears to be increasing among the researchers and the results are promising. Hence, as the future work a new set of features based on textons can be developed for the small bowel images to extract the texture information and improve the accuracy.

Color space	Extracted features	Unsupervised classification	Supervised classification
RGB	Contrast Energy Homogeneity Entropy	52%	94.7%
Lab	Contrast Energy Homogeneity Entropy	53%	78%
Lab+RGB	Contrast Energy Homogeneity Entropy	56%	79%
Lab	Energy Homogeneity (normalized features)	76%	76%
Lab	Energy (third subband, normalized features)	72%	84%
RGB	Energy Homogeneity	61%	78%
Lab+RGB	Energy Homogeneity	65%	88%

Table 3.5: Comparison of the results of unsupervised classification method with supervised classification for different feature sets and different color spaces.

Chapter 4

Unsupervised Learning in Hearing Aids Signal Analysis

4.1 Audio classification for hearing aids

SPEECH and environmental audio signals are important sources of information in our everyday communication, and can provide information about the location or environment of the captured scene or event. Having approximately 10% of the world population suffering from some sort of hearing loss, one of the important applications of audio classification is in hearing aids for hearing impaired people. Users of hearing aids are forced to listen under a variety of noise conditions and in most cases simple amplification cannot help hearing-impaired listeners. Such devices amplify the noise as well as the desired signal. Consequently, numerous signal enhancement algorithms have been proposed for digital hearing aids. To overcome this problem, the hearing aid should be able to detect the audio classes which the incoming signals belong to, and then change the hearing aid parameters accordingly. The first step to achieve this goal is the ability to quickly and correctly classify the audio signals in the environment.

There is a growing body of evidence that different hearing aid characteristics that can operate efficiently under different listening conditions are desirable [71]. In a survey obtained by Kochkin [32] from 2323 hearing aid users it was observed that less than one third of the hearing aid users were satisfied with their hearing aid if the device worked properly in

only three or fewer environments while over 91% of the users were satisfied if the hearing aid worked wherever it was needed. Thus if the hearing aid can be automatically adjusted according to the listening conditions substantially better user satisfaction would be expected.

4.2 Audio signal classification

Audio signal classification is one the tasks that humans perform effortlessly all the time. Differentiating the voice of a singer from the music, understanding heavily accented speech, recognizing a voice on the telephone, telling the difference between a helicopter sound and a car sound, discriminating the speakers voice from the background noise are some of the auditory tasks that we do every day without even considering them. However, duplicating this capability on machines takes an intensive effort. In the area of machine learning and artificial intelligence, analysis and discrimination of the audio signals is one the research areas that has been active for a long time and is not completely solved yet. There is a wide range of applications for the classification of audio signals. Speech processing for security applications and human computer interaction, multimedia data management and distribution, security, biometrics and bioacoustics are some of the applications of audio signals classification [33].

4.2.1 Taxonomy of audio signals

Before discussing different existing classification and analyzing techniques, it is important to define a taxonomy of auditory signals. Audio signals can be sorted into classes from different viewpoints. However, the taxonomy presented here is based on the origin of the signal. Figure 4.1 shows the taxonomy of the audio classes used in this work as a reference. The audio signals used in this work can be divided into two main groups, signals that have a natural origin and those which are human made or artificial sounds. Natural signals are then subdivided into human signals (or speech), which in turn consists of male and female speakers and non-human sounds, which include bird, animal and insects. On the other hand, human made sounds consist of two main categories: machine sounds, which in turn are divided into helicopter and aircraft, and musical instruments such as piano, flute and

drum. Other taxonomies with higher resolution can be obtained in many ways for example by subdividing the human speech into pathological or normal or by dividing musical sounds into different musical genres such as pop, rock, etc. In this work however, we confine our attention to the taxonomy given in Figure 4.1.

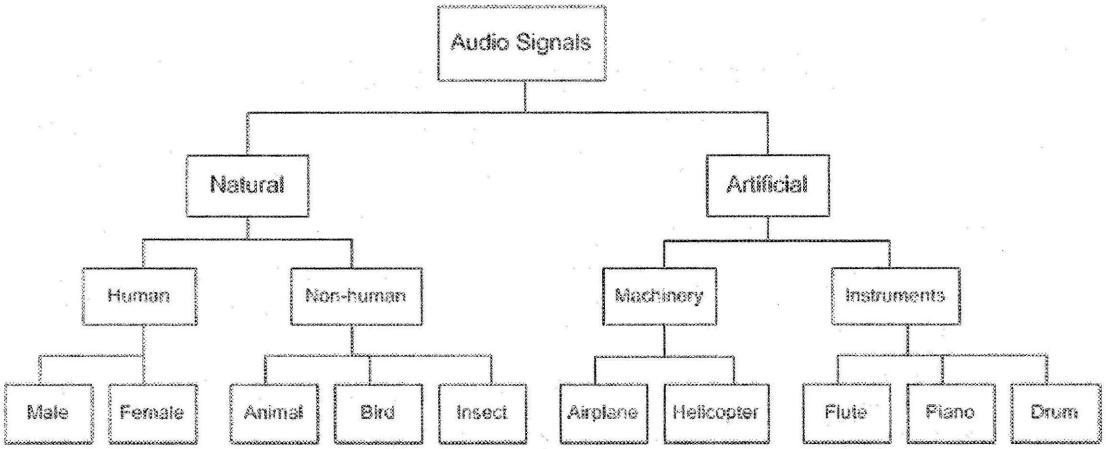

Figure 4.1: Taxonomy of audio signals used in this work

4.2.2 Audio signal classification

Audio signal classification consists of extracting physical and perceptual features from an audio signal (or one segment of the signal) and categorizing the signal into one the given audio classes. Audio classification in general is a wide area of research and a large amount of research has been done on it in the last decade. Most of the research works in the area can be divided into three main categories: speech, music and audio scene analysis. Each of these topics will be discussed in more detail below:

Speech analysis

A considerable part of research in the area of audio signals has been devoted to speech analysis and classification. Speech analysis is a wide area of research itself. The following areas are some of the major branches of the speech analysis in the literature.

Speech recognition: Speech recognition is one of the oldest and the most fundamental speech classification problems. The goal is to convert the words from the human speech into a readable text. Speech recognition has a wide range of applications in the areas such as health care, military, security, telephony and enabling people with disabilities. References [72] [73] and [74] are some of the comprehensive works in the area of speech processing history and proposed methods and solutions.

Pathological speech analysis: Can be used for recognition of selected types of vocal tract pathologies [75]. Various pathological conditions affect the vocal functions, which result in speech disorders. The aim of pathological speech analysis is to assess the speech disorders by using acoustic characteristics of the speech. It can be also helpful in monitoring the progress of the patient over the course of therapy [33]. Further more, it is valuable to provide the physician with a quantitative guideline for a deformation degree assessment of speech signal [76].

Speaker recognition: Speaker recognition (or sometimes called speaker verification [77]) is the identification or verification a user based on the characteristics of their voice. Compared to the speech recognition problem, where the main goal is to determine what word is uttered, the goal is to find out who the speaker is. Some of the applications of speaker recognition can be speaker authentication, identification or biometrics.

Music

As the amount of multimedia and music files is growing every day, automatic extraction of music information is gaining more importance as a way to structure and organize the increasingly large numbers of music files available digitally on the Web. Today a large portion of the audio classification literature is related to music and music information retrieval. However, most of the research in this area, fall within one of these categories:

Music content analysis: With the creation of huge music databases, the demand for fast and reliable tools for content analysis and description is growing. These analysis tools

can be used for searches, content queries, and interactive access. Amongst all possible descriptors, music genres are crucial since they have been widely used for years to organize music catalogues, libraries, and music stores [41]. A musical genre is typically characterized by the common attributes related to instrumentation, rhythmic structure, and harmonic content of the music. The music genre classification maps a taxonomy of genres, i.e., a hierarchical set of categories onto a music collection. Similar to any other classification problem, a set of features is used to decide on the music genre. Table 4.1 shows a summary of the features being used in music content retrieval today [41]. As for the classification, a number of supervised and unsupervised methods have been proposed. Shao et al. [39] use agglomerative hierarchical clustering on their music dataset. In the work by Rauber et al. [40] the growing hierarchical self-organizing map is applied to cluster data and organize them on a two-dimensional space. References [78] and [79] are examples of application of supervised classifiers where K-nearest neighbor are used in the context of genre classification. The hidden markov models (HMMs) have been used in [80] and [81]. In [82] West and Cox show the applications of linear discriminant analysis in genre classification of audio content. In [83] support vector machines are used for the classification purpose and finally [84] is an instance of the use of artificial neural networks.

