Designing a Global Framework for Non-Linear Dimensionality Reduction

Thesis Submitted in the fulfillment of the requirements for the award of the degree of

Doctor of Philosophy
in
Electronics and Communications Engineering

By
Rashmi Gupta

Under the Supervision of
Prof. Rajiv Kapoor

Faculty of Technology
Delhi College of Engineering
University of Delhi

September 2013

CERTIFICATE

It is certified that the work which is being presented by Rashmi Gupta in the thesis entitled **"Designing a global framework for non-linear dimensionality reduction"** in the fulfillment of the requirements for the award of the degree of Doctor of Philosophy and submitted to the Faculty of Technology, University of Delhi, is an authentic record of her work carried out by her under my supervision. The matter presented in this thesis has not been submitted elsewhere for the award of any other degree.

(Prof. Rajiv Kapoor)
Supervisor/Guide and
Head, ECE Department
DTU (Formerly DCE)
Delhi 110042

H.O.D, ECE, FOT,
Delhi University

DECLARATION

I hereby declare that the work which is being presented in the thesis entitled "Designing a Global Framework for Non-Linear Dimensionality Reduction" is an authentic record of my own work carried out under the supervision of Prof. Rajiv Kapoor for the award of the degree of Doctor of Philosophy and submitted to the Faculty of Technology, University of Delhi.

The research work presented in this thesis has not been submitted elsewhere for the award of any degree.

(Rashmi Gupta)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge and belief.

(Prof. Rajiv Kapoor)
Supervisor/Guide and
Head, ECE Department
DTU (Formerly DCE)
Delhi 110042

Dedicated to My Son

"Ashutosh"

ACKNOWLEDGEMENTS

I express my foremost gratitude to my supervisor Prof. Rajiv Kapoor for his meticulous guidance, logical thinking, sound advice and inspiration throughout this work. His encouragement and constructive criticism, never accepting less than my best, proved to be the pillar of my intellectual growth. His personal support and caring attitude deserves special mention. The interaction and extensive discussions with him on various aspects of dimensionality reduction helped me to learn and enjoy the subject. This work owes much to his intellect, vision, personal involvement and cheerful encouragement. I humbly acknowledge a gratitude to him.

I wish to extend my sincere thank to Prof. Raj Senani, Dean, Board of Research Committee, Faculty of Technology, Delhi University and all the technical supporting staff of Faculty of Technology, Delhi University, for their kind cooperation at each stage of my research work.

My sincere thanks are to the editors and referees of various scientific journals for their detailed review and significant comments on my research papers.

I am also extremely grateful to Prof. Jeffery Cohn, Dr Michael J. Lyons, Aleix Martinez and Robert Benavente and Dr Abhinav Dhall, for providing Cohn–Kanade JAFFE, AR and SFEW database respectively. Their valuable data helped me a lot to carry out my research work.

I thank my colleagues and seniors of MAIT, DCE and AIACT&R for their friendly support and constant encouragement. They helped me at various phases in my research work apart from their technical inputs.

I wish to thank my friends Bhawna, Nidhi, Neha and Kavita for their co-operation throughout my research work.

Thanks to my mother-in-law who is the most kind and non-demanding mother-in-law that one could ever wish for. I thank my parents, sisters and brothers for providing much-needed support and love during difficult times.

No words can express my feelings to my husband Anil Gupta without whom none of this would have been possible. It is his great understanding and boundless support that enables me to achieve my dream. Most importantly my loving thanks to

my son Ashutosh. His naughty acts and innocence always keep me fresh. As a mother, I feel deeply obliged to him, for the long hours which I have deprived him off to do my research.

Last but not the least I would like to thank the Almighty God for his blessings throughout the entire sphere of my life.

Thank you one and all.

Designing a Global Framework for Non-Linear Dimensionality Reduction

Abstract

Submitted in the fulfillment of the requirements
for the award of the degree of

Doctor of Philosophy
in
Electronics and Communications Engineering

By
Rashmi Gupta

Under the Supervision of
Prof. Rajiv Kapoor

Faculty of Technology
Delhi College of Engineering
University of Delhi

September 2013

ABSTRACT

The key tool of dimensionality reduction is that the large set of parameters or features must be summarized into a smaller set, with no or less redundancy. With the emergence of non-linear dynamic systems analysis over recent years it is clear that the conventional approaches for dimensionality reduction may be far from optimal. There are generally no techniques available, especially for finding the features corresponding to contour appearing over different muscles movement. Moreover, the available techniques considering both geometric and discriminant information simultaneously are not computationally efficient and robust. The work takes into account all of the above and few novel dimensionality reduction techniques which have been demonstrated to be more robust than conventional analysis techniques.

In this thesis, a novel framework for non-linear dimensionality reduction is designed to extract non-linear features using the concept that local is non-linear. To detect non linearity, relation between the nearest neighborhood elements of the image, have been expressed in terms of Gaussian membership functions. All the elements of the image are connected with the nearest neighborhood elements with some membership degree of the Gaussian functions. It results in the formation of number of fuzzy lattices. Fuzzy lattices deform according to the various muscles movements. Three fuzzy lattices of maximum kinetic energy corresponding to these contours are sufficient to recognize any object. The technique is based on the concept that any face can be recognized by sketching just few prominent lines corresponding to contours, which are appearing over different muscles movements.

The developed technique based on the concept of local non-linear relation has also been tested on real time data (power quality events) generated by interfacing Fluke 610000A with Laptop via data acquisition system. The generated events are detected and classified using the developed technique based on the concept of local non-linear relation. It extracts any change occurring in the patterns of power quality events. The proposed technique efficiently distinguishes various real time power quality events in a single cycle.

Additional work is an improvement over the well known non-linear dimensionality reduction techniques such as Isomap and Local Linear Embedding. All such methods are based upon the neighborhood information and require a user base input such as constant parameter epsilon or number of nearest neighborhood, which is computational burden. In proposed work, the neighborhood graph is constructed by stacking image in third dimension using Morphmap that reduces the computational complexity. The representation of the depth has been taken into account within small area in these objects by applying the intensity attenuation function.

Another work is designing a novel framework for non-linear dimensionality reduction that considers both discriminant and geometric information of data. It aims to preserve the pairwise geodesic distances between the intraclass separable pairs and to separate the interclass neighbors in the reduced embedding spaces. Based on the extracted information features, large margins between inter and intra class clusters are organized, delivering a strong interclass discriminative power.

Most of the real world data applications such as fingerprint, face or signature recognition suffer from the curse of dimensionality. In order to handle this efficiently, its dimensionality needs to be reduced without much loss of information. Principal Component Analysis is one of the most common and efficient technique for linear dimensionality reduction. However, it is not optimal for classification of data as there is no class discriminatory information in Principal Component Analysis. Thus, Linear Discriminant Analysis could be used to achieve dimensionality reduction along with classification of data classes. Linear Discriminant Analysis works well for distributions which are Gaussian. If the densities are significantly non-Gaussian, Linear Discriminant Analysis may not preserve any complex structure of the data required for the classification. The Marginal Fisher Analysis overcomes this difficulty to a large extent as it uses a different criterion for classification. Furthermore, the Marginal Fisher Analysis with suitable threshold value has been introduced for improving the recognition accuracy and detection of forged signatures.

A new method is proposed for designing wavelet statistically matched to the signal and is applied for data compression. It overcomes the difficulty of choosing the appropriate wavelet from a library of previously designed wavelets. The statistically matched wavelet is designed based on the characteristics of the power quality event

using the concept of fractional Brownian motion. It has been found that the proposed technique is better than Daubechies wavelet in the detection of power quality events. To classify the detected events, an iterative closest point algorithm is used which classifies the detected event even in the presence of outlier points and Gaussian noise.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AR	Aleix Martinez and Robert Benavente
CCA	Curvilinear Component Analysis
CE	Conformal eigenmap
CK	Cohn-Kanade
CL	Cannot Link
DAQ	Data Acquisition
DNA	Deoxyribonucleic Acid
DR	Dimensionality Reduction
EVD	Eigen Value Decomposition
EM	Expectation Maximization
FB-m	Fractional Brownian motion
fMRI	Functional Magnetic Resonance Imaging
FT	Fourier Transform
GPLVM	Gaussian Process Latent Variable Models
ICA	Independent Component Analysis
ICP	Iterative Closest Point
IMs	Induction Motors
Isomap	Isometric Mapping
JAFFE	Japanese Female Facial Expression
KE	Kinetic Energy
KPCA	Kernel Principal Component Analysis
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LDR	Linear Dimensionality Reduction
LE	Laplacian Eigenmaps
LLE	Locally Linear Embedding
LMS	Least Mean Square
LTSA	Local Tangent Space Analysis
MDS	Multidimensional Scaling

MFA	Marginal Fisher Analysis
ML	Must Link
MLA	Multi-Resolution Analysis
MLE	Maximum Likelihood Estimation
MMF	Magnetic Motive Force
Morphmap	Morphological mapping
MVU	Maximum Variance Unfolding
NLDR	Non-linear Dimensionality Reduction
NLM	Non-Linear Mapping
NPE	Neighborhood Preserving Embedding
ORL	Olivetti and Oracle Research Laboratory
PCA	Principal Component Analysis
PQ	Power Quality
RMS	Root mean square
SDE	Semidefinite Embedding
SDP	Semidefinite Programming
SFEW	Static Facial Expressions in the Wild
S-Isomap	Supervised Isomap
SLLE	Supervised LLE
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
SVD	Singular Value Decomposition
SVMs	Support Vector Machines
WT	Wavelet Transform

Chapter-1

Introduction

Chapter-1

Introduction

This chapter provides a brief introduction to dimensionality reduction. The chapter also provides a thorough literature survey and various methodologies related to the field of dimensionality reduction. The motivation behind the work carried out and respective objectives are defined. The chapter further discusses about the brief outlines of the thesis.

1.1 Dimensionality Reduction

High-dimensional datasets present many mathematical challenges as well as some opportunities and are bound to give rise to new theoretical developments. One of the problems with high-dimensional datasets is that, in many cases all the measured variables are not important for understanding the underlying phenomena of interest. While certain computationally expensive novel methods can construct predictive models with high accuracy from high-dimensional data. It is still a point of interest in many applications to reduce the dimensions of the original data prior to modeling of the dataset. The fundamental assumption which forms the basis for dimensionality reduction is the fact that the sample actually lies on a manifold of smaller dimension than the original data space. The goal of dimensionality reduction is to find a representation of that manifold which allows projection of the data vectors on it and obtaining a low-dimensional, compact representation of the data. Often, the original representation of the data is redundant for several reasons. Firstly, many of the variables have a variation smaller than the measurement noise and thus irrelevant. Secondly, many of the variables are correlated with each other and hence a new set of uncorrelated variables have be found. Thus, in many situations it should be possible to somehow strip of the redundant information thereby producing a more economic representation of the data. Dimensionality reduction is important in many domains,

such as data analysis, visualization, data compression, pattern recognition and classification of high-dimensional data.

The next section presents a literature survey reflecting the work done in the field of dimensionality reduction.

1.2 Related Work

Hotelling in [1] proposes a linear dimensionality reduction (LDR) technique called Principal Components Analysis (PCA). It constructs a low dimensional representation of the data that describes maximum variance in the data as possible.

Torgerson in [2] describes the traditional technique of multidimensional scaling (MDS) also known as classical scaling technique. The input into classical scaling is similar to the input into most other multidimensional scaling techniques, i.e., a pairwise Euclidean distance matrix whose entries represent the Euclidean distance between the high dimensional data points.

Roweis in [3] presented a latent variable model called probabilistic PCA. The model uses a Gaussian prior over the latent space along with a linear Gaussian noise model. The probabilistic formulation of PCA leads to an expectation-maximization (EM) algorithm that may be computationally more efficient for very high dimensional data.

Lawrence in [4] extended probabilistic PCA by using Gaussian processes to learn non-linear mappings between the high dimensional and the low dimensional space. PCA and classical scaling have been successfully applied in a large number of domains such as face recognition, coin classification and seismic series analysis.

However, PCA and classical scaling suffer from two main drawbacks. First, in PCA, the size of the covariance matrix is proportional to the dimensionality of the data points. As a result, the computation of the eigenvectors might be infeasible for very high dimensional data. In datasets, where data vectors are less than dimensionality, the problem of large size of covariance matrix may be overcome by performing classical scaling instead of PCA. Because the classical scaling measures data vectors with the number of data points instead of number of dimensions in the data. Alternatively, iterative techniques such as PCA [1] or probabilistic PCA [3] may be employed.

Second, the cost function in PCA and classical scaling focus mainly on retaining large pairwise distances, instead of retaining the small pairwise distances.

MDS finds an embedding that preserves the inter-point distances, equivalent to PCA when those distances are Euclidean [2]. LDA, a supervised learning algorithm selects a transform matrix in such a way that the ratio of the between-class scatter and the within-class scatter is maximized [79]. PCA seeks a projection that best represent the data in a least-squares sense [5, 6]. PCA assume the existence of a linear map between the data points and the parameterization space, such a map often does not exist as many data sets contain essential non-linear structures. Applying LDR techniques to data therefore results in a distorted representation. If the input patterns are distributed more or less throughout this subspace, the eigen value spectra from these methods also reveal the data set's intrinsic dimensionality. A more interesting case arises when the input patterns lie on or near a low dimensional submanifold of the input space. In this case, the structure of the data set may be highly non-linear, and linear methods are bound to fail.

In the last decade, a large number of non-linear dimensionality reduction (NLDR) techniques have been proposed [10-13]. They all utilized local neighborhood relation to learn the global structure of non-linear manifolds. But they have quite different motivation and objective functions.

If the high dimensional data lies on or near a curved manifold, such as in the Swiss roll dataset, classical scaling might consider two data points as near points, whereas their distance over the manifold is much larger than the typical inter point distance.

Tenenbaum et al. in [14] describes an approach called Isomap that resolves this problem by attempting to preserve pairwise Geodesic distances between data points. However, Isomap faces several imperfections such as that its algorithm suffers with topological instability [15]. Isomap may construct erroneous connections in the neighborhood graph. Such short-circuiting [16] can severely impair the performance of Isomap. Several approaches have been proposed to overcome the problem of short-circuiting, by removing data points with large total flows in the shortest-path algorithm [17] or by removing nearest neighbors that violate local linearity of the neighborhood graph. Isomap may also suffer from holes in the manifold. This problem can be dealt

with by tearing manifolds with holes. Another weakness of Isomap is that it can fail if the manifold is non convex. Despite these weaknesses, Isomap has been successfully applied on tasks such as wood inspection, visualization of biomedical data, and head pose estimation.

Sch¨olkopf et al. in [18] presented Kernel PCA that is the reformulation of traditional linear PCA in a high-dimensional space. It is constructed using a kernel function. In recent years, the reformulation of linear techniques using the kernel trick has led to the proposal of successful techniques such as kernel ridge regression [19]. Kernel PCA computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix. The reformulation of PCA in kernel space is straightforward, since a kernel matrix is similar to the product of the data points in the high-dimensional space that is constructed using the kernel function.

Kernel PCA though suffers on the account is that the size of the kernel matrix is proportional to the square of the number of instances in the dataset. An approach to resolve this weakness is proposed in [20]. Also, Kernel PCA mainly focuses on retaining large pairwise distances. Kernel PCA has been successfully applied to large number of problems such as face recognition, speech recognition, and novelty detection.

Weinberger et al. in [21] proposed a technique called Maximum Variance Unfolding (MVU, formerly known as Semidefinite Embedding) that attempts to resolve this problem by learning the kernel matrix. MVU learns the kernel matrix by defining a neighborhood graph on the data and retaining pairwise distances in the resulting graph. MVU is different from Isomap in that explicitly attempts to unfold the data manifold. It does so by maximizing the Euclidean distances between the data points, under the constraint that the distances in the neighborhood graph are left unchanged (i.e., under the constraint that the local geometry of the data manifold is not distorted). The resulting optimization problem can be solved using semidefinite programming. MVU has a weakness similar to Isomap i.e., short-circuiting may impair the performance of MVU because it adds constraints to the optimization problem that prevent successful unfolding of the manifold. Despite this weakness, MVU has been successfully applied to sensor localization and Deoxyribonucleic acid (DNA) micro array data analysis.

Roweis and Saul, in [22] present a Local Linear Embedding (LLE) technique that is similar to Isomap and MVU as it constructs a graph representation of the data points. However, in contrast to Isomap, it attempts to preserve solely local properties of the data. As a result, LLE is less sensitive to short-circuiting than Isomap, since only a small number of local properties are affected if short-circuiting occurs. Furthermore, the preservation of local properties allows for successful embedding of non-convex manifolds.

The popularity of LLE has led to the proposal of linear variants of the algorithm [23-24] and to successful applications to super resolution and sound source localization. A possible explanation lies in the difficulties that LLE has when it is confronted with manifolds that contain holes. In addition, LLE tends to collapse large portions of the data very close together in the low-dimensional space, because the covariance constraint on the solution is too simple. Also, the covariance constraint may give rise to undesired re-scaling of the data manifold in the embedding [25].

Similar to LLE, Belkin and Niyogi, in [26] find a low-dimensional data representation by preserving local properties of the manifold. In Laplacian eigenmaps (LE), the local properties are based on the pairwise distances between near neighbors. LE computes a low-dimensional representation of the data in which the distances between a data point and its nearest neighbors are minimized. This is done in a weighted manner, i.e., the distance in the low-dimensional data representation between a data point and its first nearest neighbor contributes more to the cost function than the distance between the data point and its second nearest neighbor. Using spectral graph theory, the minimization of the cost function is defined as an eigen value problem. LE suffers from many of the same weaknesses as LLE, such as the presence of a trivial solution that is prevented from being selected by a covariance constraint. Despite these weaknesses, LE and its variants have been successfully applied to face recognition [27]. In addition, variants of LE may be applied to supervised or semi-supervised learning problems [28]. A linear variant of LE is presented in [29]. In spectral clustering, clustering is performed based on the sign of the coordinates obtained from LE [30].

Donoho and Grimes, in [31] present Hessian LLE that is a variant of LLE that minimizes the curviness of the high-dimensional manifold when embedding it into a

low-dimensional space, under the constraint that the low-dimensional data representation is locally isometric. This is done by an eigen analysis of a Hessian matrix that describes the curviness of the manifold around the data points. The curviness of the manifold is measured by means of the local Hessian at every data point. The local Hessian is represented in the local tangent space at the data point, in order to obtain a representation of the local Hessian that is invariant to differences in the positions of the data points. It can be shown that the coordinates of the low-dimensional representation can be found by performing an eigen analysis of an estimator Hessian matrix of the manifold Hessian.

Hessian LLE shares many characteristics with LE. It simply replaces the manifold Laplacian by the manifold Hessian. As a result, Hessian LLE suffers from many of the same weaknesses as LE and LLE. A successful application of Hessian LLE to sensor localization has been presented by [32].

Similar to Hessian LLE, Zhang and Zha, in [33] describe the approach Local Tangent Space Analysis (LTSA) which is a technique that describes local properties of the high dimensional data using the local tangent space of each data point. LTSA is based on the observation that if local linearity of the manifold is assumed there exists a linear mapping from a high dimensional data point to its local tangent space. And the linear mapping from the corresponding low dimensional data point to the same local tangent space also exits [34]. LTSA attempts to align these linear mappings in such a way that they construct the local tangent space of the manifold from the low dimensional representation. In other words, LTSA simultaneously searches for the coordinates of the low dimensional data representations and for the linear mappings of the low dimensional data points to the local tangent space of the high dimensional data. Like the other sparse spectral DR techniques, LTSA may be hampered by the presence of a trivial solution in the cost function.

The non-linear methods mentioned are efficient at visualizing artificial data sets and powerful to handle non-linear data. However, they are unsupervised method, so fails to identity the types inter or intraclass of neighborhoods and unable to handle discriminatory information. To address these issues, supervised Isomap (S-Isomap) and supervised LLE (SLLE) have been proposed by enabling the inclusion of class labels directly [35-36]. SLLE guides the discriminant learning by increasing the pre

obtained distances artificially between interclass points and leaving the distances unchanged for those intraclass points. S-Isomap drives the discriminant learning through defining a new distance metric to enhance interclass dissimilarity over intraclass similarity. The idea of SLLE and S-Isomap is to pick the neighbors of each point from the same class and then separate interclass points through improving intraclass compactness. Satisfactory results are reported if data sets are well sampled with relatively convex intrinsic geometry. SLLE and S-Isomap are not so powerful for handing multiple class real cases. Pairwise Cannot-Link and Must-Link constraints [37-38] induced from the neighborhood graph into the Isomap are incorporated to guide the discriminant manifold learning.