Musical instrument recognition: Musical instrument recognition is another aspect of music information retrieval. Such a capability may be extremely helpful in the framework of automatic musical transcription systems as well as in content-based search applications. One of the practical applications of musical instrument recognition is automatic music transcription. A typical task of classification of musical instruments consists of three phases [85] the first step is the preprocessing, which can be also referred to as pitch extraction. The next stage is the extraction of frequency information, fundamental frequencies and harmonics. These information will then be used in the third stage which is the pattern recognition and classification stage. Some of the works use the temporal information as well [86]. References [87] [88] and [89] are some of the

other existing techniques in the literature on the recognition of musical instruments.

Speech/music discrimination: Another aspect of content based audio classification that has attracted many researchers is discrimination of human speech from the music. In this process, sometimes we are more interested in extracting the speech information form the background music, for example for the purpose of performing automatic speech recognition on the soundtrack data. On the other hand, sometimes the music content is of more importance e.g. many listeners are more interested in the music on broadcast radio rather than the commercial and talk programming. The works by Hawley et al. [90] and Saunders et al. [91] are some of the previous works on this topic in the literature. Several feature sets have also been suggested for this purpose. In [92] a comparison of the proposed feature sets for speech/music discrimination (such as cepstral coefficients, amplitude features and pitch features) is presented.

Timbre	Melody/Harmony	Rhythm
texture model: model of features over texture window: 1) Simple modeling with low order statistics 2) modeling with auto regressive model 3) modeling with distribution estimation algorithms(e.g. EM estimation of a GMM of frame)	pitch function: measure of the energy in function of music notes 1) Unfolded function: describes pitch content and pitch range 2) folded function: describes harmonic content	periodicity function: measure of the periodicities of features 1) Tempo: periodicities typically in the range 0.31,5S (i.e., 20040 BPM) 2) musical pattern: periodicites between 2 and 6 s (corresponding to the length of one or more measure bar)

Table 4.1: Typical features used for music content retrieval

Audio scene classification

Audio scene analysis is the process of extracting information about the environment based on the characteristics of the received signal, and has numerous applications in multimedia processing. Hence, compared to the previously mentioned classification categories (music and speech) audio scene analysis is a more general and comprehensive task. The idea of audio scene analysis was first proposed by Bregman in [93], which is the cornerstone of this area. In his work Bregman presented a new perspective in human sound perception. The concept of audio scene analysis comes from the way that human brain works to use the sounds to build a picture from the surrounding environment, which is also called an auditory scene. There are numerous applications for audio scene analysis. Amongst all, one of the most popular applications of audio scene analysis is in the development of smart hearing aids, which will be discussed in more details in the future sections.

4.2.3 Review of the previous works

Many methods have been proposed in the area of audio signal classification with the application to hearing aids.

In [71] Kates proposes the selection of processing algorithm based on the audio information from the scene. Nordqvist and Leijon [34] introduced a hidden Markov model (HMM) based classifier for hearing aids using features derived from cepstral coefficients. In the work done by Buchler et. al [35] a variety of machine learning techniques (k-means, histogram driven Bayes classifiers, multilayer perceptrons, and HMMs) were tested and the ergodic HMMs were shown to outperform the rest of the methods. Audio content analysis at Microsoft research commonly employs Gaussian mixture models (GMM)[36], k nearest neighborhood (K-NN)[37] and support vector machine (SVM)[38] for audio classification. Other popular classifiers for audio classification include linear discriminant analysis (LDA) [33], hidden Markov models (HMM)[39] and artificial neural networks (ANN)[94].

While there is a large amount of research in the literature on the application of supervised classifiers, the use of unsupervised classifiers for audio classification is relatively unexplored.

Clustering (or unsupervised) approaches are most beneficial in cases where precise manual labeling of the data is time consuming and laborious or when the feature characteristics might change over time. As mentioned earlier, the hearing aid is expected to operate in a wide range of audio environments. Therefore, the number of audio classes and nature of the classes in the received audio signal is not predictable. In this case, a clustering approach can be beneficial to discover different audio classes that exist in the received audio signal. This step can be followed by supervision to select and amplify the desired audio class. In addition, using a clustering method has the advantage of avoiding the constraints of a fixed taxonomy, which may suffer from ambiguities and inconsistencies. In addition, considering the variety of the audio signals, some of the signals may simply not fit within a given category [41]. The use of a clustering technique makes it possible to take into account the overlap that might exist between different classes. In [39], Shao et al. use an agglomerative hierarchical clustering on the audio data set for music genre classification. Rauber et al.[40] use the growing hierarchical self organizing map to create a 2-D output for visual representation of the music data set. The classification method proposed in this work is based on the self organizing tree maps, which was explained in Chapter 2, followed by a fuzzy labeling of the data. approach allows for extraction of underlying characteristics of the data and then supervised labeling is used to interpret the discovered clusters.

The proposed methods can also be discussed from the point of feature extraction. Most of the existing method extract either temporal or spectral features for classification. A wide range of feature sets have been proposed for this purpose. In [92] a comparison of different feature sets proposed for audio classification is given. Some of the suggested features include signal energy, pitch, zero crossing rate [92] [91], Entropy modulation [95], 4 Hz modulation energy, percentage of low-energy frames, spectral rolloff point, spectral centroid, mean frequency, cepstral coefficients [96], [97] and high and low frequency slopes [71]. All the mentioned features are extracted only from time or frequency domain; however, the temporal or spectral features are not enough for representation and localization of non-stationary aspects of audio signals, such as trends, discontinuities, and repeated patterns.

Thus the features used in this work are based on joint time and frequency analysis of the signals, which is effective for revealing non-stationary characteristics of audio signals.

Work	Classification technique	Features
Nordquist et al. [34]	HMM	Delta features from cepstral coefficients
Bchler et al. [35]	k-means, MLP bayes classifier, HMM	Tonality, width, pitch variance, measures of time offset
Abu-El-Quran et al. [36]	Adaptive thresholding of feature values	4Hz modulation, low energy frames, spectral rolloff, spectral centroid, cepstral residual, pulse metric, spectral flux, zero crossing rate, variance of the low band energy
Lu et al. [37]	K-NN	High zero crossing ratio, low short time energy ratio, spectrum flux, LSP divergence,band periodicity, noise frame ratio
Guo et al. [38]	SVM	Total power, subband powers, brightness bandwidth, pitch, mel frequency cepstral coefficients (MFCC)
Shao et al. [39]	HMM	MFCC, linear prediction coefficients derived from cepstrum coefficients, delta and acceleration
Freeman et al. [94]	ANN	Mean frequency, high and low frequency slopes, envelope modulation

Table 4.2: Summary of the feature extraction and classification techniques used in the literature for audio classification

4.2.4 The proposed method

Figure 4.2 shows the block diagram of the implemented system, where the blue lines show the flow of the train data and the red lines show the flow of the test data. In the training phase each input audio segment X is passed through the adaptive time-frequency decomposition (TFD) block. The TFD matrix V is then decomposed by the use of Non-negative matrix Factorization (NMF) methods into base and coefficient matrices W and H. Then the features are processed and the desired number of features are extracted from each base vector and its corresponding coefficient vector to form the feature set f. Once this procedure is run for all the segments in the training set, the SOTM clustering technique is applied to the data to discover the clusters and computer the cluster centers C. Then a membership degree is calculated for each cluster, α, which will be used for the labeling of the test data. Each segment in the test dataset, after passing through the feature extraction block, is fed to the data labeling block, where the decision is made about which class the segment belongs to. All of these blocks will be described in more details in the future Sections.

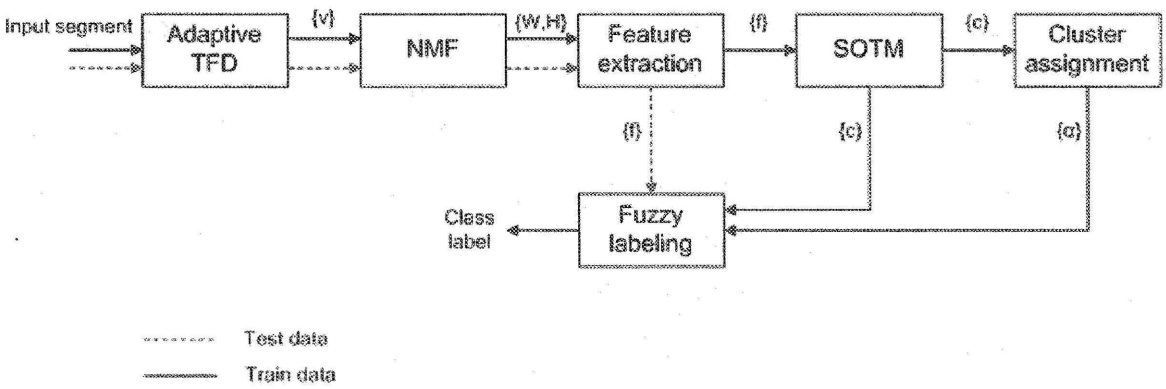

Figure 4.2: Block diagram of the feature extraction and classification

4.3 Feature extraction

The features used in this work are captured by applying the matching pursuit algorithm on the signals followed by the non-negative matrix decomposition. The concept of these two algorithms are briefly described in Sections 4.3.1 and 4.3.2. Then a feature set is created from the results of these two algorithms which is described in Section 4.3.3.

4.3.1 Matching pursuit TFD

In every day conversations, we communicate a wide range of ideas with precision. By adding or omitting a few words, we can communicate subtle differences in close meanings. This is possible due to the fact that natural human languages have large vocabularies that include words with close meanings. In the area of information processing, a low level representation of the signal must include information about distinct properties and minor differences simultaneously. However, most of the signals we deal with in real life applications (such as audio signals) are complex signals that consist of a wide scope of patterns. Precise representation of these signals with few basic functions is not an easy task [98]. This is the motivation behind the idea of projection of the signals onto large and redundant dictionaries of waveforms, which was proposed by Mallat et al. in [98]. According to this work, linear transforms (such as Fourier and wavelet) do not have the flexibility required for representing wide range of signals. Fourier transform eliminates temporal properties and hence provides a poor representation of the signals that are well localized in time. Wavelet bases also are not optimal for those signals whose Fourier transform has a narrow high frequency support. Hence, decomposing a signal on such basis, is like writing a text using a small vocabulary. Although it might be possible to express the idea, it takes extra effort and extra words to describe the unavailable words. Flexible decompositions are particularly important for those signals whose local temporal and spectral properties vary widely.

In the matching pursuit algorithm, the signal is decomposed into a linear expansion of waveforms. These waveforms belong to a redundant dictionary and are selected in order to best match the signal structure. These waveforms are called time-frequency atoms. For

example, impulses need to be decomposed using atoms that are well concentrated in time, while spectral lines are better represented by waveforms which have a narrow frequency bandwidth. Although the matching pursuit decomposition is a nonlinear algorithm, it maintains the energy conservation property like an orthogonal expansion.

When using a dictionary of time-frequency atoms, applying the matching pursuit algorithm yields an adaptive time-frequency transform. It decomposes the function $f(t)$ into a sum of complex time-frequency atoms that best match its residues. A general family of time-frequency atoms can be obtained by scaling, translating and modulating a single window function $g(t)$. By denoting $\gamma = (s, u, \xi)$, the function $g(t)$ can be defined as

$$g_\gamma(t) = 1/\sqrt{s} g(\frac{t-u}{s}) e^{i\xi t}, \tag{4.1}$$

Where $s > 0$ is the scale and ξ and u represent frequency modulation and translation respectively. The Fourier transform of $g_\gamma(t)$ can be written as [99]

$$\hat{g}(\omega) = \sqrt{s}\hat{g}(s(\omega - \xi)) e^{-i(\omega-\xi)u} \tag{4.2}$$

In this work, a dictionary of Gabor time-frequency atoms has been used. The discrete Gabor time-frequency atom can be written as

$$g_\gamma(n) = cos(i\frac{2\pi k}{N} + \phi), \tag{4.3}$$

where

$$g_s(n) = \frac{K_s}{\sqrt{s}} \sum_{p=0}^{N} 2^{1/4} e^{-\pi} (\frac{n - pN}{s})^2. \tag{4.4}$$

The constant K_s is used for normalizing the function g_s, p $(0 \leq p < N)$ is the time shift, ϕ $(0 \leq \phi < 2\pi)$ the phase shift, and $0 \leq k < N$. The decomposition of the signal f can be written as a linear expansion of the signal over a set of atoms selected from the dictionary. In order to find the atoms that best match the structure of the signal a successive approximation

of f with orthogonal projections on the elements of the dictionary is performed [99]. After n iterations, the decomposition of the signal f is given by:

$$f = \Sigma_{i=0}^{n-1} \langle R^i f, g_{\gamma i} \rangle g_{\gamma i} + R^i f, \tag{4.5}$$

where $R^i f$ is the decomposition residue after n iterations and $\langle , \rangle$ denotes the inner product of the two functions. At each stage of iteration, the algorithm selects the atom $g_{i\gamma}$ for which the inner product $\langle R^i f, g_{\gamma i} \rangle$ is maximized [99]. The energy distribution of the decomposition can be written as

$$Ef(t,\omega) = \sum_{i=0}^{n-1} |\langle R^i f, g_{\gamma i}(t,\omega) \rangle|^2 W g_{\gamma i}(t,\omega) \tag{4.6}$$

where $W g_{\gamma i}(t,\omega)$ is the Wigner distribution of the atom $g_{\gamma i}(t,\omega)$ which does not include cross terms [99].

4.3.2 Non-negative matrix factorization

Non-negative matrix factorization (NMF) is a decomposition technique proposed by Lee and Seung in [100]. The interpretation of NMF for the application of statistical analysis of the multivariate data can be described as follows: Assume V is an $m \times n$ non-negative data matrix, where n is the dimension of the data and m is the number of vectors or the number of samples in the data set. The goal is to find non-negative matrix factors $W_{m\times r}$ and $H_{r\times n}$ to approximate the matrix V, such that

$$V \approx WH, \tag{4.7}$$

and also

$$v \approx Wh, \tag{4.8}$$

where v and h are the corresponding columns of V and H respectively. This means each data vector v can be approximated by a linear combination of the columns of W weighted

by the components of h. Therefore W can be considered as a set of basis vectors that are optimized for the linear approximation of the data in V.