In contrast to the traditional linear techniques, the non-linear techniques have the ability to deal with complex non-linear data. On the other hand, such approaches also has several limitations such as the solutions do not yield an estimate of the underlying manifold's dimensionality, the geometric properties preserved by these embedding are difficult to characterize and the resulting embeddings sometimes exhibit an unpredictable dependence on data sampling rates and boundary conditions. Moreover, the original LLE, Isomap and LE cannot deal with the out-of-sample problem [40] directly. Out-of-sample problem states that only the low dimensional embedding map of training samples can be computed but the samples out of the training set cannot be calculated at all. Hessian LLE is a variant of LLE that learns distance-preserving embeddings, with theoretical guarantees of asymptotic convergence [41].

NLDR methods attempt to describe a given high-dimensional set of points as a low dimensional manifold by means of a non-linear map preserving certain properties of the data. Many NLDR techniques attempt to find a low dimensional representation for the data while preserving local properties. For example, the LLE algorithm tries to preserve the representation of each data point as a linear combination of its neighbors [22]. The Laplacian Eigenmaps algorithm uses the Laplacian operator for selecting low dimensionality coordinate functions based on its eigen functions [86]. The diffusion map generalizes this framework in the context of analysis of diffusion processes, making it more robust to non-uniform sampling density [87]. The Hessian LLE tries to use the proximity graph for finding coordinate functions that have a

minimal response to the Hessian operator of the surface, obtaining a truly locally linear mapping.

Apart from these methods, Local Binary Pattern (LBP) as a novel low-cost image descriptor for texture classification has also been introduced to the field of facial expression analysis [47, 48]. LBP can efficiently encode the texture features of micro-pattern information in the face image which is effective information for both face recognition and facial expression recognition applications. The kernel sliced inverse regression algorithm generates a non-linear learning model in the original input pattern space. It reproduces kernel Hilbert space setting and emphasize on combining with other linear algorithms [49]. Vector based representations ignore the spatial structure of the image data which may be very useful for visual recognition. In tensor representation, Discriminant multi-linear projections are pursued to construct the Discriminant embedding [50].

The graph based DR algorithms on facial expression recognition are compared by employing leave one-subject-out strategy for cross validation [51]. A commonly used approach to improve robustness in classifying expression is to combine the results of several different methods [52]. Two hybrid facial expression recognition systems are proposed that employ the one-against-all classification strategy [53]. The first system decomposes the facial images into linear combinations of several basis images using Independent Component Analysis (ICA). Subsequently, the corresponding coefficients of these combinations are fed into Support Vector Machines (SVMs) that carry out the classification process. The basic description of SVMs can be phrased as a two class classification problem where data points are mapped into a high dimensional hyperspace so that they can be separated by a hyper plane [54]. A margin exists on each side of the hyper plane which is distanced to the nearest set of data points of each class. A high margin indicates good separation and good generalization. The multiclass SVMs problem solves only one optimization problem [55]. It constructs the basic facial expressions rules that separate training vectors of the class from the rest of the vectors by minimizing the objective function.

The developed techniques for non-linear dimensionality reduction have also been tested for detection of power quality (PQ) events. These events cover a broad frequency range with significantly different magnitude variations and can be non-

stationary. Thus, accurate techniques are required to detect and classify these events. The major key issues and challenges in classifying PQ events are critically analyzed and presented. The selection of suitable features is extremely important for classification of any problem. An appropriately chosen feature set reduces the burden over the classifiers. This chapter also includes a comprehensive survey of various techniques which are used in PQ event detection and classification.

Allen et al, in [56] and Altes et al, in [57] used Fourier Transform (FT) for extracting the frequency contents of the recorded signal. According to the frequency contents of the signal, some of the PQ problems can be detected. But FT is not suitable for non-stationary signals because it provides information only about the existence of a certain frequency component and does not give time information. A suitable way to obtain such information is to apply time-frequency (or time scale) signal decomposition where time-evolved signal components in different frequency bands can be obtained.

Although, Short Time Fourier Transform (STFT) can partly alleviate this problem, but it still has the limitation of a fixed window width. The trade-off between the frequency resolution and the time resolution should be determined a priori to observe a particular characteristic of the signals [58]. Due to a fixed window width, STFT is inadequate for the analysis of the transient non-stationary signals. Therefore, more powerful and efficient techniques are required to detect and analyze non-stationary disturbances.

To resolve the fixed resolution problem of STFT, many researchers have proposed the use of the Wavelet Transform (WT) approach to analyzing the power system disturbances [59], [60], [61], [62]. The WT approach prepares a window that automatically adjusts to give proper resolutions of both the time and the frequency. In this approach, a larger resolution of time is provided to high-frequency components of a signal, and a larger resolution of frequency to low-frequency components. These features make the WT well suited for the analysis of the power system transients caused by various PQ disturbances. A model using adaptive wavelet networks has been proposed for the PQ event detection [63]. The reconstructed version of the original signal has been used by discarding all of the coefficients of few higher level details to detect and localize the disturbances in the presence of noise [64]. But

discarding all of the coefficients of higher resolution levels has a risk of losing high-frequency transient features, completely or partially, depending upon their travel beyond the discarded scales. An approach has been proposed for the PQ disturbances classification based on the wavelet transform and self organizing learning array system [65]. The discrete wavelets transform and artificial neural network with fuzzy logic has been exploited for the characterization and classification of PQ events [66]. The authors have used Hilbert transform for feature extraction of distorted waveform [67, 68]. The Hilbert Transformer gives a better approximate only if the signal approaches a narrow band condition.

Dash et al, in [69] introduced S-transform as a new PQ signal analysis and feature extraction tool. Chilukuri and Dash, in [70] extracted the features of seven simulated signals by calculating the minima and maxima of the S-transform absolute matrix. The S-Transform based probabilistic neural network model has been proposed [71-72]. Similar to the STFT, S-transform also requires significant amount of computational resources. This is due to the fact that the S-transform matrix is calculated by performing the inverse.

Keeping in view the trends, developments and limitations mentioned above, the work in the thesis has been carried out.

1.3 Motivation

From the review of literature survey, it is observed that further there is a lot of scope for improvement of existing techniques and development of novel techniques in the field of dimensionality reduction. All the reported work has the issues like higher order of complexity, presence of trivial optimal solutions, and large number of features. The systems need to be more robust and to have other discriminant features to work as a real system. The development of other linear as well as non-linear dimensionality reduction techniques is required which seek for better criteria for classification among various datasets. The above said issues create limitations to the reported technique for dimensionality reduction. In view of this author has endeavored to take up the following objectives that contribute to the field of dimensionality reduction.

1.4 Objectives

(i) To develop a general framework for dimensionality reduction that considers both discriminant and geometric information of data.

(ii) To improve stress function for efficiently handle non-linear manifolds.

(iii) To develop Morphomap for non-linear dimensionality reduction techniques.

(iv) To improve the Marginal Fisher's Analysis for detection of forged signatures.

(v) To design a novel framework for non-linear dimensionality reduction to extract prominent features using fuzzy lattices.

(vi) Development of wavelet matched to the signal for application of data compression.

1.5 Outline of Thesis

The rest of the thesis is organized as follows:

Chapter 2 introduces improved Marginal Fisher Analysis for signature recognition. Linear Discriminant Analysis fails when the discriminatory information is not in the mean but rather in the spread of the data. The Marginal Fisher Analysis overcomes this difficulty to a large extent as it uses a new criterion for classification. Moreover, the Marginal Fisher Analysis with suitable threshold value has been introduced for improving the recognition accuracy and detection of forged signatures.

Chapter 3 discusses and compares spatial distance preserving techniques for reducing the dimensionality of the data. The preservation of the pairwise distances measured in a data set ensures that the low dimensional embedding inherits the geometric properties of the data like local neighborhood relationships.

Chapter 4 reviews and compares the graph-based methods for manifold learning. The methods are efficient at visualizing artificial data sets and powerful to handle non-linear data. However, these are unsupervised methods and hence fails to identity the interclass or intraclass types of neighborhoods and unable to handle discriminatory information. To address these issues, constraint Isomap is proposed in this chapter that provides geometrical as well as discriminatory information of data. It

enhanced both interclass separation and intraclass compaction and delivering clear separation on the manifold embedding of multiple classes.

Chapter 5 describes the extensions and analysis of local non-linear techniques. The original LLE, Isomap and Laplacian eigenmaps cannot deal with out of sample problem and cannot preserve local feature such as angle. To overcome the limitations of existing methods, extensions of local non-linear techniques have been discussed in this chapter. In the first proposed method, a low dimensional embedding is constructed that maximally preserves angles between nearby data points. Second proposed method, minimizes the cost function of a local non-linear technique for dimensionality reduction under the constraint that the mapping from the high-dimensional to the low-dimensional data representation is linear.

Chapter 6 focuses on characterizing the strict efficient solution to recognize prominent features of a person. The method is based on the concept of non-linear relation between the nearest neighborhood elements of the image. To detect non-linearity, Gaussian membership functions have been used which results in formation of fuzzy lattices. Three fuzzy lattices having maximum energy are selected to extract the prominent features. Compared to earlier framework for analyzing high dimensional data that lie on or near a low dimensional manifold the proposed method has interesting property of representing any object with small set of features.

Chapter 7 seeks to provide the generation of PQ events. The generated events are detected and classified using the very new developed technique based on the concept of local non-linear relation. It extracts any change occurring in the patterns of PQ events. The proposed technique efficiently distinguishes various real time PQ events in a single cycle.

Chapter 8 intends to compress the PQ data using statistically matched wavelet. The proposed method overcomes the difficulty of choosing the appropriate wavelet for a given application and presents a new approach for compression of PQ data. To classify the detected events, Iterative Closest Point algorithm is used which classifies detected event even in presence of outlier points and Gaussian noise.

Chapter 9 focuses on improvement over the well known non-linear dimensionality reduction techniques such as Isomap and LLE. Both of the methods require a user base input or calculation of nearest neighborhood elements which is

itself a computational burden. In the proposed method, the neighborhood graph construction is done by stacking image in third dimension using Morphological mapping that reduces the computational burden to a great extent.

Chapter 10 summarizes the major findings of the entire investigations which also emphasizes the future scope for research in this area.

Chapter-2

Linear Dimensionality Reduction Techniques and their Extension

Chapter-2

Linear Dimensionality Reduction Techniques and their Extension

Most of the real world data applications such as fingerprint, face or signature recognition suffer from the curse of dimensionality. In order to handle this efficiently, its dimensionality needs to be reduced without much loss of information. Principal Component Analysis (PCA) is one of the most common and efficient technique for linear dimensionality reduction. However, it is not optimal for classification of data as there is no class discriminatory information in PCA. Linear Discriminant Analysis (LDA) could be used to perform dimensionality reduction along with classification of data classes. LDA works well for distributions which are Gaussian or similar in nature. If the densities are significantly non-Gaussian, LDA may not preserve any complex structure of the data required for the classification. The Marginal Fisher Analysis (MFA) overcomes this difficulty to a large extent as it uses a different criterion for classification. Moreover, the improved MFA has increased the recognition accuracy and detection of forged signatures by setting a suitable threshold value.

2.1 Introduction

From automated speech recognition, fingerprint identification, optical character recognition, DNA sequence identification and much more, it is clear that reliable accurate pattern recognition by machine would be immensely useful [89]. But the real world data suffers from the problem of high dimensions. With the advent of high quality signature capture devices, signature is attracting more attention as a biometric to develop practical applications. The difficulties inherent to signature based authentication are related to the great variability of signatures. Furthermore, the forgers can reproduce signatures with high resemblance to the user's signatures. Hence, main concern is the dimensionality reduction that can be beneficial not only

for reasons of computational efficiency but also to improve the accuracy of the analysis. The chapter is organized as follows: Section 2.2 describes the theoretical characteristics and algorithms of the spatial distance preservation based techniques for dimensionality reduction. Section 2.3 details the improved Marginal Fishers Discriminant Analysis as an extension of linear dimensionality reduction. Section 2.4 discusses the results of the experiments on all described techniques for dimensionality reduction. Finally, conclusions are drawn in Section 2.5.

2.2 Linear Dimensionality Reduction Techniques

In this section, linear dimensionality reduction (LDR) techniques such as PCA, LDA and improved MFA have been discussed for signature recognition.

2.2.1 Principal Component Analysis

The goal of PCA is to reduce the dimensionality of the data while retaining as much as possible the variation present in the original dataset. PCA allows computing a linear transformation that maps the data from a high dimensional space to a lower dimensional sub-space [5-6]. Dimensionality reduction leads to information loss. Thus, it aims to preserve as much information as possible. PCA projects the data along the directions where the data varies the most. These directions are determined by the eigenvectors of the covariance matrix corresponding to the highest eigenvalues.

2.2.1.1 Algorithm: PCA

Suppose $x_1, x_2, ..., x_M$ are N x 1 vectors.

Step 1: Find mean of the data:

$$\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i$$

Step 2: Find zero mean matrix:

$$\Phi_i = x_i - \bar{x}$$

Step 3: Form the matrix $A = [\Phi_1 \Phi_2 ... \Phi_M]$, then compute co-variance matrix (C) that characterize the scatter of data:

$$C = \frac{1}{M} \sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T$$

Step 4: Compute eigenvalues and eigenvectors of C:

$$(C - \lambda_i I)\, e_i = 0$$

Step 5: Select eigenvectors corresponding to the highest eigenvalues:

$$e_i = V$$

Step 6: Projecting higher dimensional data into lower dimensional space.

$$\Phi_i = V^T \Phi_i$$

PCA is one of the most common and efficient technique for LDR. It minimizes the reconstruction error. Although, PCA efficiently reduces the number of dimensions but it does not have any class discriminatory information. Thus, data classification cannot be done using PCA. Also, reduction in dimension can only be achieved if the original variables are correlated. If the original variables are uncorrelated, PCA only helps in ordering them according to their variance. Hence, PCA cannot be used when need is to perform the dimensionality reduction along with class discrimination. LDA technique which is discussed in the next section can be used for dimensionality reduction along with class discrimination.

2.2.2 Linear Discriminant Analysis

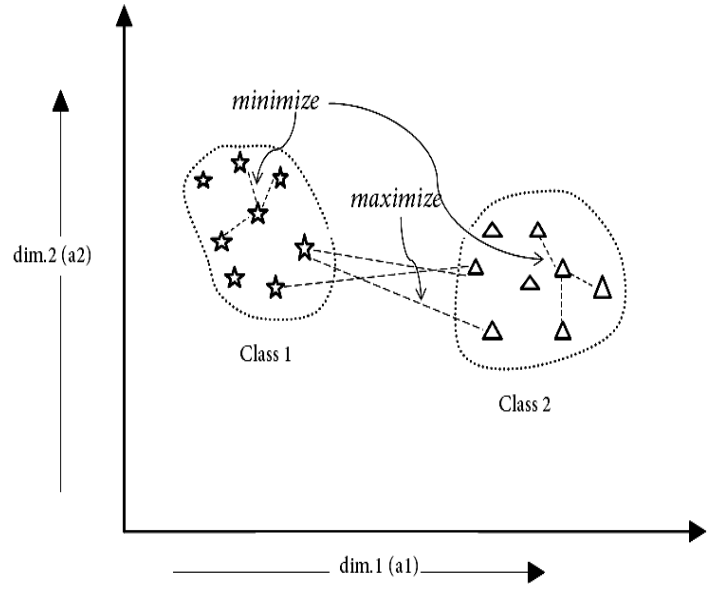
LDA is basically a method used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects. The main objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible [8]. The Fisher Linear Discriminant is defined as the linear function that maximizes the criterion function as follows:

$$J(w) = \frac{\left| u_{1p} - u_{2p} \right|^2}{S_{1p}^2 + S_{2p}^2} \tag{2.1}$$

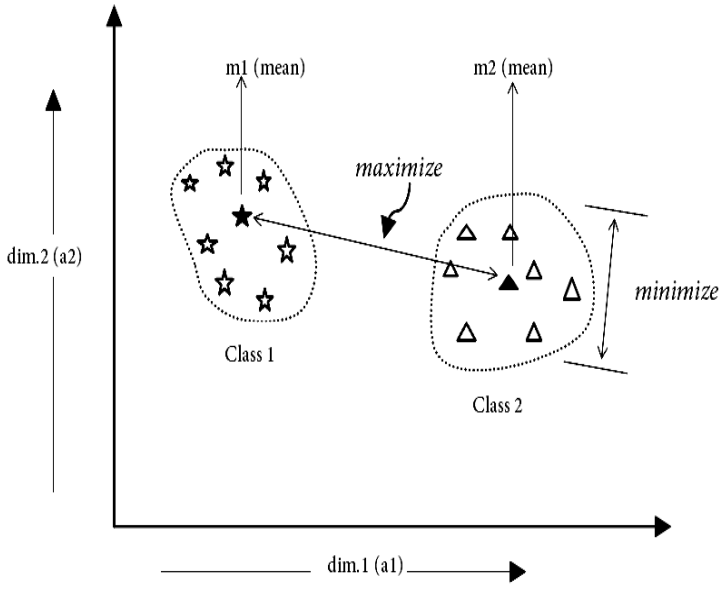
where u_{1p} is mean of the class w_i in projected feature space (y-space).

S_{ip}^2 is the variability within class w_i after projection on y space

$S_{1p}^2 + S_{2p}^2$ is the variability within two classes after projection



(a)



(b)

Figure 2.1 Graphical Illustration of Techniques (a) LDA (b) MFA

A simple illustration of LDA has been shown in Figure 2.1(a). In this Figure, it is described that LDA focuses on the means of the classes for discrimination. It separates the means of various classes as farther apart as possible and also tries to make the scatter of individual classes as compact as possible. Thus, LDA reduces the dimensionality as well as separating the two data classes as farther apart as possible.

2.2.2.1 Algorithm: LDA

Step 1: In order to find a good projection vector, there is need to define a measure of separation. The mean vector of each class in x and y space is defined as follows:

$$\mu_i = \frac{1}{N_i} \sum_{x \in w_i} x$$

where u_i is means of the class w_i in original feature space (x-space).

$$\mu_{iP} = \frac{1}{N_i} \sum_{y \in w_i} y = \frac{1}{N_i} \sum_{x \in w_i} w^T x = w^T \frac{1}{N_i} \sum_{x \in w_i} x = w^T \mu_i$$

where N_i is the number of data points in i^{th} class.

Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space.

$$\begin{aligned}(u_{1p} - u_{2p})^2 &= (w^T u_1 - w^T u_2)^2 \\ &= w^T (u_1 - u_2)(u_1 - u_2)^T w \\ (u_{1p} - u_{2p})^2 &= S_{bp} = w^T S_b w \end{aligned} \tag{2.2}$$

where S_b is between classes scatter matrix of the original features vectors and

S_{bp} is between classes scatter matrix of the projected features vectors.

Step 2: The measure of the scatter in multivariate feature space is denoted as scatter matrices can be defined as follows:

$$S_i = \sum_{x \in w_i} (x - \mu_i)(x - \mu_i)^T$$

$$S_w = S_1 + S_2,$$

where S_i is the co-variance matrix of class w_i and S_w is known as class scatter within the original feature space.

Now, the scatter of the projection y can then be expressed as a function of the scatter matrix in feature space x as follows:

$$\begin{aligned} S_{ip}^2 &= \sum_{w_i} (y - u_{ip})^2 = \sum_{w_i} (w^T X - \mu_i)^2 \\ &= \sum_{w_i} (w^T X - \mu_i)(w^T X - \mu_i)^T w \\ &= w^T S_i w \end{aligned}$$

$$S_{1P}^2 + S_{2P}^2 = w^T (S_1 + S_2) w = w^T S_w w = S_{wp} \tag{2.3}$$

where S_{wp} is the within class scatter in the projected feature space y

Step 3: The Fisher criterion can be expressed in terms of S_w and S_b using (2.1), (2.2) and (2.3) as follows:

$$J(w) = \frac{\left| u_{1p} - u_{2p} \right|^2}{{S_{1p}}^2 + {S_{2p}}^2} = \frac{w^T S_b w}{w^T S_w w} \tag{2.4}$$

Hence $J(w)$ is a measure of the difference between class means (encoded in the between class scatter matrix) normalized by a measure of the within class scatter matrix.

Maximum $J(w)$ is found by differentiating and equating (2.4) to zero and we get:

$$\frac{\partial}{\partial w} J(w) = \frac{\partial}{\partial w} \left(\frac{w^t S_b w}{w^t S_w w} \right) = 0$$

Step 4: Solving the generalized eigenvalues problem:

$${S_w}^{-1} S_b w = \lambda w$$

Yields

$$w^* = {S_w}^{-1} (\mu_1 - \mu_2) \tag{2.5}$$

Step 5: The low dimensional embedding is given as follows:

$$Y = (w^*)^T X \tag{2.6}$$

This low dimensional embedding provides class discrimination information assuming data of each class has Gaussian distribution. Marginal fisher Analysis (MFA) technique discussed in next section can be used for dimensionality reduction along with class discrimination for all types of the data.