Usually r is selected to be smaller than n or m, so the result is a compressed version of the original matrix, V and the data vectors can be represented using fewer basis vectors. On the other hand, in order to obtain a good approximation, the basis vectors should discover the structure that is latent in the data. In this work, the NMF technique has been performed on the time-frequency matrix. Therefore n is the length of the signal, m is the frequency resolution of the constructed time-frequency matrix, and r is the decomposition order. After decomposition, W and H carry spectral and temporal characteristics of the original matrix respectively. W contains spectral structures and H contains the corresponding location of each spectral structure in the original matrix. The problem of finding W and H can be considered as a minimization of the function

$$f = \|V - WH\|^2 \tag{4.9}$$

There is a variety of strategies in the literature to find W and H [101][102]. In this work a gradient-based method proposed by Lin in [103], which uses bond-constrained optimization technique. The standard form of bound-constrained optimization problem can be expressed as [103]:

$$\begin{aligned} &\min f(x) \qquad x \epsilon R \\ &\text{subject to} \quad l_i \leq x_i \leq u_i, \quad i = 1, ..., n \end{aligned} \tag{4.10}$$

(4.11)

$$x^{k+1} = P[x^k - \alpha^k \nabla f(x^k)] \tag{4.12}$$

where

$$P[x] = \begin{cases} x_i & for \quad l_i < x_i < u_i \\ u_i & for \quad x_i \geq u_i \\ l_i & for \quad x_i \leq l_i \end{cases} \tag{4.13}$$

In [103] this technique is applied to the NMF problem. This method is computationally efficient and offers better convergence properties than the standard approach [103].

4.3.3 Feature selection

As shown in Figure 4.2, once the TFD matrix (V) is decomposed into base and coefficient matrices (W and H), a feature set is extracted from each base vector and its corresponding coefficient vector. The features are derived from coefficient vectors, base vectors, and from MP decomposition. A brief description of the features used in this work is provided here:

1. **Sparsity:** The sparsity feature is calculated for each coefficient vector,$\{h_i\}_{1\times N}$, as

$$S_{h_i} = \frac{\sqrt{N} - (\sum_{n=1}^{N} h_i(n))/\sqrt{\sum_{n=1}^{N} h_i^2}}{\sqrt{N} - 1} \tag{4.14}$$

   The value of this feature is 1, if and only if h_i contains a single non-zero component, and is zero if and only if the components are equal.

2. **Sum of derivatives:** This feature is calculated on the base vector and represents discontinuities and abrupt changes in the signal. The equation for derivation of this feature is given by

$$D_{h_i} = \sum_{n=1}^{N-1} h_i'(n)^2, \tag{4.15}$$

   where

$$h_i'(n) = h_i(n+1) - h_i(n) \tag{4.16}$$

$$n = 1, ..., N-1 \tag{4.17}$$

   The value of this feature is a measure of discontinuities. If there are discontinuities in the coefficient vector, the value is large, otherwise it is small.

3. **Moments:** The first moment of the base and coefficient vectors are also extracted. The spectral and temporal moments, $MO\omega_i$ and MOh_i, are obtained using the following equations

$$MO\omega_i = \sum_{m=1}^{M} m\omega_i(m) \tag{4.18}$$

$$MOh_i = \sum_{n=1}^{N} nh_i(n) \tag{4.19}$$

where h_i and ω_i are the base and coefficient vectors and M is the frequency resolution of the TFD.

4. **Sparsity I:** In addition to the sparsity of the coefficient vectors, the sparsity of the base vectors is also extracted. This feature represents the noisy structure of the signal and is calculated as

$$S_{\omega_i} = \frac{\sqrt{M} - (\sum_{m=1}^{M} \omega_i(n))/\sqrt{\sum_{m=1}^{M} \omega_i^2}}{\sqrt{M} - 1} \tag{4.20}$$

5. **Sparsity II:** This feature is defined as the number of samples whose value is smaller than a threshold ϵ to the total number of samples in the base vector:

$$SP_{\omega i} = \frac{\omega_i < \epsilon}{M}, \tag{4.21}$$

where $\omega_i < \epsilon$ is the number of base samples less than a small threshold and M is the total number of samples in each coefficient vector. This function is unity if and only if all the components in ω_i are greater than the threshold, and is zero if and only if all the samples are less than the threshold.

6. **Periodicity:** While the previous feature measures the scattering of the components in frequency, we still need another feature to represent the presence of harmonicity of the energy in frequency. For each base vector, the Fourier transform of the vector is calculated as

$$W_i(\nu) = |\sum_{m=1}^{M} e^{-j\frac{2\pi m\nu}{M}} \omega_i(m)| \tag{4.22}$$

where M is the length of the base vector, and $W_i(\nu)$ is the Fourier transform of the base vector ω_i. Next a second Fourier transform is performed on the base vector to obtain $W_i(\kappa)$ as

$$W_i(\kappa) = |\sum_{\nu=1}^{M/2} e^{-j\frac{2\pi\nu\kappa}{M/2}} W_i(\nu)| \tag{4.23}$$

Finally we sum up all the values of $|W(\kappa)|$ for $\kappa > m_0$, where m_0 is a small number.

$$P_{\omega_i} = \Sigma_{\kappa=m_0}^{M/4} |W_i(\kappa)| \tag{4.24}$$

The value of P_{ω_i} is large for bases whose components show strong periodic behavior, such as vowels in speech. However, for non-periodic sounds such as aircraft, the feature has lower values.

7. **Sum of derivatives:** This feature is calculated on the coefficient vectors and captures discontinuities and abrupt changes in the signal.

$$D_{\omega_i} = \sum_{m=1}^{M-1} \omega_i'(m)^2 \tag{4.25}$$

where

$$\omega_i'(m) = \omega_i(m+1) - \omega_i(m) m = 1, ..., M-1 \tag{4.26}$$

$$m = 1, ..., M-1 \tag{4.27}$$

where ω_i' is the first derivative of the coefficient vector. The value of this feature is large if the coefficient vector contains discontinuities.

8. **Projection features:** As shown in Eq 4.5, MP decomposition projects the signal onto a set of time-frequency atoms. The amount of signal energy that is projected in each iteration depends on the structure of the signal. Signals with coherent structures need less number of iterations, while signals with a non-coherent structure tend to take more iterations to get decomposed. This property is used as feature to discriminate coherent audio signals form non-coherent signals. To extract this class of features, first we calculate the difference in the projection energy between iteration i and $i+1$:

$$d_i = \tilde{a}_{i+1} - \tilde{a}_i \tag{4.28}$$

$$i = 0, ..., I-2 \tag{4.29}$$

where

$$\tilde{a} = \frac{a_{\gamma i}}{\text{Total energy of the decomposed signal}} \tag{4.30}$$

is the ratio of the projection energy at each iteration. Next, we define L_i as the sum of the energy differences:

$$L_i = d_0 + d_1 + ... + d_i \tag{4.31}$$

$$i = 0, ..., I-2 \tag{4.32}$$

L_i keeps the trend of the energy coefficients (a_i) but it is normalized and it is independent of the signal's energy. Finally, normalized coefficients (L_i) are used to calculate MP feature:

$$MP = \sum_{i=0}^{I-2} L_i \tag{4.33}$$

4.4 Classification and results

4.4.1 classification methodology

The classification method used in this work is based on the SOTM clustering algorithm. The proposed method, which is a fusion of supervised and unsupervised classification, consists of two stages. In the first stage the SOTM clustering algorithm is applied to the training dataset. Since the data is represented to the SOTM in a random manner, the formation of the clusters might be slightly different for each run. In fact, some of the discovered clusters include one or very limited number of samples. Therefore, those clusters in which the number of samples is smaller than a threshold will be eliminated. The value of this threshold in this work was adjusted to be 5% of the total number of samples in the train data set. Next a membership matrix, $M_{m\times n}$, is calculated based on the distribution of each class in different clusters, where m is the number of clusters and n is the number of classes. Each entry in the membership matrix , m_{ij}, (which we call membership coefficient) indicates the probability of a vector in the cluster i to belong to the j_{th} class.

$$M = \begin{bmatrix} m_{11} & m_{12} & ... & m_{1n} \\ m_{21} & m_{22} & ... & m_{2n} \\ \vdots & \vdots & & \vdots \\ m_{m1} & m_{m2} & ... & m_{mn} \end{bmatrix} \tag{4.34}$$

where

$$m_{ij} = p(\theta_j|C_i) \tag{4.35}$$

These coefficients will be used in the calculation of the fuzzy membership degree for each of the test vectors. Each segment is represented using 15 feature vectors. By using this approach less weight is associated with the vectors that are in the overlap regions. In the second stage, each of the feature vectors representing a test signal is assigned to one of the cluster centers found in the previous stage based on the minimum Euclidean distance criterion. For each test signal, the scatter vector S is defined as

$$S = [s_1, ..., s_c] \tag{4.36}$$

where s_i is the number of the representing vectors for a test signal that fall within the i_{th} cluster and C is the number of clusters. Finally the probability of a signal belonging to the j_{th} class is calculated according to the distribution of its representing feature vectors in different clusters and can be written as:

$$\Phi(j) = S.M(j) \tag{4.37}$$

4.4.2 Results

The audio data set used in this work consists of 192 signals of about 3s duration, with a sampling rate of 22.05 KHz and a resolution of 16 bits per sample. Table 4.3 shows different sound classes in the data set and the number of signals in each class.