2.3 Extension of Linear Dimensionality Reduction Techniques

LDA is developed based on the assumption that the data of each class has a Gaussian distribution. However, the property of Gaussian distribution often does not exist in real world problems. To overcome this limitation of LDA, MFA is proposed by developing a new criterion that characterized the intraclass compactness and the interclass separability to obtain the optimal transformation [90-91].

2.3.1 Marginal Fisher Analysis and its Extension

In MFA, the intrinsic graph characterizes the intraclass compactness, which connects each data point and its neighboring points of the same class. The interclass graph characterizes the interclass separation, which connects the marginal points. In the low-dimensional space, MFA tries to keep neighboring points close if they have the same label and prevents points of other classes from entering the neighborhood [90].

The difference between the criteria of LDA and classical MFA for class discrimination is clearly illustrated in Figure 2.1 (a) and (b). LDA focuses on the means of the classes for discrimination as it separates the means of various classes as farther apart as possible and also tries to make the scatter of individual classes as compact as possible at the same time. MFA takes care of the boundaries of the neighboring classes to avoid intermixing. It makes the neighboring points as farther apart as possible if they have different class labels, and hence, increasing the inter class separability. It keeps the neighboring points as close as possible if they have same class label, and hence, reducing intraclass scatter. This criterion makes MFA independent of type of data distribution.

Here, the threshold value has been defined for detection of forged signatures.

$$\text{Threshold } (t) = \min\left[\frac{\{(d_1 + d_2 + ... + d_n) - (d_m)\}}{(n-1)} - d_m + (S.Dev.) + (d_{\max} - d_m)\right] \quad (2.7)$$

where d_m = minimum among the Euclidean distance

$d_{\max}$ = maximum among the Euclidean distance

$d_1, d_2, ..., d_n$ = Euclidean distances between the projected sample point of the pattern to be tested and the projected samples of the patterns existing in the database.

2.3.1.1 Algorithm: Improved MFA

Step 1: Find the K nearest neighbors (KNN) per datapoint.

Step 2: Construct intraclass graph:

$$w_{ij}^{w} = \begin{Bmatrix} 1; & if\ x_i \in N_{k1}^{+}(x_j)\ or\ x_j \in N_{k1}^{+}(x_i) \\ 0; & \in otherwise \end{Bmatrix} \quad (2.8)$$

$N_{k1}^{+}(x_j) = k_1$ nearest neighbourhood of x_j of same class

Intraclass compactness is characterized by:

$$J_w(A) = \sum_{i,j} \left\| A^T x_i - A^T x_j \right\|^2 w_{ij}^{w} = 2\left[tr\left\{ A^T \left(X\left(D^w - W^w \right) X^T \right) A \right\} \right]$$

$$= 2\left[tr\left\{ A^T \left(X\left(L^w \right) X^T \right) A \right\} \right] = 2\left[tr\left\{ A^T \left(Z^w \right) A \right\} \right] \quad (2.9)$$

Where Xi is the data matrix, Z^w is the intraclass compactness matrix.

D^w is the diagonal matrix with $d_{ij}^{w} = \sum_j w_{ij}^{w}$

Step 3: Construct interclass graph:

$$w_{ij}^{b} = \begin{Bmatrix} 1; & if\ x_i \in N_{k2}^{-}(x_j)\ or\ x_j \in N_{k2}^{-}(x_i) \\ 0; & \in otherwise \end{Bmatrix} \quad (2.10)$$

$N_{k2}^{-}(x_j) = k_2$ nearest neighbours of x_j of different classes

Interclass seperability is defined by:

$$J_b(A) = \sum_{i,j} \left\| A^T x_i - A^T x_j \right\|^2 w_{ij}^{b} = 2\left[tr\left\{ A^T \left(X\left(D^b - W^b \right) X^T \right) A \right\} \right]$$

$$= 2\left[tr\left\{ A^T \left(X\left(L^b \right) X^T \right) A \right\} \right] = 2\left[tr\left\{ A^T \left(Z^b \right) A \right\} \right] \quad (2.11)$$

Where Z^b is the interclass seperability matrix

D^b is the diagonal matrix with $d_{ij}^b = \sum_j w_{ij}^b$

Step 4: Eigen decomposition of the matrix $Z_w^{-1} Z^b$

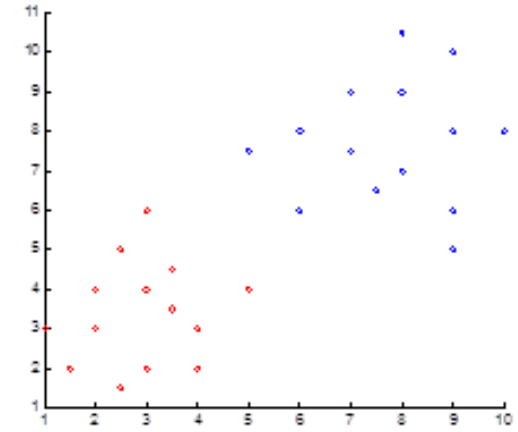
Step 5: Check for detection of forged signatures using (2.7)

For every test, if $d_1 = d_m$ and $t_i > t$, it corresponds to recognized pattern. If $t_i < t$, it corresponds to the forged signature.

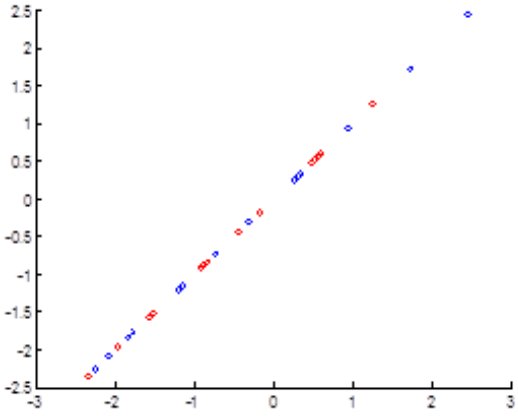
2.4 Results

The performance of PCA, LDA and improved MFA techniques has been checked for classification of data of two classes and signature recognition. MFA with suitable threshold value has also been tested for detection of forged signatures. The data points of two different classes represented by different colors have been shown in Figure 2.2(a). The red dot shows data belongs to one class and blue dot shows data belongs to other class. The data is two dimensional (2-D) and objective is to convert this into one dimension (1-D) by projecting the sample points in such a way that the two classes could be discriminated as clearly as possible.

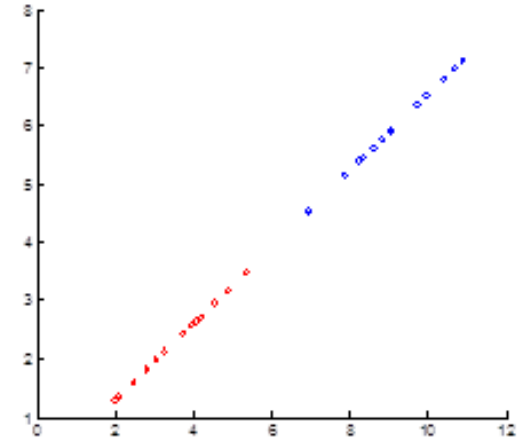
The projected data points with reduced dimension after applying PCA and LDA on data samples have been shown in Figure 2.2 (b) and Figure 2.2 (c) respectively. From the Figures, it is observed that there is an intermixing among the data samples of different classes in PCA. In LDA, data samples of different classes are well separated. Thus, performance of LDA is much better than PCA for classification.



(a)



(b)



(c)

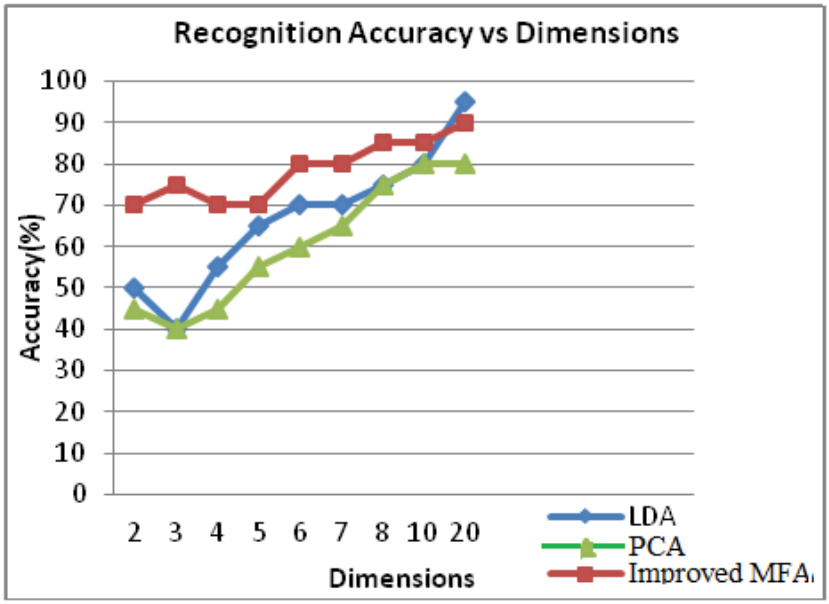
Figure 2.2 Projection of sample points in (a) original 2-D space (b) 1-D space using PCA c) 1-D space using LDA

Figure 2.3 Sample images from database of signature images

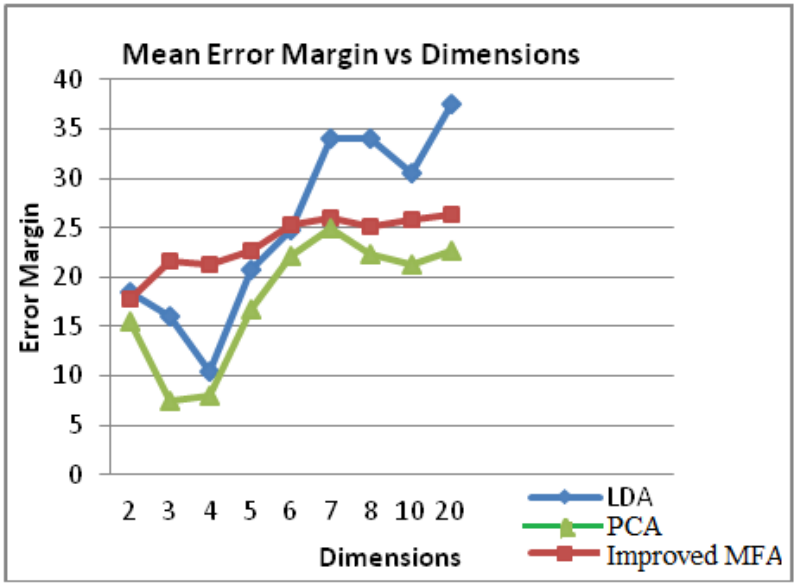
Figure 2.4 Sample images from database of forged signature images

The proposed algorithm is also tested for signature recognition and detection of forged signatures. Thirty signature classes have been taken as database and each class consists of twenty signatures. Only a small variation in the size, position, and orientation of the objects in the images are allowed. The recognition is based on the Euclidean distances of the projected sample point of the testing pattern to the projected samples of the existing patterns. Sample images from signature databases and forged signature images are shown in Figure 2.3 and 2.4 respectively.

The sample point corresponding to the minimum Euclidean distance from the pattern to be tested is selected as the recognized pattern. If $d_i = d_m$ (where d_i is any value among $d_1, d_2, ..., d_n$) and $t_i > t$, it corresponds to the recognized pattern.



(a)



(b)

Figure 2.5 Plots of PCA, LDA and Improved MFA (a) Recognition Accuracy (b) Error Margin

Table 2.1 Comparison between PCA, LDA and Improved MFA for signature recognition

Methods	Dimensions	2	3	4	5	6	7	8	10	20
PCA	Accuracy (%)	45	40	45	55	60	65	75	80	80
	Mean error margin	16.49	7.11	7.85	16.7	21.13	24.35	23.05	22.60	24.52
	Execution time (sec.)	7.95	8.61	8.60	8.92	9.1	9.75	9.76	9.90	10.94
LDA	Accuracy (%)	50	40	55	65	70	70	75	80	85
	Mean error margin	17.79	16.21	10.5	20.7	24.73	34.15	34.05	30.60	37.56
	Execution time (sec.)	7.90	8.31	8.65	8.72	8.81	8.95	9.02	9.10	9.90
Improved MFA	Accuracy (%)	70	75	70	70	80	80	85	85	90
	Mean error margin	17.75	21.65	21.29	22.66	25.33	26.02	25.13	25.80	26.35
	Execution time (sec.)	9.10	9.35	9.76	9.96	10.25	10.75	11.16	11.60	12.51

The error margin in recognition gives the value that if subtracted from t_i, gives the correct result. Thus, more the error margin, less prone is the result to error. The negative value in error margin means that the particular result is incorrect. The results of recognition accuracy, mean error margin and execution time for PCA, LDA and improved MFA are depicted in Table 2.1.

The recognition accuracy and error margin are ploted with respect to the number of dimensions in Figures 2.5 (a) and (b) respectively. From the Figure 2.5 and Table 2.1, it can be observed that improved MFA outperforms both PCA and LDA for all the dimensions (from 2 to 20) in terms of signature recognition. The results of mean error margin in improved MFA are better than LDA in lower dimensions. Improved MFA is performing better than both PCA and LDA at little extra computational most.

2.5 Conclusions

PCA is an efficient technique for linear dimensionality reduction. However, it is not optimal for classification of data. LDA finds the projection that maximizes the ratio of between-class scatter to the within-class scatter to achieve class discrimination. While in MFA the optimal projection is obtained by maximizing the ratio of inter-class separation to intra-class compactness. For classification, performance of MFA is significantly better than LDA. In LDA, it is assumed that the data of each class has a Gaussian distribution while MFA is applicable for all types of data. Moreover, the Marginal Fisher Analysis with suitable threshold value improves the recognition accuracy and detection of forged signatures.

Chapter-3

Spatial Distance Preservation based Techniques

Chapter-3

Spatial Distance Preservation based Techniques

This chapter deals with methods that reduce the dimensionality of data by using distance preservation as criteria. The preservation of the pairwise distances measured in a data set ensures that the low dimensional embedding inherits the main geometric properties of the data, like the global shape or local neighborhood relation. In this chapter, spatial distance preserving techniques such as Multidimensional Scaling, Sammon's non-linear mapping and Curvilinear Component Analysis have been discussed and compared for dimensionality reduction. They rely on different optimization procedures to determine the embedding.

3.1 Introduction

The distance preservation is the first criterion used to achieve dimensionality reduction in a non-linear way [92]. In the linear case, simple criterion like maximizing the variance preservation or minimizing the reconstruction error, combined with a basic linear model, lead to robust methods like PCA. In the non-linear case, the use of same simple criteria requires the definition of more complex data models, which is a little bit difficult. In this context, distance preservation appears as a non-generative way to perform dimensionality reduction [93]. The motivation behind distance preservation is that any manifold can be fully described by pairwise distance. Hence if low-dimensional representation can be built in such a way that the initial distances are reproduced, then the dimensionality reduction is successfully achieved. If close points are maintained and if far points remain far, then the initial manifold and its low dimensional embedding share the same shape [94-95]. This is the basic approach, which is used in techniques discussed in this chapter. This chapter is organized as follows: Section 3.2 describes the details of CCA and Sammon's non-linear mapping. Experimental results are shown in Section 3.3. Finally in Section 3.4, the conclusions are drawn.

3.2 Spatial Distance Preservation based Techniques

Spatial distances, like the Euclidean distance, are the most natural way to measure distances in the real world. The word spatial indicates that these metrics compute the distance of two separating points of space, without considering to any other information like the presence of a submanifold. In this computation, only the coordinates of two points matter. The next three sections review the spatial distance preservation based techniques.

3.2.1 Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) is a spatial distance preserving technique and is a strictly linear technique. In its classical version, metric MDS preserves pairwise scalar products instead of pairwise distances. It relies on a simple generative model. More precisely, only an orthogonal axis separates the observed variable in x and the latent ones stored in y**,** can be represented mathematically as follows:

$$y = Wx \tag{3.1}$$

Here, **x** is the data points in high dimensional data space whose dimensions is D, y is the data point in embedded space whose dimension is d, and W is the transformation matrix of size D x d.

Let us consider that the total numbers of data points are N. Then in matrix form it can be written as follows:

$$\mathrm{J}_w(A) = \sum_{i,j} \left\| A^T x_i - A^T x_j \right\|^2 w_{ij}^w = 2\left[tr\left\{ A^T \left(X\left(D^w - W^w \right) X^T \right) A \right\} \right] \tag{3.2}$$

$\mathrm{X} = [\ldots, \mathrm{x(i)}, \ldots, \mathrm{x(j)}, \ldots]$

The scalar product between vectors $\mathrm{x(i)}$ and $\mathrm{x(j)}$ is given by:

$$s_x(i,j) = s(x(i), x(j)) = \langle x(i).x(j) \rangle \tag{3.3}$$

The eigen value decomposition of the Gram matrix **S** is done as follows:

$$S = U \Lambda U^T \tag{3.4}$$

$$= (U\Lambda^{1/2})(\Lambda^{1/2}U^T)$$

$$= (\Lambda^{1/2}U^T)(\Lambda^{1/2}U^T)$$

where '**U**' is an N-by-N orthonormal matrix and 'Λ' is an N-by-N diagonal matrix containing the eigen values. If the eigen values are sorted in descending order, then the estimated d-dimensional latent variables can be computed as follows:

$$\hat{Y} = I_{d\times N}\Lambda^{1/2}U^T \tag{3.5}$$

To minimize the error, Error Stress Function is defined as follows:

$$E = \sum_{i,j=1}^{N} (S_x(i,j) - \langle \hat{Y}(i).\hat{Y}(j)\rangle)^2 \tag{3.6}$$

If instead of Gram matrix, given data consists of Euclidean distance then it has to be converted into Gram matrix. In terms of norms, Euclidean distance can be defined as Scalar Product given below:

$$\begin{aligned} d_x^2(i,j) &= \|\mathbf{x}(i) - \mathbf{x}(j)\|^2 \\ &= \langle (\mathbf{x}(i) - \mathbf{x}(j)).(\mathbf{x}(i) - \mathbf{x}(j)) \rangle \\ &= \langle \mathbf{x}(i).\mathbf{x}(i) - 2\langle \mathbf{x}(i).\mathbf{x}(j)\rangle + \langle \mathbf{x}(j).\mathbf{x}(i)\rangle \rangle \\ &= s_x(i,i) - 2s_x(i,j) + s_x(j,j) \end{aligned}$$

Thus, Scalar Product can be computed as follws:

$$s_{\mathbf{x}}(i,j) = -\frac{1}{2}\left(d_{\mathbf{x}}^2(i,j) - \langle \mathbf{x}(i).\mathbf{x}(i)\rangle - \langle \mathbf{x}(j).\mathbf{x}(j)\rangle \right) \tag{3.7}$$

Assume that pairwise distances are squared and stored in an N x N matrix **D**, then:

$$\mathbf{D} = [d_x^2(i,j)]_{1 \le i,j \le N} \tag{3.8}$$

The operation of D is called as Double Centering. It simply consists of subtracting from each entry of **D,** the mean of the corresponding row and the mean of the corresponding column, and adding back the mean of all entries. In matrix form it can be written as follows:

$$\mathbf{S} = -\frac{1}{2}(\mathbf{D} - \frac{1}{\mathbf{N}}\mathbf{D}\mathbf{1}_N\mathbf{1}_N^T - \frac{1}{\mathbf{N}}\mathbf{1}_N\mathbf{1}_N^T\mathbf{D} + \frac{1}{N^2}\mathbf{1}_N\mathbf{1}_N^T\mathbf{D}\mathbf{1}_N\mathbf{1}_N^T) \tag{3.9}$$

This is the transformation of Euclidean distance into Gram Matrix.

3.2.1.1 Algorithm: MDS

Step 1: If available data consist of vectors gathered in X, then center them, compute the pairwise scalar products $S = X^T X$, and go to step 3.

Step 2: If available data consist of pairwise Euclidean distance, transform them into scalar products:

(i) Square the distances and build D.

(ii) Perform the double centering of D using (3.9), this yields S.

Step 3: Compute the eigen value decomposition using (3.4).

Step 4: The low-dimensional representation is obtained by computing the product using (3.5).

MDS is suitable for linear data. In the following section, spatial techniques for non- linear dimensionality reduction have been discussed.

3.2.2 Sammon's Non-Linear Mapping

It is a non-linear technique. The concept of Sammon's Non-Linear Mapping (NLM) is closely related to MDS. But in this case, no generative model is used only a stress function is defined. Consequently, the low dimensional representation can be totally different from the distribution of the true latent variables. Sammon's NLM minimizes the following stress function.

$$E = \frac{1}{c}\sum_{i=1, i<j}^{N} \frac{(d_x(i,j) - d_y(i,j))^2}{d_x(i,j)} \tag{3.10}$$

where $d_y(i,j)$ is a distance measured between i_{th} and j_{th} points in the low dimensional space, $d_x(i,j)$ is a distance measured between the i_{th} and j_{th} points in the high dimensional space and normalizing constant C is defined by (3.11).

$$c = \sum_{i=1, i<j}^{N} d_x(i,j) \tag{3.11}$$

The factor $\frac{1}{d_x(i,j)}$ in (3.10), which is not in case of MDS, is weighting the summed terms, which gives less importance to errors made on large distances. More precisely, the weighting factor simply adjusts the importance to be given to each distance in Sammon's stress, according to its value. The preservation of long distances is less important than the preservation of shorter ones, and therefore, the weighting factor is chosen to be inversely proportional to the distance. The optimization technique, which is used to minimize above function is Quasi-Newton optimization form, is iterative in nature. From the concept of Quasi-Newton update rule, the parameter $y_k(i)$ can be updated as follows:

$$y_k(i) \leftarrow y_k(i) - \alpha \frac{\frac{\partial E}{\partial y_k(i)}}{\frac{\partial^2 E}{\partial y_k(i)^2}} \tag{3.12}$$

where α is called as magic factor and Sammon recommends its value between 0.3 and 0.4.

Sammon's NLM minimizes a stress or error function, which is defined as follows:

$$\begin{aligned}
\frac{\partial E}{\partial y_k(i)} &= \frac{\partial E}{\partial d_y(i,j)} \frac{\partial d_y(i,j)}{\partial y_k(i)} \\
&= \frac{-2}{c} \sum_{j=1, j\neq i}^{N} \left(\frac{d_x(i,j) - d_y(i,j)}{d_x(i,j)} \right) \frac{\partial d_y(i,j)}{\partial y_k(i)} \\
&= \frac{-2}{c} \sum_{j=1, j\neq i}^{N} \left(\frac{d_x(i,j) - d_y(i,j)}{d_x(i,j)} \right) \frac{(y_k(i) - y_k(j))}{d_y(i,j)} \\
&= \frac{-2}{c} \sum_{j=1, j\neq i}^{N} \left(\frac{d_x(i,j) - d_y(i,j)}{d_x(i,j) d_y(i,j)} \right) (y_k(i) - y_k(j))
\end{aligned} \tag{3.13}$$

Similarly, the second derivative can be obtained as follows:

$$\frac{\partial^2 E}{\partial y_k^2(i)} = \frac{-2}{c} \sum_{j=1, j\neq i}^{N} \left(\frac{d_x(i,j) - d_y(i,j)}{d_y(i,j) d_x(i,j)} - \frac{(y_k(i) - y_k(j))^2}{d_y^3(i,j)} \right) \tag{3.14}$$

Sammon's NLM involves various parameters, especially due to its iterative optimization scheme. These are number of iterations and the magic factor α, also it is noteworthy that initialization may play a part in the final result.

3.2.2.1 **Algorithm:** Sammon's NLM

Step 1: Compute all pair wise distances $d_x(i, j)$ in the D-dimensional data space.

Step 2: Initialize the d-dimensional coordinates of all points $y(i)$, either randomly or on the hyper plane spanned by the first d principal components of the data set.

Step 3: Compute the right hand side of (3.12) for the coordinates of all points $y(i)$ and update the coordinates of all points of $y(i)$.

Step 4: Return to step 3 until the value of stress function does not decreases further.

Compared to classical metric MDS, Sammon's NLM linear mapping can efficiently handle non-linear manifolds, at least if they are not too heavily doped.

3.2.3 Curvilinear Component Analysis

The Curvilinear Component Analysis (CCA) is the first method to combine vector quantization [93] with a non-linear dimensionality reduction achieved by distance preservation. Like dimensionality reduction, vector quantization can be defined as a way to reduce the size of a data set. However, instead of lowering the dimensionality of the observation, vector quantization reduces the number of observation.

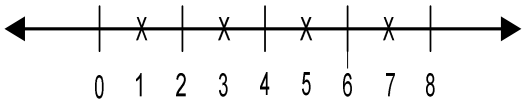


Figure 3.1One Dimensional Vector Quantization

Vector quantization is basically an optional preprocessing of the data. It can be applied to reduce the number of data points in large databases. For small databases or sparsely sampled manifolds, however, it is often better to skip vector quantization in order to fully exploit the available information. In order to reduce the data points, round-off value or mean value between the various data points can be taken. The concept of vector quantization is shown in Figure 3.1

CCA minimizes a stress or error function, which is defined as follows:

$$E = \frac{1}{2} \sum_{i=1, j=1}^{N} ((d_x(i,j) - d_y(i,j))^2 F_\lambda(d_y(i,j)) \tag{3.15}$$

To maintain the global shape of the manifold, preserving the short distance is required as compared to longer distance. Thus, F_λ is typically chosen as a monotonically decreasing function of its argument. While in Sammon's stress function, the weighting depends on the constant distance measured in the data space.

The optimization procedure, which is used to determine the minimization of (3.15) can be calculated as follows:

$$\frac{\partial E}{\partial y_k(i)} = \frac{\partial E}{\partial d_y} \frac{\partial d_y}{\partial y_k(i)} \quad \frac{\partial E}{\partial y_k(i)} = \frac{\partial E}{\partial d_y} \frac{\partial d_y}{\partial y_k(i)}$$

$$\nabla_{y(i)} E = \sum_{j=1}^{N} (d_x - d_y)(2F_\lambda(d_y) - (d_x - d_y)F'_\lambda(d_y)) \frac{y(j) - y(i)}{d_y} \tag{3.16}$$

where $\nabla_{y(i)} E$ represents the gradient of E with respect to vector $y(i)$. The minimization of E by a gradient descent gives the following update rule.

$$y(i) \leftarrow y(i) - \beta \nabla_{y(i)} E \tag{3.17}$$

where β is a positive learning rate scheduled according to the Robbins-Monro condition.

The embedding of highly folded manifolds requires focusing on short distances. Longer distances have to be stretched in order to achieve the unfolding and their

contribution must be lowered in stress function E. Therefore, F_λ is usually chosen as a positive and decreasing function. For example

$$F_\lambda(d_y) = \exp\left(-\frac{d_y}{\lambda}\right) \tag{3.18}$$

where λ controls the decrease.

3.2.3.1 Algorithm: CCA

Step 1: Perform the vector Quantization to reduce the size of data set, if needed.

Step 2: Compute all pairwise Euclidean distances $d_x(i, j)$ in the high-dimensional data space.

Step 3: Initialize the d-dimensional coordinates of all the points $y(i)$, either randomly or on the hyper plane spanned by the first principal components. Let q be equal to 1.

Step 4: Give the learning rate β and the neighborhood width λ their scheduled value of epoch no. q.

Step 5: Select a point $\mathbf{y}(i)$, and update all other ones according to update rule using (3.17).

Step 6: Return to step 5 until all points $y(i)$ have been selected exactly once during the current epoch.

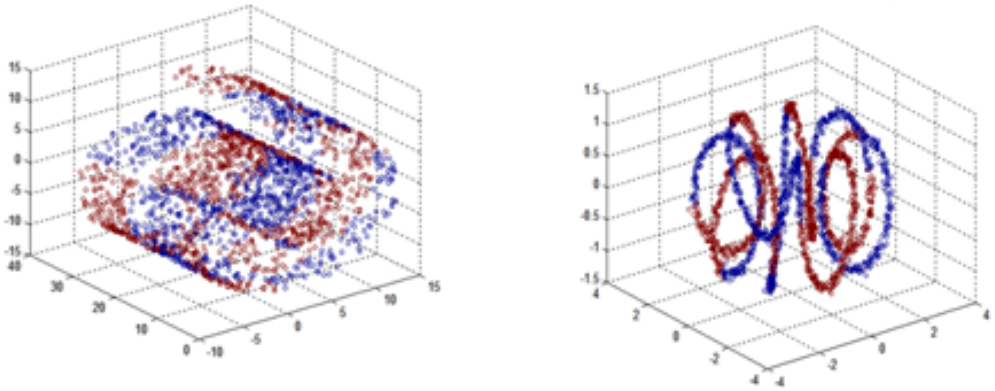
Step 7: Increase q, and if convergence is not reached return to step 4.

Comparing with Sammon's NLM, CCA proves much more flexible, mainly because the user can choose and parameterize the weighting function F_λ. This allows limiting the range of considered distances and focusing on the preservation of distances on a given scale only. Moreover the weighting function F_λ depends on the distances measured in the embedding space, this results in tearing some regions of the manifold. This is better solution than crushing the manifold.

3.3 Results

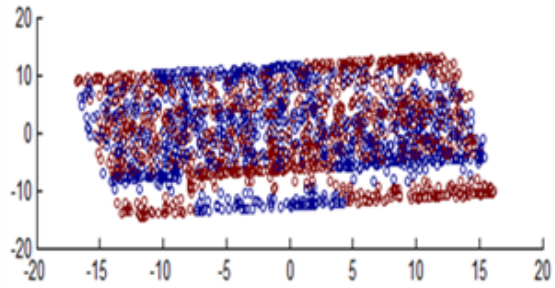
In this section, the results of the spatial distance preservation based techniques

such as MDS, Sammon's non-linear mapping and CCA are tested on artificially generated data set.

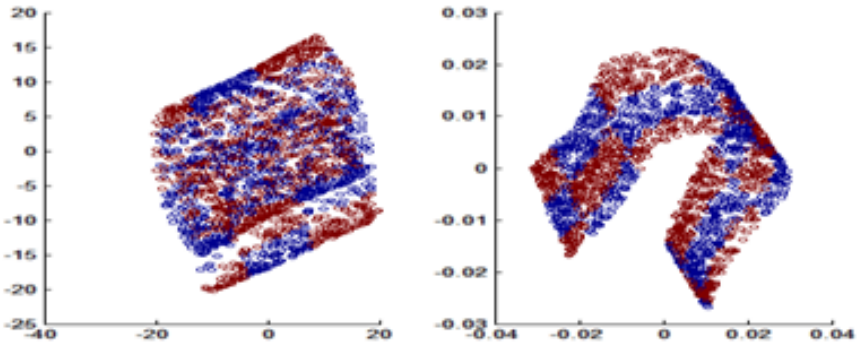


(a) (b)

Figure 3.2 Artificial generated data set (a) Swiss Roll (b) Helix



(a)



(b) (c)

Figure 3.3 Results of Dimensionality Reduction on Swiss Roll Data Set (a) MDS (b) Sammon's NLM (c) CCA

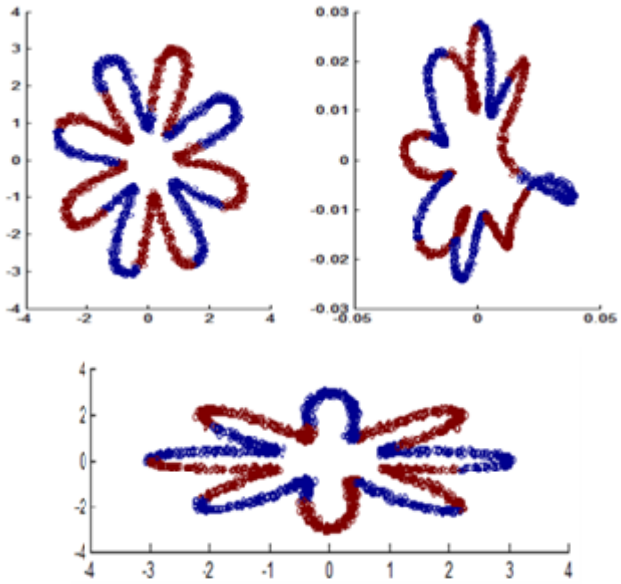


Figure 3.4 Results of Dimensionality Reduction on Helix Data Set (a) MDS (b) Sammon's NLM (c) CCA

Figure 3.2 shows the artificially generated dataset such as Swiss roll and helix. The artificial generated data set consists of 2,000 samples. The result of MDS, CCA and Sammon's NLM on Swiss roll and Helix Data set are shown in Figures 3.3 and 3.4 respectively. From results, it is observed that MDS is not able to unfold the manifold. There is overlapping between the data points in embedded space as compare to original data space. In case of Helix data set, the results of Sammon's NLM are disappointing. But the results of CCA are much more convincing. From Figures 3.3 and 3.4, it is observed that the results are almost superposition free in CCA. Euclidean distance between the data points is maintained in a much closer sense in the embedding space too. This is the first criteria from visualization point of view. Other parameters of comparison are time complexity and space complexity.

Table 3.1 Performance Comparison between MDS, NLM and CCA

	MDS	Sammon's NLM	CCA
Vector Quantization	No	No	Could be used
Space Complexity	$O(N^2)$	$O(N^2)$	$O(P^2)$
Time Complexity	$O(N^2D)$	$O(N^2d)$	$O(N^2P)$
Embedding	disappointing	turns are superposed	superposition free

Table 3.1 depicts the summary of comparison between the three techniques. In CCA, space complexity is much less as compared to Sammon's NLM because the number of data points gets reduced after vector quantization. Here, data points are reduced from N to P. Time complexity for both the data set is also less in CCA as compared to Sammon's NLM and MDS.

3.4 Conclusions

Compared to MDS, Sammon's NLM can efficiently handle non-linear manifolds, at least if they are not heavily doped. As a main drawback, NLM lacks the ability to generalize the mapping to new points. Another shortcoming of NLM is its optimization procedure, which may be slow or inefficient for some data sets. Thus, comparative analysis of Sammon's NLM and CCA shows that CCA is much more flexible because the user can choose and parameterize the weighting function. From a computational point of view, the optimization procedure of CCA works much better than the quasi-Newton rule of NLM. On the other hand, the interpretation of CCA error criteria is difficult, since weighing function is changing when CCA is running.

Chapter-4

Graph based Techniques and their Extensions

Chapter-4

Graph Based Techniques and their Extension

The graph based techniques have been used extensively for non-linear dimensionality reduction. In this chapter, graph-based techniques namely Isomap, Maximum Variance Unfolding, Local linear Embedding and Laplacian Eigenmaps have been reviewed and compared for manifold learning. These techniques are efficient at visualizing artificial data sets and powerful to provide geometrical information of data. However, these are unsupervised technique, so fails to identity inter or intraclass types of neighborhoods and unable to provide discriminatory information. To address these issues, constraint Isomap is proposed that provides geometrical as well as discriminatory information of data.

4.1 Introduction

In contrast to the traditional linear techniques, the non-linear techniques have ability to deal with complex non-linear data. For real world data, the non-linear dimensionality reduction techniques may offer an advantage, because real world data is likely to form a highly non-linear manifold. Previous studies have shown that non-linear techniques outperform their linear counter parts on complex artificial tasks. For instance, the swiss roll dataset comprises a set of points that lie on a spiral like two-dimensional manifold that is embedded within a three-dimensional space. A vast number of non-linear techniques are perfectly able to find this embedding, whereas linear techniques fail to do so.

Motivated by the lack of a systematic comparison of graph based dimensionality reduction techniques, this chapter presents a comparative study of the most important graph based dimensionality reduction techniques: Isomap, Maximum Variance Unfolding, Locally Linear Embedding, and Laplacian Eigenmaps. Though capable of revealing highly non-linear structure, graph-based techniques for manifold learning are based on highly tractable polynomial time optimizations such as shortest path

problems, least squares fits, semidefinite programming, and matrix diagonalization. In [35, 36], S-Isomap and SLLE has been proposed by inclusion of class label. In this work, pairwise Cannot-Link (CL) and Must-Link (ML) constraints [38] induced from the neighborhood graph into the Isomap are incorporated to guide the discriminant manifold learning. More importantly, pairwise constraints sets are flexible in regulating the supervised information.

The chapter is organized as follows: Section 4.2 describes the theoretical characteristics and algorithms of the graph based techniques for dimensionality reduction. Section 4.3 describes the constraint Isomap as an extension of graph based techniques for non-linear dimensionality reduction. Section 4.4 discusses the results of experiments of all described techniques on artificially generated dataset and real face datasets. Moreover, it identifies weaknesses and points of improvement of the mentioned non-linear techniques. Finally, conclusions are drawn in Section 4.5.

4.2 Graph Based Techniques

The non-linear dimensionality reduction techniques find meaningful hidden low dimensional structure in high dimensional space. If the data is confined to a low dimensional subspace, then simple linear methods can be used to discover the subspace and estimate its dimensionality. More generally, though, if the data lies on or near a low dimensional submanifold, then its structure may be highly non-linear, and linear techniques are bound to fail. Graph based techniques have recently emerged as a powerful tool for non-linear dimensionality reduction and manifold learning. These techniques are able to reveal low dimensional structure in high dimensional data from the top or bottom eigenvectors of specially constructed matrices. To analyze data that lies on a low dimensional submanifold, the matrices are constructed from sparse weighted graphs whose vertices represent input patterns and whose edges indicate neighborhood relations. This section provides the review of four such graph based learning algorithms Isomap, Maximum Variance Unfolding, Local Linear Embedding, and Laplacian Eigenmaps for non-linear dimensionality reduction.

4.2.1 Local Linear Embedding

In Local Linear Embedding (LLE), the local properties of the data manifold are constructed by writing the high-dimensional data points as a linear combination of their nearest neighbors [22].

4.2.1.1 Algorithm: LLE

The input **X** is a matrix

$X = \{X_1, X_2, ...X_N\}$, Where $\mathbf{X_i} \in R^D$

Where D is the number of dimensions of the input data

The output **Y** is a matrix

$Y = \{Y_1, Y_2, ...Y_N\}$, Where $\mathbf{Y}_i \in R^d$ and d << D

Where d is the number of dimensions of the output data.

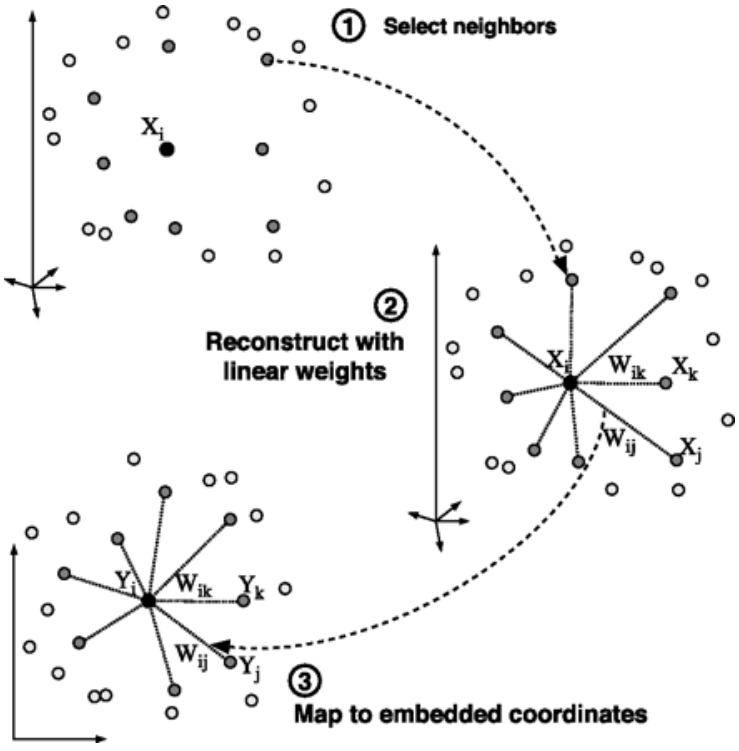


Figure 4.1 Pictorial Representation of LLE

For each vector $\mathbf{x_i}$, following three steps has to be repeated:

Step 1: Find K nearest neighbors $\{Xi_1, X_{i2}, ...X_{iN}\}$, per data point using Euclidean distance as shown in Figure 4.1.

Step 2: It is assumed that the manifold is well sampled, i.e., there are enough data and each data point and its nearest neighbors lie on or close to a locally linear patch of the manifold. The $\mathbf{x_i}$ can be approximated by a linear combination of its neighbors. The weight matrix W has to be found between pair of neighbors using (4.1).

$$W = \{W_{ij}, i = 1, 2, ..., N; j = 1, 2, ..., K\}, \tag{4.1}$$

That minimizes cost function given as follows:

$$\varepsilon(W) = \sum_{i=1}^{N} \left| \mathbf{X_i} - \sum_{j=1}^{k} W_{ij} \mathbf{X_{ij}} \right|^2 \tag{4.2}$$

Under the conditions: $\sum_{j=1}^{k} W_{ij} = 1$ and $W_{ij} = 0$ if $\mathbf{x_j}$ is not the neighbor of $\mathbf{x_i}$.

Step 3: Find d-dimension embedding vector using weights, which minimizes the cost function given as follows:

$$\Phi(Y) = \sum_{i=1}^{N} \left| \mathbf{Y_i} - \sum_{j=1}^{k} W_{ij} \mathbf{Y_j} \right|^2 \tag{4.3}$$

By creating sparse matrix M given by (4.4), compute the bottom *q+1* eigen vectors of M.

$$M = (1 + W)' * (1 + W) \tag{4.4}$$

The bottom *q+1* eigen vectors are reduced dimensions of input, which contain almost all the important information.