Airplane	Animal	Bird	Drum	Female	Flute	Helicopter	Insect	Male	Piano
20	20	20	20	20	15	17	20	20	20

Table 4.3: Different audio classes in the data set and the number of signals in each class

MP-TFD with the frequency resolution of $M = 250$ is constructed for each audio signal. Once the time-frequency matrix (TFM) is extracted, NMF with decomposition order of 15

($r = 15$) is performed on each TFM. Next, a feature vector comprised of nine features is extracted from each base and coefficient vector.

$$F = \{S_{h_i}, D_{h_i}, MO_{\omega_i}, MO_{h_i}, S_{\omega_i}, SP_{\omega_i}, P_{\omega_i}, D_{\omega_i}, MP\} \tag{4.38}$$

Finally, SOTM is applied on the training dataset and the number of valid clusters is calculated for each classification scenario. One of the advantages of using SOTM is that unlike other clustering approaches such as fuzzy C-means, the exact number of clusters is not needed to be determined beforehand. The clusters are formed as the data is presented to the network and the number and size of the clusters is determined by the parameters such as the hierarchial control function ($H(t)$) and the learning rate ($\alpha(t)$). The initial values of these functions are appointed according to the dataset. In the next stage, the membership coefficients are calculated for each cluster based on the distribution of the train signals. In the test stage, each of the test signals are assigned to one of these cluster centers based on the minimum Euclidean distance measure. Finally, the class label of each signal is determined by the weighted sum of the feature vectors falling within each cluster multiplied by the membership coefficients. Another point to be discussed here is that since the data is represented to the SOTM in a random manner, the number and the shape and size of the clusters might vary each time the clustering algorithm is run on the data. However, since there is not a one to one correspondence between the clusters and the audio classes, this fact has no considerable impact on the total performance of the classifier. In addition, the results of the several are averaged to further eliminate this effect.

One of the most important classification tasks for a hearing aid system is to discriminate human speech form environmental noise. Therefore, in the first scenario the data set consists of signals from human speech and environmental sounds. The human category includes 20 signals from male speakers and 20 signals from female speakers and environmental sounds include 10 bird, 10 aircraft, 10 piano and 10 animal signals. Table 4.4 shows the results for this classification task where an accuracy of 96% has been achieved. As it can be seen from the confusion matrices, the system demonstrates high accuracy in discrimination of human

voice from other audio signals. The achieved true positive rate shows that all human voice signals have been classified correctly. In addition, the overall accuracy rate for classification scenarios that include discrimination of human voice is very high. Furthermore, in order to evaluate the efficiency of the system to discriminate human voice in particular environments, two other classification tasks have been defined. In the first case, an accuracy of 98% has been achieved in discrimination of human voice from the musical instruments. This capability could be useful in recognizing and separation of human voice from the background music in a song or at the concert. The second classification task was defined as discrimination of human voice from natural sounds, where an accuracy of 96% has been achieved. Furthermore, the proposed method was applied to other classification scenarios such as natural vs artificial sounds and musical instruments vs aircraft. The results of these classification tasks are provided in Tables 4.7 and 4.9.

Table 4.5 shows the overall obtained accuracy rate and the data set used for each classification scenario.

	Human	Non-human	Total
Human	40 (100%)	0 (0%)	40 (100%)
Non-human	3 (7.5%)	37 (92.5%)	40 (100%)

Table 4.4: Confusion matrix for classifying human vs non-human audio signals

Classification scenario	Data set	Accuracy rate
Human/non-human	Non-human:aircraft, piano, animal, bird Human: male, female	96%
Human/Music	Human:male, female Music:piano,flute,drum	98%
Natural/Artificial	Natural:male, female, bird, animal, insect Artificial: helicopter, airplane, piano, flute, drum	81%
Human/Nature	Human:male, female Nature:animal, insect, bird	96%
Music/Aircraft	Music:piano, flute, drum Aircraft:helicopter, airplane,	92%

Table 4.5: Different audio classes in the data set and the number of signals in each class

	Human	Musical instruments	Total
Human	40 (100%)	0 (0%)	40 (100%)
Musical instruments	1 (2%)	39 (98%)	40 (100%)

Table 4.6: Confusion matrix for classifying human speech vs musical instruments

	Natural	Artificial	Total
Natural	50 (100%)	0 (0%)	50 (100%)
Artificial	19 (34%)	36 (66%)	55 (100%)

Table 4.7: Confusion matrix for classifying natural vs artificial sounds

	Human	Nature	Total
Human	20 (100%)	0 (0%)	20 (100%)
Nature	3 (15%)	17 (75%)	20 (100%)

Table 4.8: Confusion matrix for classifying human vs nature sounds

	Musical instruments	Aircraft	Total
Musical instruments	34 (75%)	6 (15%)	40 (100%)
Aircraft	0 (0%)	37 (100%)	50 (100%)

Table 4.9: Confusion matrix for classifying musical instrument vs aircraft sounds

Chapter 5

Conclusion

IN this work the application of unsupervised learning for analysis and classification of biomedical signals was investigated. Although there are many works on the application of supervised learning techniques for classification of biomedical data, exploring the application of unsupervised learning methods can be beneficial in many ways. Building a reliable supervised classifier requires a large enough, precisely labeled dataset. However, some biomedical datasets are very large and manual labeling of the data can be extremely costly and time consuming. In such cases, unsupervised learning methods can be used to find the natural groupings (e.g in audio classification)that exist in the dataset and then a physician can label the discovered groups. Furthermore, unsupervised techniques posses more flexibility in situations where the characteristics of the data change over time or the the number of classes is not known beforehand. For example, consider the audio classification task in a hearing aid device. The audio signals that are received by the device contain different audio classes depending on the audio environment. Audio classes that exist in an indoor environment can be different from those that are found in an outdoor environment or at the concert or at a lecture. In such situations where the number and the nature of the classes are not known, a clustering method might perform better than a supervised classifier that is tuned to detect specific classes. In addition, unsupervised classifiers can be used to get some insight about the structure of the data and select more efficient feature extraction methods.

Two classification methods based on clustering techniques was applied to two separate

biomedical signal classification problem. In Chapter 3, fuzzy C-means clustering was applied for classification of small bowel capsule endoscope images and in the Chapter 4 classification of audio signals for hearing aids was investigated. Despite the different classification tasks in the Chapter 3 and Chapter 4, there are commonalities for the two databases. First, the signals in both databases are non-stationary. Second, in both scenarios we are dealing with a large volume of data and lastly in both cases the real-time performance of the algorithms is important. For the hearing aid application, the need for real-time performance is more obvious. No hearing aid user would be interested in a device that amplifies the audio signals with delay. In the case of capsule endoscopy, the real-time performance becomes more critical in the design of the next generation of capsule endoscopes, or the "smart" capsule endoscopy, where the capsule itself contains the drugs and can release the drug wherever it is required in the gastrointestinal tract.

Based on the nature of the classification task in Chapter 4, where the number of audio classes is not known, a classification method based on SOTM clustering algorithm was used to discriminate different audio classes. The advantage of SOTM over other clustering techniques such as fuzz C-means is that in this approach the number of clusters is not required beforehand and this makes the SOTM more suitable for this audio classification task. The discussion and conclusion for each of the chapters is provided in following sections.

5.1 Classification of small bowel images

5.1.1 Results and discussion

In Chapter 3, fuzzy C-means clustering was applied to the problem of detecting abnormalities in the small bowel capsule endoscopy images.

Initially the images were converted to Lab color space. The Lab color space is a perceptually uniform color space and the Euclidean distance measure performs better in this color space. The results provided in Table 3.3 show that the classification accuracy in this color space is better than the rates obtained in the RGB space.

A feature extraction method based on wavelet coefficients and cross co-occurrence ma-

trices was applied to the images. Since the abnormalities might occur at random locations in the image, SIDWT was used for the wavelet decomposition to extract shift-invariant features. The combination of wavelet coefficients and cross co-occurrence matrices was shown to be efficient in the previous works. Four types of features were extracted from the CCM to represent texture characteristics, including energy, homogeneity, texture and contrast. Since the feature extraction process was performed on the three color planes of the image, the extracted features contain color information as well. Different combinations of features were evaluated and the results was provided in Table 3.3. The results for a supervised classifier, which is LDA in this case, is also provided for the same feature set. As it can be observed from the table, the best performance for unsupervised classification was achieved with energy and homogeneity features in the Lab color space. The confusion matrix and receiver operating curve for this feature set is provided in Fig 2.6 and Table 2.2.

An accuracy rate of 76% was achieved for with fuzzy C-means algorithm. Although the results show higher accuracy rates for the supervised classifier, one should bare in mind that the performance of the supervised classifier can be biased by the dataset to some extent. In order for a supervised classifier to be reliable and provide good generalization, it has to be trained on a large enough dataset. However, the number of images in the small bowel data base is 75. Hence, despite the higher accuracy rate the reliability of the supervised classifier yet has to be investigated.

5.1.2 Future work

Although the accuracy rate obtained in this work is acceptable for an unsupervised classifier, other alternatives and modifications can be sought to improve the performance of the system.

In the feature extraction stage, wavelet decomposition followed by the CCM was used to extract color and texture information. Although CCMs have been used successfully in the previous works, they might not be the best solution for small datasets since a large amount of data is generated after the calculations. Hence, a large amount of averaging and down sampling has to be done to decrease the number of features to a reasonable number and this

could cause the loss of information.

Among other texture analysis methods, textons are shown to be effective in representing textural information. Textons have already been used in for texture analysis in biological and biomedical images and have shown promising results. Thus, one of the subjects of the future research work would be to examine alternative feature extraction methods such as textons.

5.2 Classification of audio signals

5.2.1 Results and discussion

In Chapter 4 a classification method based on SOTM clustering algorithm was applied to the classification of audio signals for the hearing aid application. The SOTM is a newly emerged clustering method, which has been already used for segmentation of biological images. In this work however, the classification method is a fusion of supervised and unsupervised classification. Unlike most of the previous works in this area, the features extracted in this work were based on time-frequency analysis of the signals followed by the matching pursuit TFD. Due to the non-stationary nature of the audio signals, temporal or spectral features can not effectively represent localized features of the audio signals such as trends, discontinuities and repeated patterns. TF features on the other hand, are more suitable to capture and represent characteristics of the audio signals. The proposed method was tested under different classification scenarios such as human/non-human, human/music , natural/artificial, human/nature etc. The classification was performed on a database of 10 different audio classes including 20 aircraft, 20 animal, 20 bird, 20 drum, 20 female, 15 flute, 17 helicopter, 20 insect, 20 male and 20 piano signals.

The classification results provided in Table 2.5 show high accuracy rates for most classification scenarios. An accuracy of 96% was achieved for discrimination of human vs non-human sounds, which is the most common classification scenario considered for the hearing aid.

Many methods have been proposed for audio classification for hearing aid. However, most

of the existing papers in the literature address the problem of discrimination of the human voice from the background noise. Although this would be desired capability in a hearing aid, it is not enough for other listening situations such as outdoor, lecture , concert etc. The problem of audio scene analysis is rather a general problem that can be the ultimate goal for the hearing aids.

The classification method used in this work is based on SOTM clustering algorithm. Hence, the number of audio classes is not needed to be known beforehand. This makes the proposed method suitable for the problem of audio scene analyis for hearing aid where the number of audio classes vary under different listening situations.

An efficient classification algorithm that can perform effectively in different audio environments could have a definite application in the hearing aids. According to several surveys, a considerable number of hearing aid users are not satisfied with the performance of their hearing aid since it amplifies the background noise as well as the desired signal. In addition, it has been observed in similar studies that if the quality of the hearing aids can be improved, substantially better user satisfaction can be expected.

5.2.2 Future work

The proposed classification method was tested in different classification scenarios and high accuracy rates were achieved. Nevertheless, the following suggestions can be applied to improve the performance and reliability of the system.

- Although the number of audio classes is not needed beforehand in the classification process, the number of discovered clusters is determined by the parameters in the SOTM algorithm such as $H(t)$ (the hierarchical control function) and $\alpha(t)$ (the reset parameter). The initial values for these parameters affect the number of the discovered clusters and the variance of the samples within each cluster. In this work, these values were adjusted according to the performance of the classifier. Thus,a future improvement for this system would be to find a way to automatically calculate the optimal value of these parameters from the statistical characteristics of the data and

with regard to the classification results.

- The number of clusters found by the SOTM,or any other clustering algorithm in general, does not always represent the actual number of groupings that exist within the dataset. Therefore, a cluster validation technique has to be performed on the results of the clustering to evaluate the validity of the discovered clusters. In this work after the clustering stage, the clusters whose number of samples were smaller than 5% of the total number of samples in the dataset, were recognized as invalid clusters and were eliminated. This threshold was determined based on the performance of the classifier. However, there are more advanced cluster validity techniques that can be adapted for this purpose. So, another area for future work could be to find the best cluster validity measure that optimizes the performance of the classifier.

- In the SOTM algorithm, the representation of the data to the network is in a random manner. Therefore, the results of the clustering might be slightly different for each time the algorithm is run on the dataset. In this work the result of the several runs are averaged to calculate the final results. However, a more robust solution would be to make modifications to the SOTM algorithm or data representation so that the clustering results do not depend on the order in which the data is fed to the SOTM.

- In Chapter 4 different classification scenarios were proposed and tested. The proposed scenarios are based on the taxonomy provided in Fig 4.1 and common listening situations. Another topic for further research in this area would be to design more classification tasks that are tailored for the hearing aid application.

Bibliography

[1] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification.* Wiley New York, 2001.

[2] A. Bousbia-Salah, A. Belouchrani, and A. Cichocki. Application of time-frequency distributions to the independentcomponent analysis of ECG signals. In *IEEE International Symposium on Signal Processing and its Applications (ISSPIT'01)*, volume 1, 2001.

[3] W. Zhou and J. Gotman. Removal of EMG and ECG artifacts from EEG based on wavelet transform and ICA. In *26th IEEE Annual International Conference of the Engineering in Medicine and Biology Society,IEMBS'04.*, volume 1, 2004.

[4] P. Gao, E.C. Chang, and L. Wyse. Blind separation of fetal ECG from single mixture using SVD and ICA. In *Proc. the Joint Conference of the 4th International Conference on Information, Communications and Signal Processing, and the 4th Pacific Rim Conference on Multimedia (ICICS-PCM'03)*, volume 3, pages 1418–1422.

[5] C. Bigan. Chaotic cardiac arrhythmia detection by ICA and nonlinear dynamic processing of ECG signal. In *IEEE International Symposium on Intelligent Signal Processing*, pages 117–120, 2003.

[6] C.A. Joyce, I.F. Gorodnitsky, and M. Kutas. Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology*, 41(2):313–325, 2004.