LLE attempts to preserve solely local properties of the data. As a result, LLE is less sensitive to short-circuiting than Isomap, because only a small number of local properties are affected if short-circuiting occurs. Furthermore, the preservation of local properties allows for successful embedding of non-convex manifolds. However, LLE is reported to fail in the visualization of even simple synthetic biomedical datasets. LLE performs worse than Isomap in the derivation of perceptual motor

actions. A possible explanation lies in the difficulties that LLE has when confronted with manifolds that contain holes. In addition, LLE tends to collapse large portions of the data very close together in the low-dimensional space, because the covariance constraint on the solution is too simple [25]. Also, the covariance constraint may give rise to undesired rescaling of the data manifold in the embedding.

4.2.2 Laplacian Eigenmaps

Similar to LLE, Laplacian Eigenmaps (LE) finds a low-dimensional data representation by preserving local properties of the manifold [26].

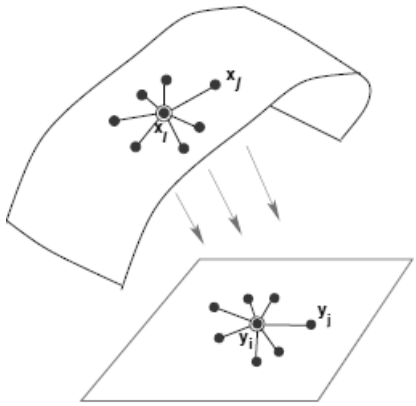


Figure 4.2 Pictorial representation of LE

LE computes a low-dimensional representation of the data, in which the distances between a data point and its k nearest neighbors are minimized as shown in Figure 4.2. This is done in a weighted manner, i.e., the distance in the low-dimensional data representation between a data point and its first nearest neighbor contributes more to the cost function than the distance between the data point and its second nearest neighbor. Using spectral graph theory, the minimization of the cost function is defined as an eigen problem.

4.2.2.1 Algorithm: LE

The LE algorithm can be described as follows:

Step 1: It constructs a neighborhood graph, in which every data point x_i is connected to its k nearest neighbors.

Step 2: For all points x_i and x_j in graph that are connected by an edge, the weight of the edge is computed using the Gaussian kernel function defined by (4.5), which leads to a sparse adjacency matrix **W**.

$$w_{ij} = e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}} \tag{4.5}$$

Step 3: In the computation of the low-dimensional representations, the cost function that is minimized is given asfollows:

$$\Phi(\mathbf{Y}) = \sum_{ij} \left\| \mathbf{y_i} - \mathbf{y_j} \right\|^2 w_{ij} \tag{4.6}$$

In the cost function, large weights w_{ij} correspond to small distances between the high-dimensional data points x_i and x_j. Hence, the difference between their low-dimensional representations y_i and y_j highly contributes to the cost function. As a consequence, nearby points in the high-dimensional space are put as close together as possible in the low-dimensional representation.

The computation of the degree matrix **M** and the graph Laplacian **L** of the graph **W** allows for formulating the minimization problem in (4.6) as an eigen problem [117]. The degree matrix **M** of **W** is a diagonal matrix, of which the entries are the row sums of **W (**i.e., $m_{ij} = \sum_j w_{ij}$). The graph Laplacian **L** is computed by **L = M−W**. It can be shown that the following holds.

$$\Phi(\mathbf{Y}) = \sum_{ij} \left\| \mathbf{y_i} - \mathbf{y_j} \right\|^2 w_{ij} = 2\mathbf{Y^T L Y} \tag{4.7}$$

Hence, minimizing $\Phi(\mathbf{Y})$ is proportional to minimizing $\mathbf{Y^T L Y}$ subject to $\mathbf{Y^T M Y = I}_n$, a covariance constraint that is similar to that of LLE. The low-dimensional data representation can thus be found by solving the generalized eigen value problem for the d smallest nonzero eigen values using (4.8).

$$\mathbf{Lv} = \lambda\, \mathbf{Mv} \tag{4.8}$$

The d eigenvectors $\mathbf{v}_i$ corresponding to the smallest nonzero eigen values form the low dimensional data representation **Y**.

Laplacian eigenmaps suffers from many of the same weaknesses as LLE, such as the presence of a trivial solution that is prevented from being selected by a covariance constraint. Despite these weaknesses, Laplacian eigenmaps has been successfully applied to face recognition and the analysis of functional magnetic resonance imaging (fMRI) data. In addition, variants of Laplacian eigenmaps may be applied to supervised or semi-supervised learning problems.

4.2.3 Maximum Variance Unfolding

Maximum Variance Unfolding (MVU), formerly known as Semidefinite Embedding learns the kernel matrix by defining a neighborhood graph on the data and retaining pairwise distances in the resulting graph [21]. MVU explicitly attempts to unfold the data manifold by maximizing the Euclidean distances between the data points, under the constraint that the distances in the neighborhood graph are left unchanged as shown in Figure 4.3. The resulting optimization problem can be solved using semidefinite programming (SDP).

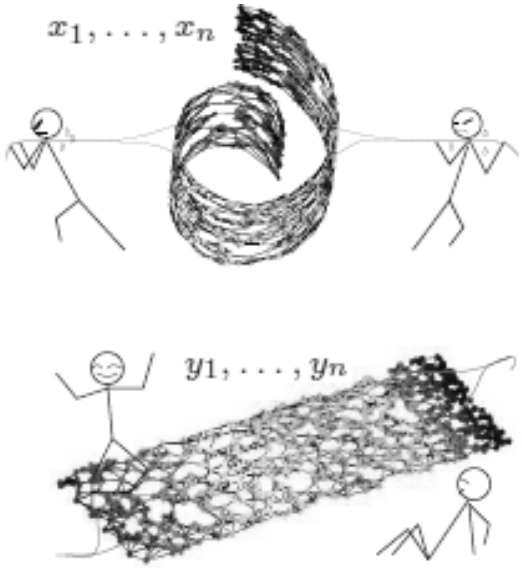


Figure 4.3 Pictorial representation of MVU

4.2.3.1 Algorithm: MVU

Step 1: MVU starts with the construction of a neighborhood graph G, in which each data point x_i is connected to its k nearest neighbors x_{ij} *(j = 1, 2, . . . , k).*

Step 2: Subsequently, MVU attempts to maximize the sum of the squared Euclidean distances between all data points, under the constraint that the distances inside the neighborhood graph G are preserved. In other words, MVU performs the following optimization problem.

Maximize $\sum_{i,j} \left\| \boldsymbol{y_i} - \boldsymbol{y_j} \right\|^2$ subject to (4.9),

$$\sum_{i,j} \left\| \boldsymbol{y_i} - \boldsymbol{y_j} \right\|^2 = \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2 \quad \text{for } \forall (i,j) \in G \tag{4.9}$$

Step 3: MVU reformulates the optimization problem as a SDP [41] by defining the kernel matrix **K** as the outer product of the low-dimensional data representation **Y**. The optimization problem then reduces to the following SDP, which learns the kernel matrix **K**.

Maximize trace (**K**) subject to (4.10), (4.11), and (4.12),

$$k_{ij} + k_{jj} - 2k_{ij} = \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2 \quad \text{for } \forall (i,j) \in G \tag{4.10}$$

$$\sum_{i,j} k_{i,j} = 0 \tag{4.11}$$

$$k \succ 0 \tag{4.12}$$

The low-dimensional data representation $\boldsymbol{Y}$ is obtained by performing an eigen decomposition of the kernel matrix $\boldsymbol{K}$ that is constructed by solving the SDP. MVU has a weakness that short circuiting may impair the performance of MVU, because it adds constraints to the optimization problem that prevent successful unfolding of the manifold. Despite this weakness, MVU is successfully applied to sensor localization and DNA microarray data analysis.

4.2.4 Isomap

Classical scaling has proven to be successful in many applications, but it suffers from the fact that it mainly aims to retain pairwise Euclidean distances, and does not take into account the distribution of the neighboring data points.

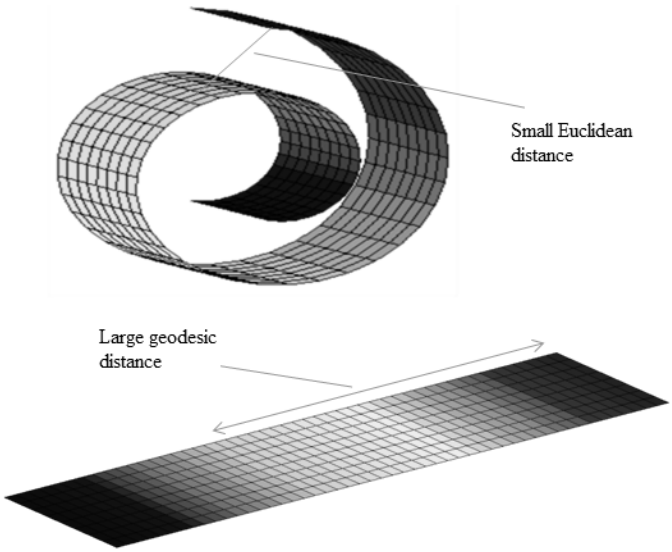


Figure 4.4 Illustrations of Geodesic and Euclidean Distance

If the high-dimensional data lies on or near a curved manifold, such as in the swiss roll dataset, classical scaling might consider two data points as near points, whereas their distance over the manifold is much larger than the typical inter point distance as shown in Figure 4.4.

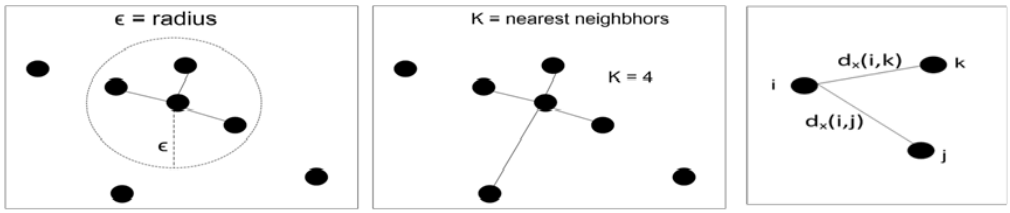
Isomap [14] is a technique that resolves this problem by attempting to preserve pairwise Geodesic (or curvilinear) distances between data points. The Geodesic between two points is defined as the shortest curve on the manifold connecting the two points. These techniques are efficient at visualizing data sets and are powerful to handle non-linear data. Overall, it is observed that Isomap find coordinates on lower dimensional manifold that best preserve Geodesic distances instead of Euclidean distances. In Isomap, the Geodesic distances between the data points are computed by constructing a neighborhood graph, in which every data point is connected with its k nearest neighbors in the dataset. The shortest path between two points in the graph forms an estimate of the Geodesic distance between these two points and can easily be

computed using Floyd's or Dijkstra's shortest-path algorithm [74, 88]. The Geodesic distances between all data points are computed, thereby forming a pairwise Geodesic distance matrix. The low-dimensional representations are computed by applying classical scaling on the resulting pairwise Geodesic distance matrix.

4.2.4.1 Algorithm: Isomap

Isomap algorithm can be described in three steps as follows:

Step 1: Neighbors of each point are determined.



(a) **(b)** **(c)**

Figure 4.5(a) Neighbors with ε radius approach (b) Neighbors with k-nearest neighborhood approach (c) edges of weight d_X (i.j) between neighboring points

The neighbors are chosen as points, which are within the ε distance or using k-nearest neighbor approach as shown in Figure 4.5 (a) and (b) respectively. These neighborhood relations are represented as a weighted graph over the data points as shown in Figure 4.5(c).

For selecting the neighborhood points, two techniques such as ε-radius and k-neighborhood are used. In ε -radius method, radius is fixed and all the data points, which are coming under this radius are considered as a neighborhood points. While in k-neighborhood method, number for k is fixed like 4-5 and k-number of data points which are nearest to the selected data points is considered as a neighborhood points for the selected data point.

After selecting the neighborhood point, next task is to assign weight to the edges, which are connecting the neighborhood data points. The edge weight in this case is equal to the Euclidean distance between the data points, which is illustrated in Figure 4.5 (c).

Step 2: Isomap estimates the Geodesic distances between all pairs of points on the manifold by computing their shortest path distance in the graph.

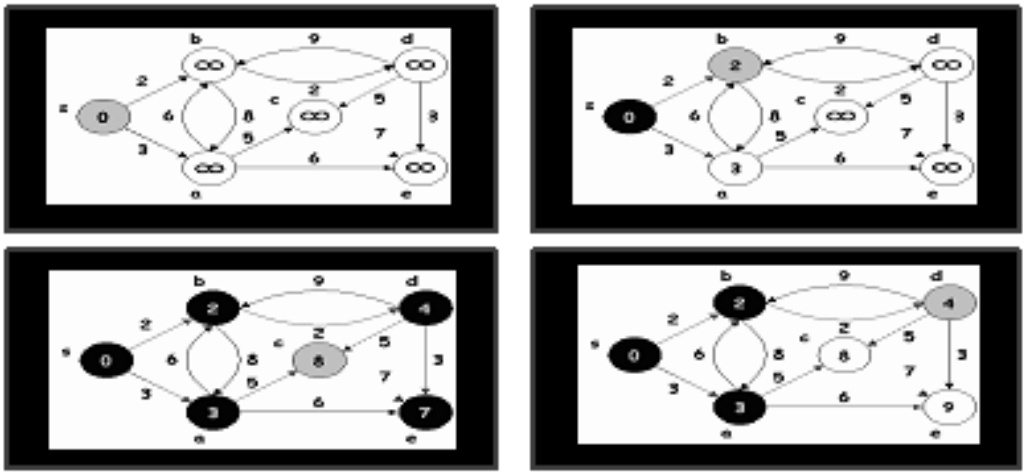


Figure 4.6 Dijkstra algorithm

The shortest path can be found by using Dijkstra algorithm as shown in Figure 4.6. If *i* and *j* are nearest neighbors then relations are represented as (4.13). If pair of points is not nearest neighbors the relations are represented as (4.14).

$$d_G(i,j) = d_x(i,j), \quad \text{Neighboring } i, j \tag{4.13}$$

$$d_G(i,j) = \infty \qquad \text{Otherwise} \tag{4.14}$$

Using shortest path algorithm, main task is to assign some finite values to all the edges of the graph. It can be done as follows:

$$for \quad k = 1, 2, ..., N$$

$$d_G(i,j) = \min\left\{d_x(i,j), d_x(i,k) + d_x(k,j)\right\} \tag{4.15}$$

Step 3: The final step applies classical MDS to the matrix of graph distances constructing an embedding of the data in a d-dimensional Euclidean space that best preserves the manifold's estimated intrinsic geometry. The coordinate vectors are chosen to minimize the cost function as given below:

$$E = \left\|\tau(\mathbf{D_G}) - \tau(\mathbf{D_Y})\right\|_L 2 \tag{4.16}$$

Where $D_Y(i,j) = \left\| y_i - y_j \right\|$

$D_G(i,j) = d_G(i,j)$ and

$\tau(D) = \frac{-1}{2}(I - \frac{1}{N})D^{.2}(I - \frac{1}{N})$

where D_Y, denotes the matrix of Euclidean distance in the embedded space and $\tau(D_Y)$, is the corresponding Euclidean inner product matrix. $\tau(D_G)$ is the shortest path inner product matrix. In a least square sense, Isomap expects $\mathbf{Y}^{\mathbf{T}}\mathbf{Y}$ to be close to $\tau(D_G)$. The τ operator converts distances to inner products, which uniquely characterize the geometry of the data in a form that supports efficient optimization.
The global minimum of (4.16) is achieved by setting the coordinates y_i to the top d eigenvectors of the matrix $\tau(\mathbf{D_G})$. The true dimensionality of the data can be estimated from the decrease in error as the dimensionality of Y is increased.

An important weakness of the Isomap algorithm is its topological instability [15]. Isomap may construct erroneous connections in the neighborhood graph. Such short circuiting [16] can severely impair the performance of Isomap. Several approaches have been proposed to overcome the problem of short circuiting by removing datapoints with large total flows in the shortest path algorithm [74, 88] or by removing nearest neighbors that violate local linearity of the neighborhood graph [75]. A second weakness is that Isomap may suffer from hole in the manifold. This problem can be dealt with by tearing manifolds with holes [16]. A third weakness of Isomap is that it can fail if the manifold is non-convex [14]. Despite these three weaknesses, Isomap is successfully applied on tasks such as wood inspection, visualization of biomedical data, and head pose estimation.

Graph based techniques are efficient at visualizing artificial data sets and powerful to handle non-linear data. However, these are unsupervised method, so cannot make use of any supervised prior information for discrimination. Also, they fails to identity inter or intraclass types of neighborhoods and unable to handle multiple class real problems. To address these issues constraint Isomap is proposed that is described in the next section.

4.3 Extension of Graph Based Techniques

In this section, constrained Isomap is proposed for visualization and non-linear dimensionality reduction. It considers both discriminant information and geometrical information of data. Here, pairwise constraints (PCs) are proposed that can provide more supervision information compared with the class labels.

4.3.1 Constraint Isomap

The constraint Isomap for manifold learning is proposed that is based on constraint margin maximization [CMM] criteria [38]. The pairwise CL and ML constraints are used to specify the types of neighborhoods. ML constraint helps in increasing the compactness of neighboring pairs while high separation between inter classes is achieved through CL constraint. In this way, interclass dissimilarity and intraclass compactness is introduced. Constraint Isomap computes the shortest path distances over constrained neighborhood graphs and guides the non-linear dimensionality reduction through separating the interclass neighbors. As a result, large margins between both inter and intraclass clusters are delivered and enhanced compactness of intra cluster points is achieved at the same time. Since this technique integrates the pairwise constraints and exhibits large margins between different clusters, it is referred as constrained Isomap.

Let $X = [x_1, x_2, ..., x_N]$ is the data set and $l(x_i) \in \{1, 2, 3, ..., C\}$, where $i = 1, 2, 3, ..., N$ are the class labels.

With the help of k nearest neighbor search, neighbor of each point are determined based on Euclidean distance between the points.

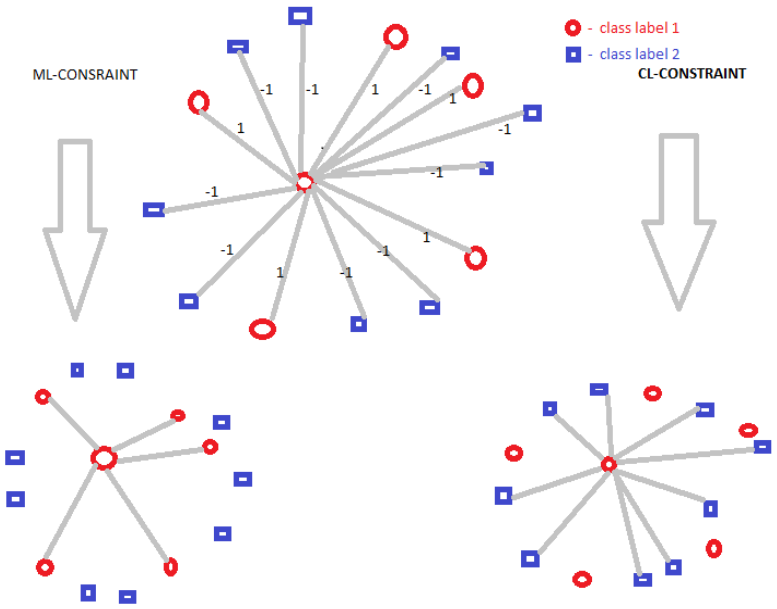


Figure 4.7 ML and CL constraints sets for constraint Isomap

Figure 4.7 shows the ML and CL constraints set for constraint Isomap that can be defined as follows:

$$s_{ML} = \left\{ (x_i, x_j) \middle| e(x_i, x_j) = 1, v_i \in V, v_j \in V, l(x_j) = l(x_i) \right\} \tag{4.17}$$

$$s_{CL} = \left\{ (x_i, x_j) \middle| e(x_i, x_j) = -1, v_i \in V, v_j \in V, l(x_j) \neq l(x_i) \right\} \tag{4.18}$$

The weights $\in \{0,1,-1\}$ to the edge $e(x_i, x_j) \in E$ linking x_i and x_j are defined as follows:

$e(x_i, x_j) = 1$, if i and j are neighbors and belong to same class

$e(x_i, x_j) = -1$, if i and j are neighbors and belong to different class

$e(x_i, x_j) = 0$, if i and j are not neighbors

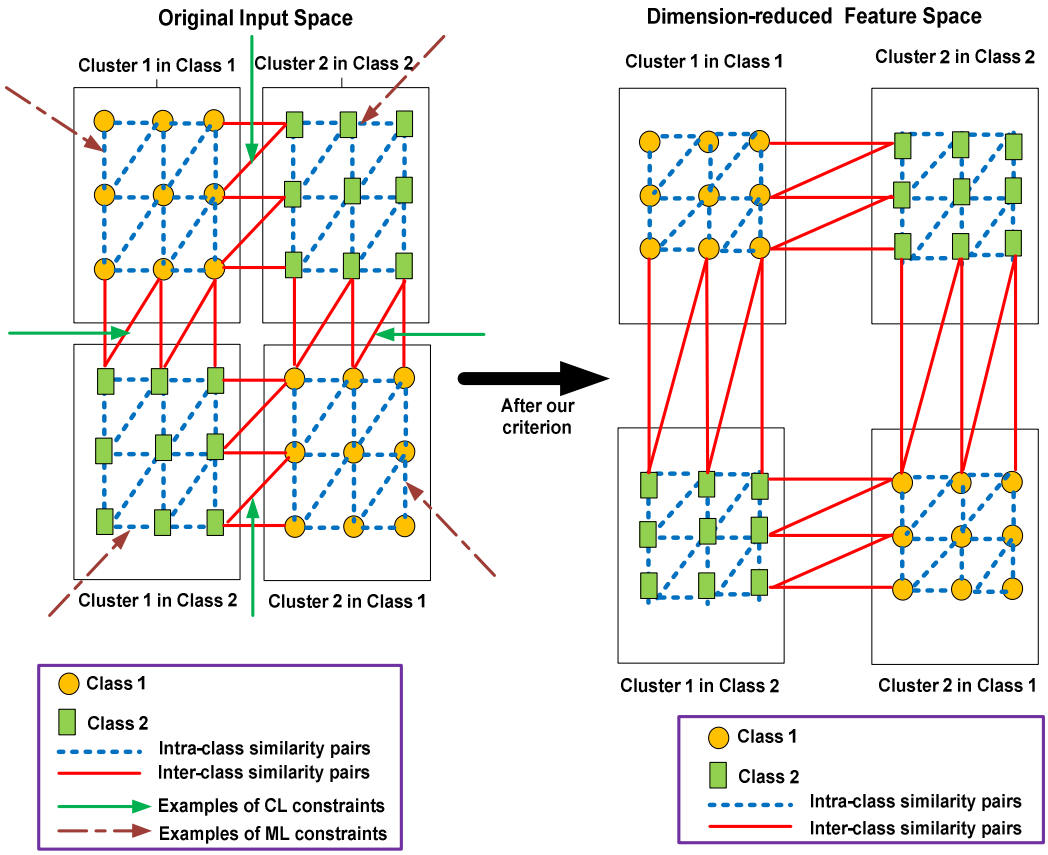


Figure 4.8 Graphical representations of proposed Constraint Isomap criteria

In this technique, graph is divided into two separate graphs according to the weight of the edge. ML-constraint graph is constructed by removing edges with negative weights and CL-constraint graph by removing edges with positive weights. From Figure 4.7, it can be seen that when negative edges are removed from the graph G, ML constrained neighborhood graph G_{ML} is constructed. Similarly when positive edges are removed from the graph G, CL constrained neighborhood graph G_{CL} is constructed. Figure 4.8 clearly shows that constraint Isomap creates margin between two different clusters. For efficient dimensionality reduction and feature extraction, it is desired that the compactness of neighboring pairs constrained by ML can be enhance, while high separation between neighboring pairs constrained by CL can be achieved. Cluster is defined as a set of data points consisting of similar data, which belong to same class. Figure 4.8 shows two clusters of two different class i.e. class1

and class 2. Data of same class are compacted and data of different class are separated by applying concept of constraint Isomap.

4.3.3.1 **Algorithm:** Constraint Isomap

Constraint Isomap algorithm can be described in three steps as follows:

Step 1: Neighbors of each point are determined. The neighbors are chosen as points, which are within the ε distance or using k-nearest neighbor approach as shown in Figure 4.5(a) and (b) respectively. Accordingly the ML and CL constraint set can be computed and constraint neighborhood graphs G_{ML} and G_{CL} are constructed,

The weights $d_X^{ML}(x_i,x_j)=\|x_i-x_j\|$ and $d_X^{CL}(x_i,x_j)=\|x_i-x_j\|$ are set on the edges of the graph G_{ML} and G_{CL} where $d_X^{ML}(x_i,x_j)$ is the Euclidean distance between two points in Must-Link constraint graph and $d_X^{CL}(x_i,x_j)$ is the Euclidean distance between two neighboring points in cannot-Link constraint graph.

Step 2: In step 2 like Isomap, the Geodesic distances $d_M^{ML}(x_i,x_j)$ and $d_M^{CL}(x_i,x_j)$ between all pairs of constraint points on the manifold are estimated or two independent graphs G_{ML} and G_{CL} respectively. For ML-graph, goal is to increase the compactness between the points. It can be done by normalizing edge weight.

$$d_X^{ML}(x_i,x_j)\ to\ \tilde{d}_X^{ML}(x_i,x_j)=d_X^{ML}(x_i,x_j)\Big/\max(d_X^{(ML)}) \tag{4.19}$$

Where $d_X^{(ML)}$ is the largest value of edge weight in $d_X^{ML}(x_i,x_j)$.

By doing so compactness between the points are increased for ML-constraint graph.

Step 3: Constraint Isomap uses the trace ratio optimization [39] to the matrices $D_G^{ML}=\{\tilde{d}_G^{ML}(x_i,x_j)\}\ and\ D_G^{CL}=\{d_G^{CL}(x_i,x_j)\}$ for computing the embedding of the original samples in a reduced d-dimensional Euclidean space. Isomap seeks the matrix $\tau(D_Y)$ over all points to be as close to $\tau(D_G)$ as possible. As a result, for some complex distributed real data sets, the interclass neighbors are likely to be congregated in the reduced output space. Unlike Isomap, to deliver large margins for

interclass discrimination, constraint Isomap handles the ML and CL constrained points independently. For the ML constraint set, constraint Isomap optimizes the following criterion.

$$J_{ML}(Y) = \min_{Y} \sum_{(x_i, x_j) \in ML} \left\| d_G^{ML}(x_i, x_j) - d_Y^{ML}(y_i, y_j) \right\|^2 \tag{4.20}$$

For CL constraint graph, the goal is to maximize the distance between the two points. It can be explained mathematically as follows:

$$J_{CL}(Y) = \max_{Y} \sum_{(x_i, x_j) \in CL} \left\| d_G^{CL}(x_i, x_j) - d_Y^{CL}(y_i, y_j) \right\|^2 \tag{4.21}$$

From the concept of MDS, (4.20) and (4.21) can be rewritten as follows:

$$\min_{Y} \sum_{(x_i, x_j) \in ML} \left\| d_G^{ML}(x_i, x_j) - d_Y^{ML}(y_i, y_j) \right\|^2 = \max Y \, \Gamma(D_G^{ML}) \, Y^T \tag{4.22}$$

Similarly,

$$\max_{Y} \sum_{(x_i, x_j) \in CL} \left\| d_G^{CL}(x_i, x_j) - d_Y^{CL}(y_i, y_j) \right\|^2 = \min Y \, \Gamma(D_G^{CL}) \, Y^T \tag{4.23}$$

To implement constraint Isomap, both (4.22) and (4.23) have been implemented using the concept of Trace Ratio Optimization given as follows:

$$Y^* = \max_{YY^T = I} \frac{tr \, Y \tau(D_G^{ML}) Y^T}{tr \, Y \tau(D_G^{CL}) Y^T} \tag{4.24}$$

Subject to constraint $YY^T = \mathrm{I}$.

Trace Ratio Optimization tries to maximize the ratio of two traces. In this ratio, numerator represent in between scatter, which measures how well the classes are separated in the projected space. The denominator part represents the within scatter, which measure how well the intraclass points are closed. Thus, the entire concept of constraint Isomap works on reducing the distance between intraclass point and at the same time increasing the distance between interclass points.

4.4 Results

All the graph based techniques are implemented on artificial generated datasets. The artificial datasets on which the algorithms are implemented are swiss roll, helix and twin peaks dataset. The datasets are specifically selected to investigate how the dimensionality reduction techniques deal with data that lies on a low dimensional manifold that is isometric to Euclidean space and data lying on a low dimensional manifold that is not isometric to Euclidean space.

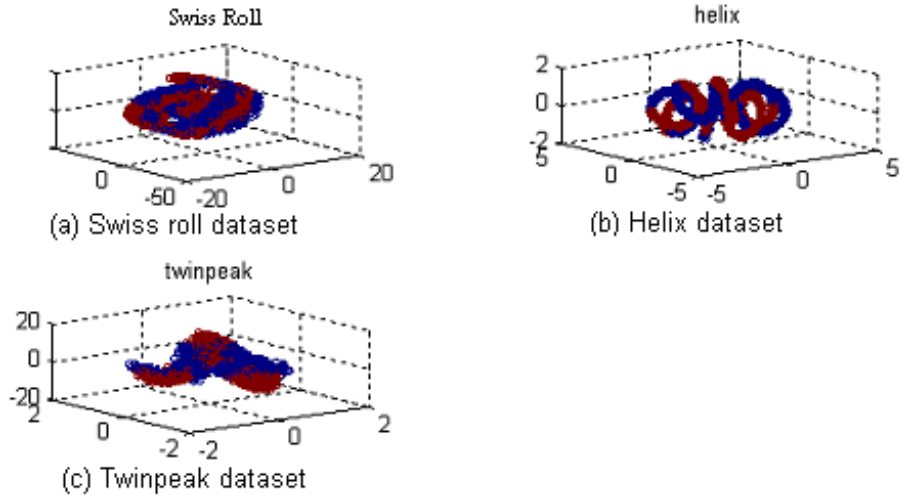


Figure 4.9 Artificially generated datasets (a) Swiss Roll (b) Helix (c) Twinpeak

Table 4.1: Parameter values for the experiments

Technique	**Parameter settings**
Isomap	$5 \leq k \geq 15$
MVU	$5 \leq k \geq 15$
LLE	$5 \leq k \geq 15$
Laplacian eigenmaps	$5 \leq k \geq 15$ $\sigma = 1$

Figure 4.9 shows artificial generated datasets. All artificial datasets consist of 5,000 samples. The experiments are run for all parameter settings listed in Table 4.1. The parameter k is the nearest neighbors of data point and σ is the bandwidth in LE.

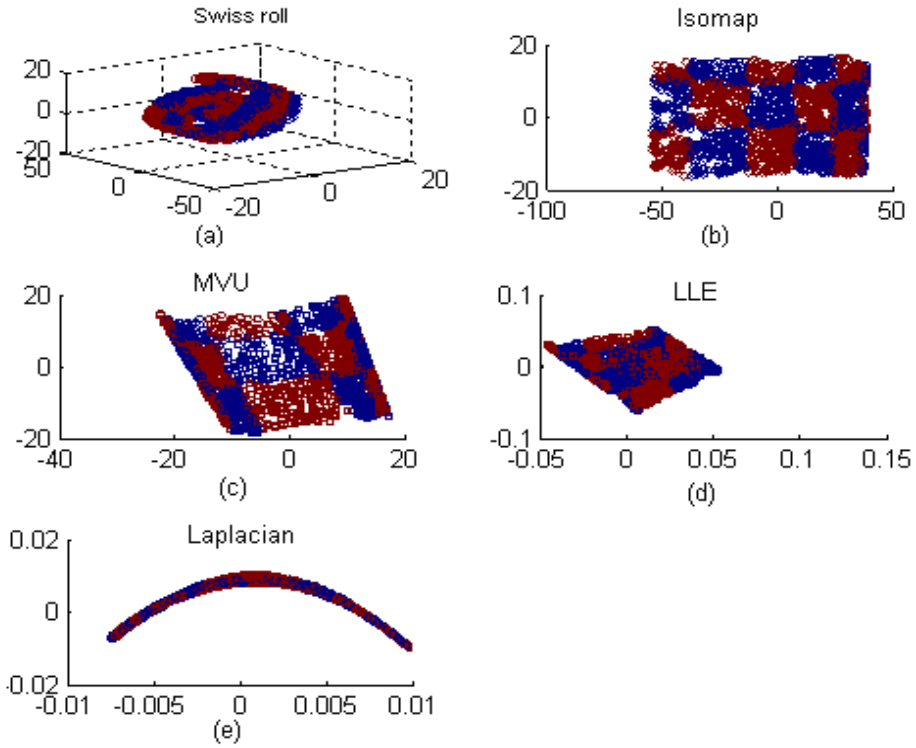


Figure 4.10 Results of dimensionality reduction techniques on Swiss roll dataset (a) Swiss roll (b) Isomap (c) MVU (d) LLE (e) LE

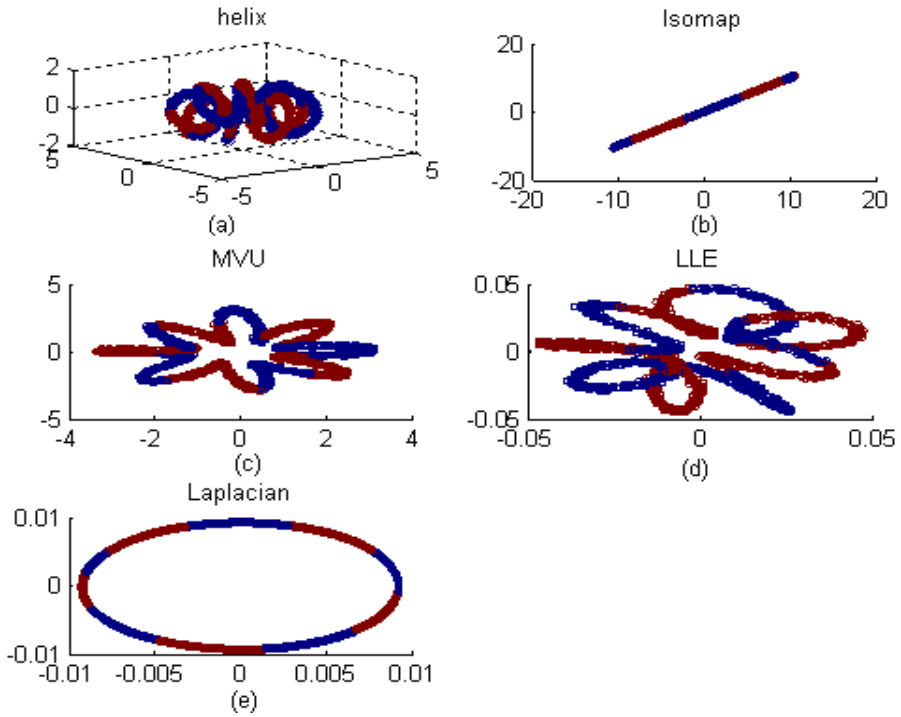


Figure 4.11 Results of dimensionality reduction techniques on helix dataset (a) helix (b) Isomap (c) MVU (d) LLE (e) LE

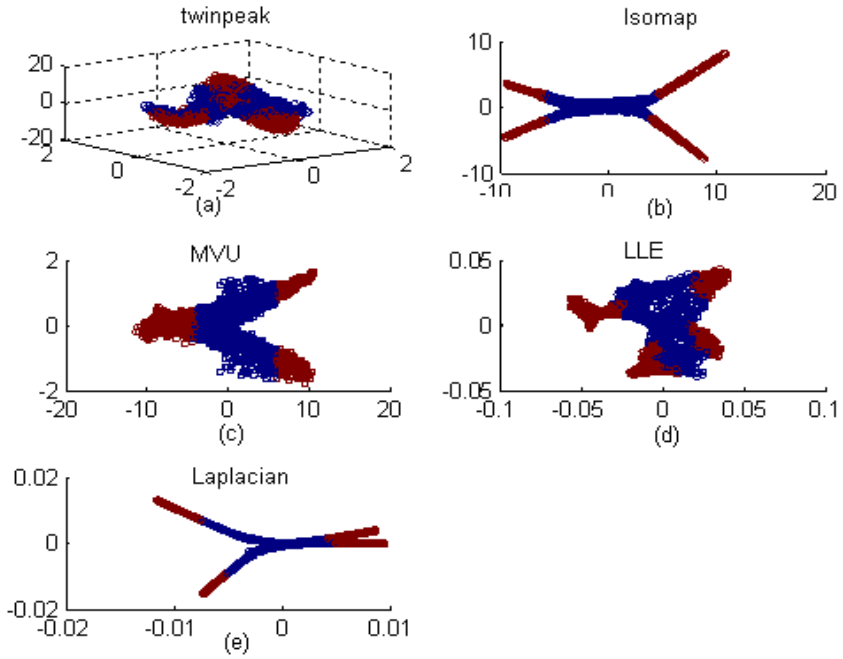


Figure 4.12 Results of dimensionality reduction techniques on twinpeak dataset (a) twinpeak (b) Isomap (c) MVU (d) LLE (e) LE

Figures 4.10, 4.11 and 4.12 show the results of graph based non-linear dimensionality reduction techniques on swiss roll, helix and twinpeak datasets. The results reveal the strong performance of dimensionality reduction techniques based on neighborhood graphs (Isomap, MVU, LLE, and LE). The performance of LLE on the helix dataset is notably worse than its performance on the swiss roll dataset. The other techniques based on neighborhood graphs (Isomap, MVU, and LE) perform strong on the helix dataset, despite the non-isometric nature of the dataset.

Table 4.2 Computational and memory complexity

Technique	Parameters	Computational	Memory
Isomap	k	$O(n^3)$	$O(n^2)$
MVU	k	$O((nk)^3)$	$O((nk)^3)$
LLE	σ, k	$O(pn^2)$	$O(pn^2)$
Laplacian eigenmaps	k	$O(pn^2)$	$O(pn^2)$

In Table 4.2, the four dimensionality reduction techniques are listed by three general properties: (1) the main free parameters that have to be optimized, (2) the

computational complexity of the main computational part of the technique, and (3) the memory complexity of the technique.

For property 1, Table 4.2 shows that the objective functions of most of the non-linear dimensionality reduction techniques have free parameters that need to be optimized. Moreover, LLE uses a regularization parameter in the computation of the reconstruction weights. The presence of free parameters has both advantages and disadvantages. The main advantage of the presence of free parameters is that they provide more flexibility to the technique, whereas their main disadvantage is that they need to be tuned to optimize the performance of the dimensionality reduction technique.

For properties 2 and 3, Table 4.4 provides insight into the computational and memory complexities of the techniques. The computational complexity of a dimensionality reduction technique is determined by (1) properties of the dataset such as the number of data points n and their dimensionality D, and (2) by parameters of the techniques, such as the target dimensionality d and the number of nearest neighbors k. Isomap performs an eigen analysis of an n × n matrix using a power method in $O(n^3)$. Because these full spectral techniques store a full n × n kernel matrix, the memory complexity of these techniques is $O(n^2)$.

In addition to this, MVU solves a SDP with nk constraints. Both the computational and the memory complexity of solving an SDP are cube in the number of constraints. Since there are nk constraints, the computational and memory complexity of the main part of MVU is $O((nk)^3)$.

Sparse spectral techniques (LLE and Laplacian eigenmaps) perform an eigen analysis of an n × n matrix. However, for these techniques the n × n matrix is sparse, which is beneficial, because it lowers the computational complexity of the eigen analysis. Eigen analysis of a sparse matrix has computational complexity $O(pn^2)$, where p is the ratio of nonzero elements in the sparse matrix to the total number of elements. The memory complexity is $O(pn^2)$ as well.

The above analysis provides some insight into the differences between Isomap, MVU, Laplacian eigenmaps, and LLE. The metrics induced by Isomap and MVU are related to Geodesic and local distances, respectively, on the submanifold, from which the input patterns are sampled. On the other hand, the metric induced by the graph

Laplacian is related to the commute times involve all the connecting paths between two nodes on a graph, not just the shortest one. The kernel matrix induced by LLE is roughly analogous to the square of the kernel matrix induced by the graph Laplacian.

The effectiveness of the constraint Isomap is tested on real databases from Olivetti and Oracle Research Laboratory (ORL), Brendan, Aleix Martinez and Robert Benavente (AR) databases [96-98]. The ORL database contains 400 face images of 40 persons. The images are captured at different time and have different variations including facial expressions (open/closed eyes, smiling/ not smiling), and facial details (glasses/ no glasses) against a dark homogenous background. The face images of 40 persons have been used that created 40 class problem. The AR database contains 116 face images, 26 images are available for each person. The AR database is captured with different expressions, illumination conditions and occlusions (scarf and sunglasses). In the AR database, besides the lighting from the left and the right, lighting from both sides of each face is also adopted. The Brendan's faces database is represented by 1965 face images taken from sequential frames of small video.