[7] W. Zhou, J. Zhou, H. Zhao, and L. Ju. Removing eye movement and power line

artifacts from the EEG based on ICA. In *27th Annual International Conference of the Engineering in Medicine and Biology Society. IEEE-EMBS'05.*, pages 6017–6020, 2005.

[8] I. Navarro, B. Hubais, and F. Sepulveda. A comparison of time, frequency and ICA based features and five classifiers for wrist movement classification in EEG signals. In *27th Annual International Conference of the Engineering in Medicine and Biology Society. IEEE-EMBS'05.*, pages 2118–2121, 2005.

[9] L.K.L Joshua and J.C Rajapakse. Extraction of event-related potentials from EEG signals using ICA with reference. In *Proc. IEEE International Joint Conference on Neural Networks. IJCNN'05.*, volume 4, 2005.

[10] MPS Chawla, HK Verma, and V. Kumar. ECG modeling and QRS detection using principal component analysis. In *3rd International Conference On Advances in Medical Signal and Information Processing. MEDSIP'06. IET*, pages 1–4, 2006.

[11] R. Yamada, J. Ushiba, Y. Tomita, and Y. Masakado. Decomposition of electromyographic signal by principal component analysis of wavelet coefficients. In *IEEE EMBS Asian-Pacific Conference on Biomedical Engineering.*, pages 118–119, 2003.

[12] J.U. Chu, I. Moon, S.K. Kim, and M.S. Mun. Control of multifunction myoelectric hand using a real-time EMG pattern recognition.

[13] J.U. Chu, I. Moon, and M.S. Mun. A real-time EMG pattern recognition system based on linear-nonlinear feature projection for a multifunction myoelectric hand. *IEEE Transactions on Biomedical Engineering*, 53(11):2232–2239, 2006.

[14] J. Nadal and RB Panerai. Classification Of Cardiac Arrhythmias Using Principal Component Analysis Of The ECG. In *Proc. the Annual IEEE International Conference of the Engineering in Medicine and Biology Society.*, volume 13.

[15] Y. Wenyu, L. Gang, L. Ling, and Y. Qilian. ECG analysis based on PCA and SOM. In *Proc. the International Conference on Neural Networks and Signal Processing.*, volume 1, pages 37–40.

[16] H. Zhang and L.Q. Zhang. ECG analysis based on PCA and support vector machines. In *Proc. the International Conference on Neural Networks and Brain. ICNN&B'05*, volume 2, pages 743–747.

[17] N. Takano, H.G. Puurtinen, M. Rautiainen, J. Hyttinen, and J. Malmivuo. ECG source location clustering based on position vectors and forward transfer matrices. *Computers in Cardiology,*, pages 313–316, 2002.

[18] O.R Pacheco and F. Vaz. Integrated system for analysis and automatic classification of sleep EEG. In *Proc. the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 4, pages 2062–2065, 1998.

[19] D. Wu and W. Wan Tan. Genetic learning and performance evaluation of interval type-2 fuzzy logic controllers. *Engineering Applications of Artificial Intelligence*, 19(8):829–841, 2006.

[20] A.B Geva and D.H Kerem. Forecasting generalized epileptic seizures from the EEG signal by wavelet analysis and dynamic unsupervised fuzzy clustering. *IEEE Transactions on Biomedical Engineering*, 45(10):1205–1216, 1998.

[21] AB Ajiboye and R.F. Weir. A heuristic fuzzy logic approach to EMG pattern recognition for multifunctional prosthesis control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(3):280–291, 2005.

[22] A.B Ajiboye and R.F Weir. Fuzzy c-means clustering analysis of the EMG patterns of six major hand grasps. In *Proc. the 9th International Conference on Rehabilitation Robotics. ICORR'05.*, pages 49–52.

[23] K. Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007.

[24] M.L Giger, N. Karssemeijer, and S.G Armato. Guest editorial computer-aided diagnosis in medical imaging. *IEEE Transactions on Medical Imaging*, 20(12):1205–1208, 2001.

[25] A. Khademi and S. Krishnan. Multiresolution Analysis and Classification of Small Bowel Medical Images. In *Proc. 29th Annual IEEE International Conference of the Engineering in Medicine and Biology Society. EMBS'07.*, pages 4524–4527, 2007.

[26] B. Li and M.Q.H. Meng. Analysis of the gastrointestinal status from wireless capsule endoscopy images using local color feature. In *Proc. IEEE International Conference on Information Acquisition. ICIA'07.*, pages 553–557, 2007.

[27] J. Bonnel, A. Khademi, S. Krishnan, and C. Ioana. Small bowel image classification using cross-co-occurrence matrices on wavelet domain. *Biomedical Signal Processing and Control*, 4(1):7–15, 2009.

[28] D.J.C. Barbosa, J. Ramos, and C.S. Lima. Detection of small bowel tumors in capsule endoscopy frames using texture analysis based on the discrete wavelet transform. In *30th IEEE Annual International Conference of the Engineering in Medicine and Biology Society. EMBS'08.*, pages 3012–3015, 2008.

[29] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968.

[30] S. Kochkin. 10-year customer satisfaction trends in the US hearing instrument market. *Hearing Review*, 9.

[31] S. Kochkin. " Why my hearing aids are in the drawer": The consumers' perspective. *Hearing Journal*, 53(2):34–42, 2000.

[32] S. Kochkin. MarkeTrak III identifies key factors in determining consumer satisfaction. *Hearing Journal*, 45:39–39, 1992.

[33] K. Umapathy and S. Krishnan. Feature analysis of pathological speech signals using local discriminant bases technique. *Medical and Biological Engineering and Computing*, 43(4):457–464, 2005.

[34] P. Nordqvist and A. Leijon. An efficient robust sound classification algorithm for hearing aids. *The Journal of the Acoustical Society of America*, 115(6).

[35] M. Bchler, S. Allegro, S. Launer, and N. Dillier. Sound classification in hearing aids inspired by auditory scene analysis. *EURASIP Journal on Applied Signal Processing*, 18:2991–3002, 2005.

[36] Adaptive Feature Selection for Speech/Music Classification. *IEEE 8th Workshop on Multimedia Signal Processing.*

[37] L. Lu, H.J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE transactions on speech and audio processing*, 10(7):504–516, 2002.

[38] G. Guo and S.Z. Li. Content-based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Networks*, 14(1):209–215, 2003.

[39] X. Shao, C. Xu, and M.S. Kankanhalli. Unsupervised classification of music genre using hidden Markov model. In *Proc. IEEE International Conference on Multimedia and Expo, ICME'04.*, volume 3.

[40] A. Rauber, E. Pampalk, and D. Merkl. Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity. In *Proc. International Society for Music Information Retrieval Conference ISMIR*, pages 71–80, 2002.

[41] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[42] A.K. Jain and R.C. Dubes. *Algorithms for clustering data.* Printice Hall, 1988.

[43] AK Jain, RPW Duin, and J. Mao. "Statistical pattern recognition: A review". *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):4–37, 2000.

[44] A.K Jain, M.N Murty, and P.J Flynn. Data clustering: a review. *ACM computing surveys*, 31(3), 1999.

[45] D. Judd, PK McKinley, and AK Jain. "Large-scale parallel data clustering". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):871–876, 1998.

[46] S.K Bhatia and J.S Deogun. "Conceptual clustering in information retrieval". *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 28(3):427–436, 1998.

[47] C. Carpineto and G. Romano. "A lattice conceptual clustering system and its application to browsing retrieval". *Machine Learning*, 24(2):95–122, 1996.

[48] H. Frigui and R. Krishnapuram. "A robust competitive clustering algorithm with applications incomputer vision". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):450–465, 1999.

[49] H.M. Abbas and M.M. Fahmy. "Neural networks for maximum likelihood clustering". *Signal Processing*, 36(1):111–126, 1994.

[50] M.J. Kyan. Unsupervised learning through dynamic self-organization: Implications for microbiological image analysis. In *PhD thesis, School of Electrical and Information Engineering University of Sydney*, 2007.