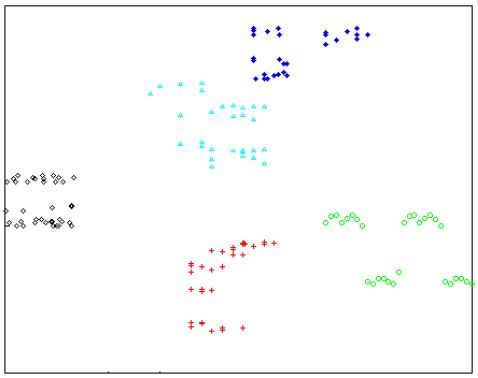
(a)

(b)

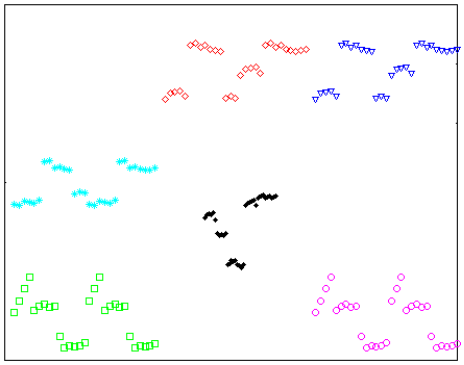
(c)

Figure 4.13 Sample facial expressions from (a) ORL database (b) AR database (c) Brendan database

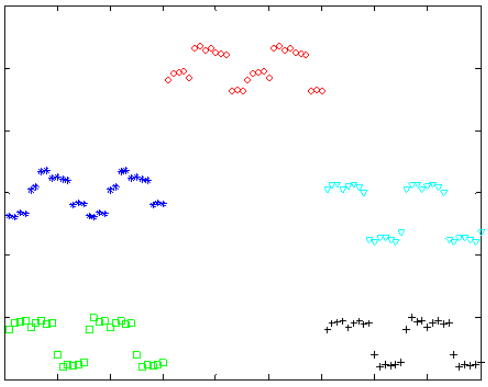
The images constitute various poses and expressions (neutral, smiling, laughter, sad, anger and surprise) of Brendan's faces. Sample face images from the ORL database, AR database and Brendan's database are displayed in Figure 4.13 (a), (b) and (c) respectively.



(a)



(b)



(c)

Figure 4.14 Result of 2-D embedding of (a) ORL face database (b) AR face database (c) Brendan face database

The 2-D embedding of ORL face images and AR face images are shown in Figure 4.14 (a) and (b) respectively, where each point corresponds to a face. From Figure 4.14 (a) and (b), it is observed that the face images of same person are grouped together and of different persons are separated from each other. At the same time, different face expressions of a person are separated from each other and similar ones grouped together. Similarly, the 2-D embedding of Brendan's face database is shown in Figure 4.14 (c), where each point corresponds to a face and it is observed that the different face expressions are separated from each other and similar ones grouped together.

Constraint-Isomap highlights the natural clusters of the faces and show separate clusters between dissimilar faces. They make similar face of the same individual lie in the vicinity of the face image space and make dissimilar faces from different individuals appear far away in their reduced embedding spaces. Compared with the other techniques, constraint-Isomap increases the margins between images of different persons and at the same time enhanced compactness between similar face images of a person.

The k means algorithm has been applied to the embedded data for computing clustering accuracy. Here, k is set to 20 for each method. The constraint Isomap is compared with LLE, Laplacian and Isomap algorithms.

Table 4.3 Clustering accuracy of ORL database

Techniques	**Clustering accuracy (%)**
LLE	62
LE	64
Isomap	58
Constraint Isomap	81

Table 4.4 Clustering accuracy of AR Database

Techniques	**Clustering accuracy (%)**
LLE	61
LE	64
Isomap	52
Constraint Isomap	79

Table 4.5 Clustering accuracy of Bredan's Database

Techniques	Clustering accuracy (%)
LLE	52
LE	53
Isomap	48
Constraint Isomap	75

The experimental results of clustering accuracy of ORL, AR and Brendan face database are tabulated in Tables 4.3-4.5 respectively. The comparative results clearly demonstrate that the constraint Isomap works well to separate faces of different persons and gather the faces of same person. Experimental results clearly demonstrate that for all the databases, constraint Isomap outperforms other DR techniques for clustering accuracy. LLE cannot separate faces of different persons visually, their clustering results, which are comparable, are worse in this data set. The results of the LLE, Laplacian eigenmaps and Isomap algorithms are comparative with each other in all the cases. The clustering performance is greatly improved with constraint-Isomap, compared with the other techniques. Isomap is efficient in visualizing synthetic dataset but usually delivers unsatisfactory results in real datasets while constraint Isomap is powerful for handling multiple-class real problems.

4.5 Conclusions

Each of the graph based techniques for non-linear dimensionality reduction has its own advantages and disadvantages. In several datasets, the correct features could only be extracted with the non-linear approaches. However, Isomap is more reliable and accurate than LLE. It is less sensitive to the chosen neighborhood size as well as to noisy or sparse data. Further, it is usually sufficed with smaller neighborhood sizes, which leads to faster calculations. Advantage of MVU is its flexibility to be adapted to particular applications as it determines the best kernel from the data. On the other side, MVU proves to be drastically slow due to the complexity of SDP step. Although Laplacian eigenmaps use a Gaussian kernel function to define local neighborhoods, the relationship between their uses of this neighborhood has yet to be explored.

Isomap is efficient in visualizing synthetic dataset but usually delivers

unsatisfactory results in real datasets while maximizing margin constraint Isomap is powerful for handling multiple-class real problems. A maximizing margin constraint Isomap enhanced both interclass separation and intraclass compaction. Clustering results obtained by constraint-Isomap are better than other graph based techniques. From the experimental results, it is observed that constraint Isomap is delivering clear separation on the manifold embedding of multiple classes.

Chapter-5

Extensions of Local Non-Linear Techniques

Chapter-5

Extensions of Local Non-Linear Techniques

In this chapter, Conformal Eigenmaps (CE) and Neighborhood Preserving Embedding (NPE) have been proposed as extensions of local non-linear techniques. Existing non-linear dimensionality reduction techniques, such as LLE and Laplacian eigenmaps are not explicitly designed to preserve local features such as angles. In proposed CE technique, a low dimensional embedding is constructed that maximally preserves angles between nearby data points. The embedding is derived from the bottom eigenvectors of LLE by solving an additional problem in semidefinite programming (SDP). In second proposed technique, NPE minimizes the cost function of a local non-linear technique for dimensionality reduction under the constraint that the mapping from the high-dimensional to the low-dimensional data representation is linear. The idea is to modify the LLE by introducing a linear transform matrix. The effectiveness of the proposed techniques is demonstrated on synthetic datasets. Experimental results on several data sets demonstrate the merits of proposed techniques.

5.1 Introduction

In the last decade, a large number of non-linear techniques for dimensionality reduction have been proposed [10-13]. In previous chapter, several manifold embedding based non-linear techniques such as LLE [22], Isomap [14] and LE [26] are discussed. They all utilized local neighborhood relation to learn the global structure of non-linear manifolds. But they have quite different motivations and objective functions. In contrast to the traditional linear techniques, the non-linear techniques have the ability to deal with complex non-linear data. On the other hand, such approaches also have several limitations such as the solutions do not yield an estimate of the underlying manifold's dimensionality. The geometric properties preserved by these embedding are difficult to characterize and the resulting

embeddings sometimes exhibit an unpredictable dependence on data sampling rates [99]. Moreover, the original LLE, Isomap and LE cannot deal with the out-of-sample problem directly [40]. Out-of-sample problem states that only the low dimensional embedding of training samples can be computed but the samples out of the training set cannot be calculated at all. Hessian LLE is a variant of LLE that learns isometries, or distance-preserving embeddings, with theoretical guarantees of asymptotic convergence [31], [41]. Like LLE, however, it does not yield an estimate of the underlying manifold's dimensionality.

In this work, an extended analysis has been provided to remedy the key deficiencies of LLE and LE. It is shown how to construct a more robust, angle-preserving embedding from the spectral decompositions of these algorithms as well as linear approximations to local non-linear techniques. The rest of this chapter is organized as follows: Section 5.2 described the proposed techniques, CE and NPE. Experimental results are shown in Section 5.3. Finally, conclusions are drawn in Section 5.4.

5.2 Extensions of Local Non-Linear Techniques

The capability of local non-linear techniques to successfully identify complex data manifolds has led to the proposal of its several extensions. The original LLE, Isomap and LE cannot deal with out of sample problem and cannot preserve local feature such as angle. To overcome the limitations of existing techniques, extensions of local non-linear techniques have been discussed in this section.

5.2.1 Conformal Eigenmaps

A conformal mapping is a transformation that preserves the angles between neighboring datapoints when reducing the dimensionality of the data. Conformal Eigenmaps (CE) are based on the observation that local non-linear techniques for dimensionality reduction do not employ information on the geometry of the data manifold that is contained in discarded eigenvectors correspond to relatively small eigen values. Conformal Eigenmaps initially perform LLE (or alternatively, another local non-linear technique for dimensionality reduction) to reduce the high-

dimensional data D to a dataset of dimensionality m. Conformal Eigenmaps use the resulting intermediate solution in order to construct a d-dimensional embedding (where d < m < D) that is maximally angle-preserving.

A conformal map is a low-dimensional embedding where the angles formed by three neighboring points in the original high dimensional dataset are equal to the angles between those same three points in the embedding. Consider the point x_i and its neighbors x_j and x_k in d-dimensional space. Also, consider z_i, z_j and z_k to be the images of those points in the final embedding. If the transformation is conformal map then the triangle formed by the *x* points would have to be similar to that formed by the z points. In the triangle formed by the *x* points the expression $|x_i - x_k|$ represents the length of one side of the triangle while the expression $|z_i - z_k|$ represents the corresponding side in the embedding. Since, the triangles are similar there must exist $\sqrt{s_i}$ such that:

$$\sqrt{s_i} = \frac{|x_j - x_k|}{|z_j - z_k|} = \frac{|x_i - x_k|}{|z_i - z_k|} = \frac{|x_i - x_j|}{|z_i - zj|} \tag{5.1}$$

$$s_i = \frac{|x_j - x_k|^2}{|z_j - z_k|^2} = \frac{|x_i - x_k|^2}{|z_i - z_k|^2} = \frac{|x_i - x_j|^2}{|z_i - zj|^2} \tag{5.2}$$

It is usually not possible to find a perfect embedding where all of the triangles are exactly similar to each other. Therefore, the goal is to find a set of *z* coordinates such that the triangles are as similar as possible. This leads to the following minimization.

$$\min_{z,s_i} \sum_{j,k} \left(|z_j - z_k|^2 - s_i |x_j - x_k|^2 \right)^2 \tag{5.3}$$

Where x_i represents the initial points and z_i denotes the points in the final embedding. Let y_i represent the points in the embedding produced by LLE (or LE). The y_i points represent an intermediate step in the algorithm and so go part of the way

to solving for z_i. Once LLE has produced, goal of the algorithm becomes a search for a transformation matrix L such that $z = Ly_i$ where the z_i value satisfy the minimization in (5.3).

$$\min_{L,s_i} \sum_{j,k} \left(\left| Ly_i - Ly_k \right|^2 - s_i \left| x_j - x_k \right|^2 \right)^2 \tag{5.4}$$

This should be done for all points x_i and with the condition that the points x_j and x_k are the neighbors of x_i.

$$\min_{L,s_i} \sum_{i} \sum_{j,k} \eta_{ij}\eta_{ik} \left(\left| Ly_i - Ly_k \right|^2 - s_i \left| x_j - x_k \right|^2 \right)^2 \tag{5.5}$$

Where η is an indicator variable. Its value is $\eta_{ij} = 1$, only if x_j is neighbor of x_i, otherwise $\eta_{ij} = 0$. The value for s_i can be calculated via least squares and the initial minimization (5.5) can be rewritten as follows:

$$\begin{aligned} &\text{Minimize} && t \\ & && P \succ 0, \\ &\text{Such that} && trace(P) = 1, \\ & && \begin{pmatrix} 1 & RVec(P) \\ RVec(P)^T & t \end{pmatrix} \succ 0, \end{aligned} \tag{5.6}$$

Where $P = L^T L$, t is an unknown scalar, *I* and *R* are m^2 *X* m^2 matrices, *I* denote the identity matrix, while *R* depends on $\{x_i, y_i\}_{i=1}^{n}$, but is independent of optimization variables *P* and *t*. The condition $trace(P) = 1$ is added to avoid the trivial solution where *P=0*. The optimization is an instance of SDP problem over elements of unknown matrix *P* [41]. After solving the SDP, the matrix can be decomposed back into $L^T L$ and the final embedding can be found by $z = Ly_i$ for all *i*. Conformal Eigenmaps introduced the interesting idea of using spectral methods like LLE and LE to find the low dimensional manifold and further modifying the output to produce a conformal map. Another extension of local non-linear dimension reduction technique is discussed in the section that follows:

5.2.2 Neighborhood Preserving Embedding

Neighborhood Preserving Embedding (NPE) is the linear approximation to local non-linear technique. In contrast to traditional linear techniques such as PCA, local non-linear techniques for dimensionality reduction are capable of successful identification of complex data manifolds such as swiss roll. This capability is due to the cost functions that are minimized by local non-linear dimensionality reduction techniques, which aim at preserving local properties of the data manifold. However, in many learning settings, the use of a linear technique for dimensionality reduction is desired, when an accurate and fast out-of-sample extension is necessary, when data has to be transformed back into its original space, or when one wants to visualize the transformation that is constructed by the linear dimensionality reduction technique. NPE is a technique that aims at combining the benefits of linear and local non-linear techniques for dimensionality reduction. It is done by finding a linear mapping that minimizes the cost function of LLE. NPE minimizes the cost function of a local non-linear technique for dimensionality reduction under the constraint that the mapping from the high dimensional to the low dimensional data representation is linear.

Similar to LLE, NPE starts with the construction of a nearest neighbor graph, in which every datapoint is connected to its nearest neighbors. The weights of the edges in the graph are computed and subsequently solves the generalized eigen value problem.

Given a set of points $X = \{X_1, X_2, ...X_N\}$, in $\mathbb{R}^D$,

NPE attempts to seek an optimal transformation matrix $\boldsymbol{P}$ to map high-dimensional data $\boldsymbol{X}$ onto a low-dimensional data Y, such that

$$\boldsymbol{Y = P^T X}$$

Where $Y = \{Y_1, Y_2, ...Y_N\}$, in $\mathbb{R}^d$ (d << D), in which the local neighborhood structure of X can be preserved.

NPE algorithm can be stated in three steps.

Step 1: Constructing an adjacency graph: Let G denote a graph with m nodes. The i^{th}

node corresponds to the data point x_i. There are two ways to construct the adjacency graph.

(i) K nearest neighbors (KNN): Put a directed edge from node *i* to *j* if x_j is among the K nearest neighbors of x_i.

(ii) ε neighborhood: Put an edge between nodes *i* and j if $\left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\| \leq \varepsilon$

The graph constructed by the first technique is a directed graph, while the one constructed by the second technique is an undirected graph. In many real world applications, it is difficult to choose a good ε. In this work, the KNN method is adopted to construct the adjacency graph. When computational complexity is a major concern, one may switch to ε neighborhood method.

Step 2: Computing the weights**:** In this step, the weights on the edges are computed. Let W denote the weight matrix with W_{ij} having the weight of the edge from node *i* to node *j*, and 0 if there is no such edge. The weights on the edges can be computed by minimizing the following objective function:

$$\min \sum_i \left\| \mathbf{X_i} - \sum_j W_{ij} \mathbf{X_{ij}} \right\|^2 \tag{5.7}$$

Where $\sum_j W_{ij} = 1, \quad j = 1, 2, ..., m$, and weight matrix can easily be obtained by minimizing the cost function (5.7).

Step 3: Computing the projections: In NPE, if data points **x** in space $\mathbb{R}^D$ can be reconstructed by **W**, then the corresponding point by y_i in space $\mathbb{R}^d$ can be reconstructed by **W** also. Therefore, the mapping transformation matrix $\boldsymbol{P}$ can be obtained by solving the following minimization problem:

$$\boldsymbol{P}_{opt} = \arg\min_P \left[\sum_i \left\| \mathbf{Y_i} - \sum_{j=1}^{k} W_{ij} \mathbf{Y_j} \right\|^2 \right] \tag{5.8}$$

$$\arg\min_A tr\left(\boldsymbol{P}^T \boldsymbol{X} \boldsymbol{M} \boldsymbol{X}^T \boldsymbol{P} \right)$$

Where $\boldsymbol{P}^T \boldsymbol{XMX}^T \boldsymbol{P} = 1, M = (1-W)^T (1-W)$

Where I represent the n x n identity matrix

By simple algebraic operations, the minimization problem of (5.8) becomes a generalized eigen value problem.

$$\boldsymbol{XMX}^T \boldsymbol{P} = \lambda \boldsymbol{XX}^T \boldsymbol{P} \tag{5.9}$$

$$\boldsymbol{Y} = \boldsymbol{P}^{\boldsymbol{T}} \boldsymbol{X} \tag{5.10}$$

where $\boldsymbol{Y}$ is a low-dimensional embedding that combines the benefits of linear and local non-linear techniques.

5.3 Results

In this section, results of proposed algorithms are presented for synthetic datasets. The datasets are specifically selected to investigate how the dimensionality reduction techniques deal with data that lies on a low-dimensional manifold.

The synthetic datasets, on which the algorithms are implemented are the swiss roll, helix, and twin peaks datasets. Figure 5.1 shows plots of these artificially generated datasets. All artificial datasets consist of 5,000 samples. The experiments are run for parameter k (nearest neighbors of data point) ranges from 5 to 15.

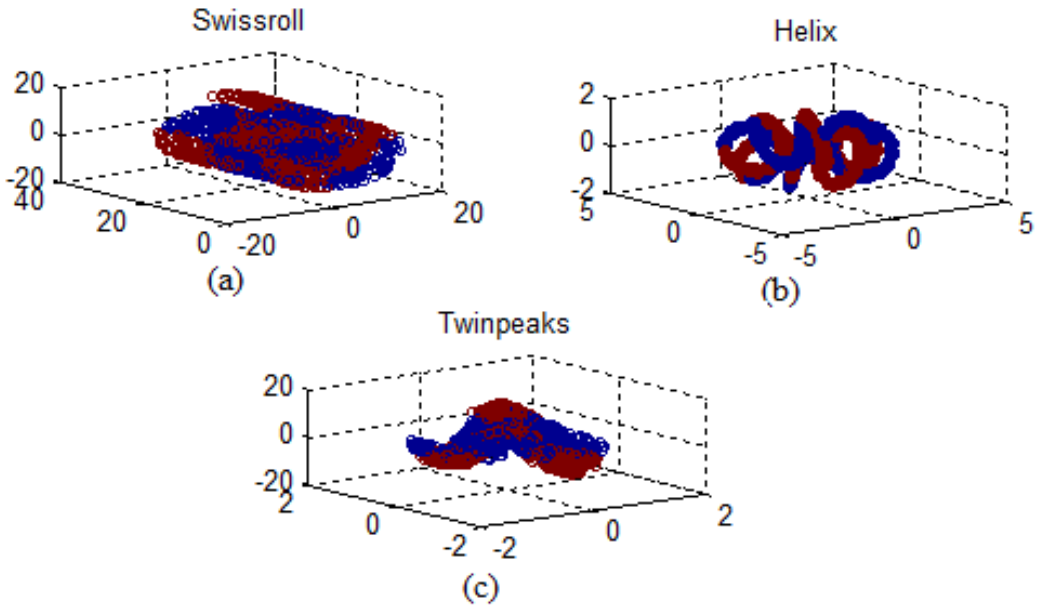


Figure 5.1 Artificially generated datasets (a) Swiss Roll (b) Helix (c) Twinpeak

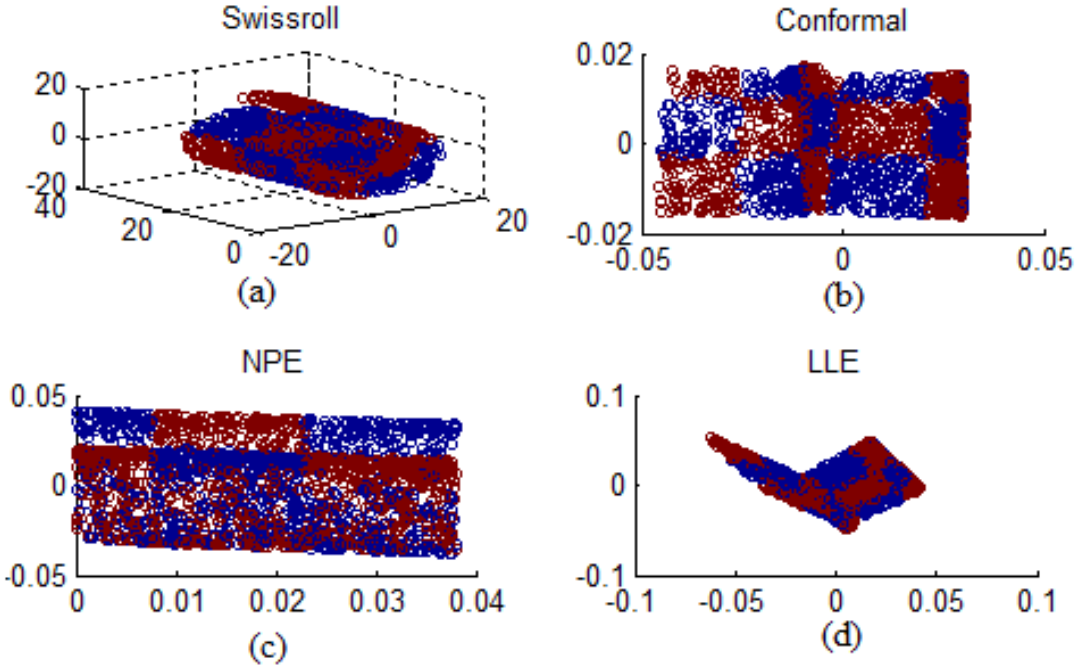


Figure 5.2 Results of dimensionality reduction on swiss roll dataset (a) Swiss roll (b) Conformal (c) NPE (d) LLE