[51] E. Backer. *Computer-assisted reasoning in cluster analysis.* Prentice Hall International Ltd. Hertfordshire,UK, 1995.

[52] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[53] H. Kong and L. Guan. Detection and removal of impulse noise by a neural network guidedadaptive median filter. In *IEEE International Conference on Neural Networks, 1995. Proceedings.*, volume 2, 1995.

[54] GA Carpenter and S. Grossberg. The ART of adaptive pattern recognition by a self-organizing neuralnetwork. *Computer*, 21(3):77–88, 1988.

[55] Given Imaging Ltd. PillCamTM SB Capsule Endoscopy - product information guide. In *World Wide Web, http://www.givenimaging.com/en-us/HealthcareProfessionals/Products/Pages/PillCamSB.aspx*, 2009.

[56] B. Kim, S. Park, C.Y. Jee, and S.J. Yoon. An earthworm-like locomotive mechanism for capsule endoscopes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems.(IROS 2005)*, pages 2997–3002, 2005.

[57] D.G. Adler and C.J. Gostout. Wireless capsule endoscopy. *Hospital Physician*, 39(5):14–22, 2003.

[58] B. Li and M.Q.H. Meng. Analysis of the gastrointestinal status from wireless capsule endoscopy images using local color feature. In *Information Acquisition, 2007. ICIA '07. International Conference on*, pages 553–557, 2007.

[59] B. Li and M.Q.H. Meng. Analysis of the gastrointestinal status from wireless capsule endoscopy images using local color feature. In *Information Acquisition, 2007. ICIA '07. International Conference on*, pages 553–557, 2007.

[60] S. Mallat. *A wavelet tour of signal processing.* Academic press, 1999.

[61] J. Liang and TW Parks. Image coding using translation invariant wavelet transforms withsymmetric extensions. *IEEE Transactions on Image Processing*, 7(5):762–769, 1998.

[62] A. Khademi. Multiresolutional analysis for classification and compression of medical images. Master's thesis, Ryerson University, Canada.

[63] S.A. Karkanis, D.K. Iakovidis, D.E. Maroulis, D.A. Karras, and M. Tzivras. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE Transactions on Information Technology in Biomedicine*, 7(3):141–152, 2003.

[64] V. Arvis, C. Debain, M. Berducat, and A. Benassi. Generalization of the cooccurrence matrix for colour images: application to colour texture classification. *Image Analysis and Stereology*, 23(1):63–72, 2004.

[65] B. Julesz. Textons, the elements of texture perception, and their interactions. 1981.

[66] M. Tuceryan and A.K. Jain. *Handbook of pattern recognition & computer vision.* World Scientific Pub Co Inc, 1999.

[67] H. Harms, U. Gunzer, and HM Aus. Combined local color and texture analysis of stained cells. *Computer vision, graphics, and image processing*, 33(3):364–376, 1986.

[68] S. Petroudi, T. Kadir, and M. Brady. Automatic classification of mammographic parenchymal patterns: A statistical approach. In *Proc. the 25th Annual IEEE International Conference of the Engineering in Medicine and Biology Society.*, volume 1, 2003.

[69] O. Tuzel, L. Yang, P. Meer, and D.J. Foran. Classification of hematologic malignancies using texton signatures. *Pattern Analysis & Applications*, 10(4):277–290, 2007.

[70] D.A. Adjeroh, U. Kandaswamy, and J.V. Odom. Texton-based segmentation of retinal vessels. *Journal of the Optical Society of America A*, 24(5):1384–1393, 2007.

[71] J.M. Kates. Classification of background noises for hearing-aid applications. *The Journal of the Acoustical Society of America*, 97:461, 1995.

[72] L. Rabiner and B.H. Juang. *Fundamentals of speech recognition.* 1993.

[73] J.C. Junqua and J.P. Haton. *Robustness in automatic speech recognition: fundamentals and applications.* Kluwer Academic Publishers Norwell, MA, USA, 1995.

[74] B. Gold and N. Morgan. *Speech & audio signal processing.* Wiley India Pvt. Ltd., 2006.

[75] A. Izworski, R. Tadeusiewicz, and W. Wszolek. Artificial Intelligence Methods in Diagnostics of the Pathological Speech Signals. *Lecture notes in computer science*, pages 740–748, 2004.

[76] Z. Han, X. Wang, and J. Wang. Pathological Speech Deformation Degree Assessment Based on Dynamic and Static Feature Integration. In *The 2nd International Conference on Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008.*, pages 2036–2039, 2008.

[77] M.A. Lund and C.C. Lee. A robust sequential test for text-independent speaker verification. *The Journal of the Acoustical Society of America*, 99:609, 1996.

[78] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[79] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *Proc. International Society for Music Information Retrieval Conference. ISMIR'05*, volume 5, 2005.

[80] N. Scaringella and G. Zoia. On the modeling of time information for automatic genre recognition systems in audio signals. In *Proc.*, pages 666–671.

[81] H. Soltau, T. Schultz, M. Westphal, and A. Waibel. Recognition of music types. In *Proc. the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98*, volume 2, pages 1137–1140.

[82] K. West and S. Cox. Finding an optimal segmentation for audio genre classification. In *Proc. 6th International Symposium on Music Information Retrieval, ISMIR'05*, pages 680–685.

[83] T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. the 6th International Conference on Music Information Retrieval (ISMIR05)*, pages 34–41.

[84] A. Berenzweig, D.P.W. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. In *Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[85] B. Kostek. Musical instrument classification and duet analysis employing music information retrieval techniques. *Proc. the IEEE*, 92(4):712–729, 2004.

[86] C. Joder, S. Essid, and G. Richard. Temporal integration for audio classification with application to musical instrument classification. *IEEE Transactions on Audio, Speech, and Language Processing, ICASSP'09*, 17(1):174–186, 2009.

[87] I. Kaminsky and A. Materka. Automatic source identification of monophonic musical instrumentsounds. In *Proc. IEEE International Conference on Neural Networks*, volume 1, pages 189–194.

[88] K.D. Martin. Toward automatic sound source recognition: identifying musical instruments. *NATO Computational Hearing Advanced Study Institute, Il Ciocco, Italy*, pages 1–12, 1998.

[89] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients andtemporal features. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'00.*, volume 2, pages 753–756.

[90] M.J. Hawley. *Structure out of Sound.* Massachusetts Institute of Technology Cambridge, MA, USA, 1993.

[91] J. Saunders, L.M. Co, and NH Nashua. Real-time discrimination of broadcast speech/music. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,ICASSP'96.*, volume 2, pages 993–996.

[92] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'99*, volume 1, pages 149–152, 1999.

[93] A.S. Bregman. *Auditory scene analysis: The perceptual organization of sound.* The MIT Press, 1994.

[94] G. Freeman, R.D Dony, and S.M Areibi. Audio Environment Classification for Hearing Aids using Artificial Neural Networks with Windowed Input. In *Proc. IEEE Symposium on Computational Intelligence in Image and Signal Processing, CIISP'07*, pages 183–188, 2007.

[95] J. Pinquier, J.L. Rouas, and R. André-Obrecht. Robust speech/music classification in audio documents. In *Proc. the 7th Seventh International Conference on Spoken Language Processing*, volume 3.

[96] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP'97*, volume 2, pages 1331–1334, 1997.

[97] N. Mesgarani, M. Slaney, and SA Shamma. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):920–930, 2006.

[98] SG Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[99] P.J. Franaszczuk, G.K. Bergey, P.J. Durka, and H.M. Eisenberg. Time–frequency analysis using the matching pursuit algorithm applied to seizures originating from the mesial temporal lobe. *Electroencephalography and clinical neurophysiology*, 106(6):513–521, 1998.

[100] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[101] I. Buciu. Non-negative matrix factorization, a new tool for feature extraction: Theory and Applications. In *Proc. the 2nd IEEE International Conference on Computers, Communications and Control, ICCCC'08.*, pages 45–52, 2008.

[102] M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, and R.J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, 52(1):155–173, 2007.

[103] C.J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.