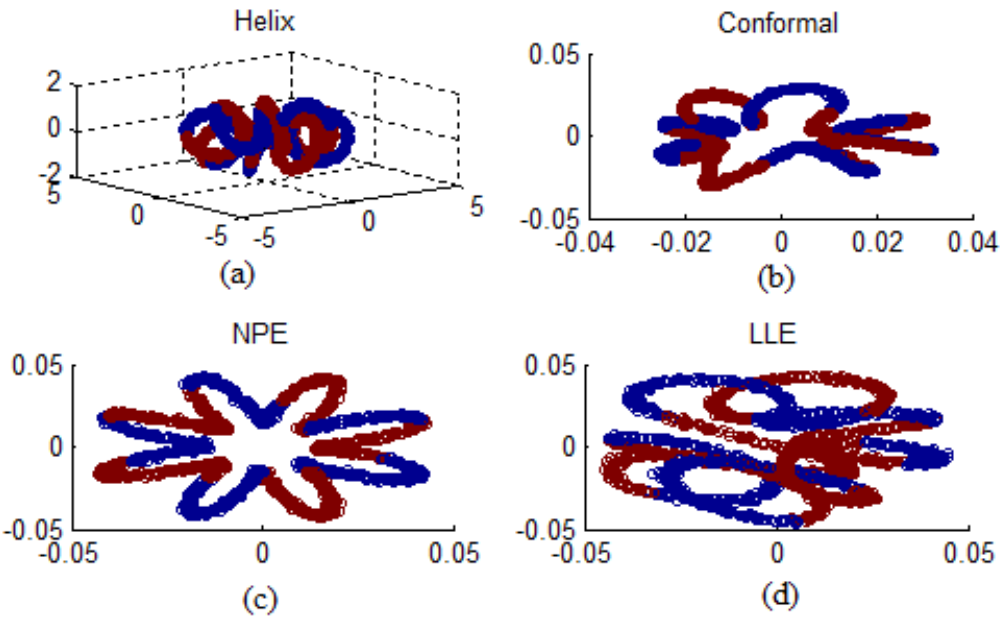


Figure 5.3 Results of dimensionality reduction on Helix dataset (a) Helix (b) Conformal (c) NPE (d) LLE

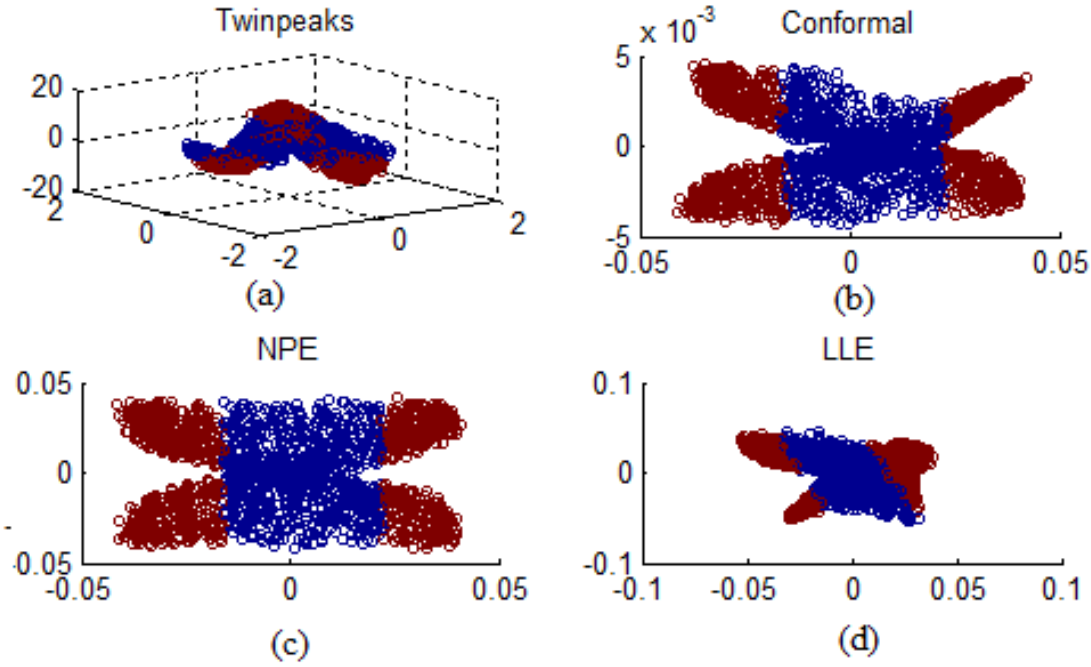


Figure 5.4: Results of dimensionality reduction on Twinpeaks dataset (a) Twinpeaks (b) Conformal (c) NPE (d) LLE

Figures 5.2-5.4 show the results of Conformal Map and NPE dimensionality reduction techniques on swiss roll, helix, and twinpeak dataset respectively. From the results, it is clear that the angle preserving embedding more faithfully preserves the shape of the underlying manifold's boundary. The distance preserving embedding of SDE has variance in more dimensions than the angle preserving embedding, suggesting that the latter has exploited the extra flexibility of conformal versus isometric maps. The semidefinite program in (5.6) mixes all of the bottom eigenvectors from LLE to obtain the maximally angle-preserving embedding.

Conformal transformations cast a new light on older algorithm, such as LLE. Viewing these bottom eigenvectors as a partial basis for functions on the data set, it is shown how to compute a maximally angle-preserving embedding by solving an additional problem in semidefinite programming. At little extra computational cost, Conformal Eigenmaps significantly extends the utility of LLE, yielding more faithful embeddings as well as a global estimate of the data's intrinsic dimensionality.

The proposed NPE is able to search a direction projected, in which neighborhood relations are preserved along the curve of the manifold. From Figures 5.2(c), 5.3(c) and 5.4(c), it is observed that NPE cannot always unfold the manifold as LLE. Furthermore, many neighbors are collapsed into a single point in the low dimensional space. The reason is that, NPE is a linear transform instead of non-linear one like LLE. Nevertheless, the NPE has favorable properties against other linear transform methods such as PCA.

Table 5.1 Performance Comparison between LLE, Conformal Map and NPE

	LLE	**Conformal Map**	**NPE**
Speed	Fast	Slow	Very Fast
Handle Curvature	Maybe	Yes	No
Handle Noise	No	Yes	No
Preserve angles	No	Yes	No

The LLE, Conformal Map and NPE techniques are compared based on various parameters such as speed, noise, non-convexity, curvature, non-uniform sampling. The performance comparison between LLE, Conformal Map and NPE is shown in

Table 5.1. First, it is observed that for swissroll dataset, LLE is pretty slow and cannot handle this data. For twin peaks, LLE fold up the corners of a plane because it introduces curvature to plane. LLE distort mapping the most. For helix dataset, LLE cannot recover the circle and noise is added to the helix sampling.

5.4 Conclusions

CE and NPE are proposed as extension of LLE and LE. In Conformal map, the three angles formed by a triangle, consisting of three neighboring points in the high dimensional space, is preserved in the lower dimensional embedding. However, the effectiveness of conformal mapping is limited by the computational complexity of SDP solver. NPE is a linear approximation to LLE. Comparing to the recently proposed manifold learning algorithms such as Isomap and LLE, NPE is defined everywhere, rather than only on the training data points. NPE is less sensitive to outliers than PCA. In general, it is observed that proposed approaches complements and extends earlier approaches at very modest extra cost.

Chapter-6

Non-Linear Dimensionality Reduction using Fuzzy Lattices

Chapter 6

Non-Linear Dimensionality Reduction using Fuzzy Lattices

Previously discussed techniques are based on the concept that local is linear while this work is based on concept of local is non-linear. To detect non linearity, relation between the nearest neighborhoods elements of the image have been expressed in terms of Gaussian membership functions. All the elements of the image are connected with the nearest neighborhood elements with some membership degree of the Gaussian functions. It results in the formation of number of fuzzy lattices. The lattices have been expressed in the form of Schrödinger equation, to find the kinetic energy (KE) used, corresponding to change occurring in the facial activity of a person. Finally, the KE embedded in three dimensional spaces is used to distinguish non-linear changes during occurrence of various facial activities. Experimental results show that proposed technique is effective in recognition of facial expressions as it focuses on extracting the non-linear features corresponding to contours of maximum energy, which are appearing over various expressions.

6.1 Introduction

Algorithms such as Isomap [14], MDS [42], LLE [22], GPLVM [43], Laplacian [44] and Hessian [45] share a common characteristic. They first include a local neighborhood structure on the data and then use this local structure to globally map the manifold to a lower dimensional space. Gabor wavelets [46] are trying to extract contour structure of face images. They do produce local features, but they suffer from the disadvantages of too much computation and very high dimension of feature space. Apart from these techniques, LBP as a novel low-cost image descriptor for texture classification has also been introduced to the field of facial expression analysis [47, 48]. LBP can efficiently encode the texture features of micro-pattern information in

the face image, which is effective information for both face recognition and facial expression recognition applications.

Another kind of technique to represent faces is to model the appearance changes of faces. Holistic spatial analysis including PCA [76], LDA [77], ICA [53] and Gabor wavelet [46] analysis have been applied to either the whole-face or specific face regions to extract the facial appearance changes while the proposed technique is based on detecting non-linear features that correspond to lattices of the maximum energy. It is based on extracting the non-linear features corresponding to contours of the face, which are appearing due to different expressions. Three fuzzy lattices having maximum energy are selected as top three dimensions of the face as they contain most of the expression related data. The chapter is organized in the following manner. The details of proposed non-linear dimensionality reduction technique using fuzzy lattices are described in Section 6.2. The multiclass SVMs used for classification is presented in Section 6.3. Section 6.4 describes the simulation results for facial expression recognition. Finally, the concluding remarks are given in Section 6.5.

6.2 Proposed Fuzzy Lattice based Technique

The features of a pattern are significant for the recognition process. The strong features of a pattern result in simple classifier design. The proposed technique is based on detection of non-linear features corresponding to contours of the face, which are appearing due to different expressions.

(a)

(b)

Figure 6.1 (a) Original image of Mahatma Gandhi (b) Sketch of Mahatma Gandhi

Figure 6.1 (a) shows the original image of Mahatma Gandhi. Figure 6.1 (b) shows the sketch of Mahatma Gandhi. From Figure 6.1 (b), it is observed that the face of a person can be represented as a sketch with minimum lines corresponding to contours of the face, which are appearing due to different expressions. Fuzzy lattice based technique is developed to extract these prominent features, which are sufficient to recognize a person.

6.2.1 Block Diagram

The block diagram of the proposed fuzzy lattice based technique for the application of facial expression recognition is shown in Figure 6.2.

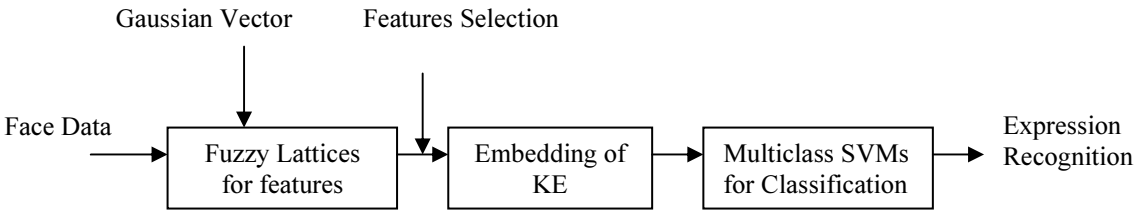


Figure 6.2 Block diagram of fuzzy lattice based technique

The Gaussian functions have been used to describe non-linear relation between the nearest neighborhood elements of the image. All the elements of the image have non-linear relation with the nearest neighborhood elements. It results in the formation of number of fuzzy lattices that have been used for feature extraction. The information generated due to any change in facial expression represents KE involved corresponding to change occurring in the facial expressions. Schrödinger equation is an important tool to ascertain the extent of the non-linearity. Thus, analysis of the fuzzy lattices to obtain the KE has been carried out using solution of Schrödinger's equation. Three fuzzy lattices having the maximum KE are selected for top three features (dimensions) of the face image as they contain most of the facial expressions related data. The KE value is embedded in three dimensional space and then multiclass SVMs have been used to distinguish various facial expressions. The mathematical details of the fuzzy lattice based technique for non-linear dimensionality reduction are described in the section that follows:

6.2.2 Mathematical Analysis

The proposed non-linear dimensionality reduction technique is based on rejecting features not related to facial expressions.

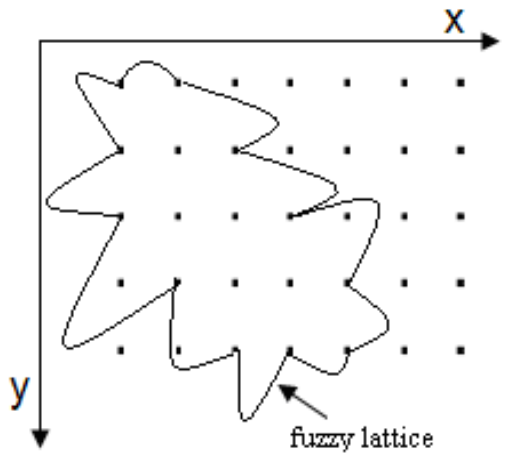


Figure 6.3 Representation of local non-linear relation

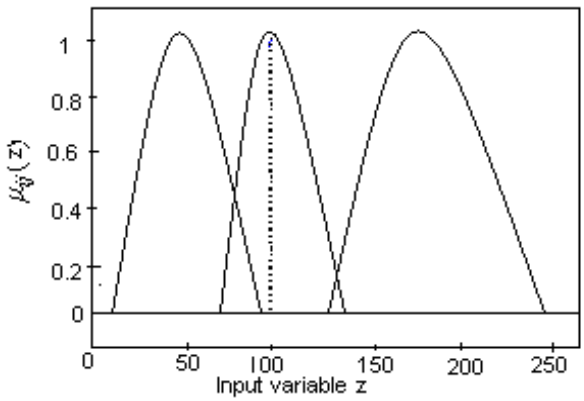


Figure 6.4 Gaussian membership functions

To capture features of the face that vary due to different expressions, the relation between the nearest neighborhood elements of the image is assumed to be non-linear as shown in Figure 6.3. The judgment of the feature amount is done by membership functions. Here, the Gaussian functions that have been used as fuzzy

membership function to describe non-linear relation between the nearest neighborhood elements that can be described as follows:

$$\mu_{ij}(z) = e^{\frac{-(z-m_i)^j}{2\sigma_i^2}} \tag{6.1}$$

where m and σ_i are the mean and variance of the membership function μ_{ij}, $i = 1,...,4$; $j = 1,...,4$ and z is the gray level values of elements of the image. Using (6.1), sixteen Gaussian functions of different center, width and shape would appear as shown in Figure 6.4. If i and j are chosen greater than four than Gaussian functions nearly equal to impulse functions, thus not useful to detect non-linear features. All the gray level values of the elements belong to the nearest neighborhood elements with some membership degree of the Gaussian function. It results in the formation of number of fuzzy lattices [78-79] as shown in Figure 6.3. The fuzzy lattices that have been used in our approach are described as follows:

$$L_d = \underset{ij}{\cup} e^{\frac{-(z-m_i)^j}{2\sigma_i^2}} \qquad \text{For } d = 1,...,D \tag{6.2}$$

where D is the number of the fuzzy lattice formed in the image. Whenever any change occurs in the pattern of the facial expressions, its lattice deform accordingly. In order to select the useful information, the gray level value of the neighboring elements should lie in the range of the Gaussian membership functions. It results in the formation of number of fuzzy lattices and some isolated elements. The isolated elements have been discarded that results in removal of unimportant data. Each fuzzy lattice corresponds to dimension (feature) of the image. Three fuzzy lattices of the maximum KE are selected as top three dimensions of the image.

Let $S = \{L_d : d = 1,..., D\}$ be the set of all the existing fuzzy lattices.

The set *S* is partitioned as follows:

$$L_1 = \underset{d}{Max}\{K.E.(L_d : L_d \in S)\} \tag{6.3}$$

$$L_2 = \underset{d}{Max}\{K.E.(S \backslash L_1)\} \tag{6.4}$$

$$L_3 = \underset{d}{Max}\{K.E.(S \backslash L_1 \cup L_2)\} \tag{6.5}$$

Eq. (6.3) shows that the fuzzy lattice of maximum KE is considered as first dimension of the image. Eq. (6.4) shows that the fuzzy lattice of maximum KE excluding L_1 is considered as second dimension of the image. Similarly, (6.5) shows that the fuzzy lattice of maximum KE excluding L_1 and L_2 is considered as third dimension of the image. The information generated due to any change in facial activity represents KE. The analysis of the fuzzy lattices to obtain the KE has been carried out using solution of the Schrödinger's equation. Therefore, to obtain the non-linearity corresponding to the particular point, the Schrödinger equation is solved at every point of control. The KE is obtained by solving second order differential equation (Schrödinger's equation). Differentiating (6.2) with respect to z, (6.6) is obtained as follows:

$$\frac{\partial L}{\partial z} = \frac{-j}{2\sigma_i^2}\left((z - m_i)^{j-1}\right) e^{\frac{-(z-m_i)^j}{2\sigma_i^2}} \tag{6.6}$$

$$\frac{\partial^2 L}{\partial z^2} = \frac{-j}{2\sigma_i^2}\left(\frac{-j}{2\sigma_i^2}(z - m_i)^{2j-2} + (j-1)(z - m_i)^{j-2}\right) e^{\frac{-(z-m_i)^j}{2\sigma_i^2}} \tag{6.7}$$

Similarly, the lattices formation occur in *x* and *y* directions where x and y denote spatial coordinate of the image. The lattice in x direction can be represented by

replacing z with x in (6.2). After replacing z with x in (6.2), $\frac{\partial^2 L}{\partial x^2}$ can be obtained as follows:

$$\frac{\partial^2 L}{\partial x^2}=\frac{-j}{2\sigma_i^2}\left(\frac{-j}{2\sigma_i^2}(x-m_i)^{2j-2}+(j-1)(x-m_i)^{j-2}\right)e^{\frac{-(x-m_i)^j}{2\sigma_i^2}} \tag{6.8}$$

Similarly, the lattice in y direction can be represented by replacing z with y in (6.2). After replacing z with y in (6.2), $\frac{\partial^2 L}{\partial y^2}$ can be obtained as follows:

$$\frac{\partial^2 L}{\partial y^2}=\frac{-j}{2\sigma_i^2}\left(\frac{-j}{2\sigma_i^2}(y-m_i)^{2j-2}+(j-1)(y-m_i)^{j-2}\right)e^{\frac{-(y-m_i)^j}{2\sigma_i^2}} \tag{6.9}$$

The KE in x, y and z directions are computed separately using (6.7), (6.8) and (6.9) respectively. Finally, the embedding of KE computed in x, y and z direction is obtained as follows:

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)L=KE \tag{6.10}$$

where x, y and z are three orthogonal dimensions of the energy space. The value of embedded kinetic energy is then fed to Multiclass SVMs for facial expression classification. The steps of whole process have been described in algorithm given as follows:

6.2.3 Algorithm

Step 1: Read in grayscale face images

Step 2: Creation of non-linear associative membership sets using (6.1)

Step 3: Fuzzy Lattice formation using non-linear membership function applied on the face images using (6.2)

Step 4: Dividing top three dimensions into three orthogonal components

Step 5: Computing KE of lattices in *x, y* and *z* directions separately using (6.7), (6.8) and (6.9)

Step 6: Embedding of the extracted KE parameter in three dimensional space using (6.10)

Step 7: Multiclass SVMs for expression classification

In this section fuzzy lattices formation and computation of their energy have been described. Then three lattices of highest energy are selected as top three features. In next section, multiclass SVMs is described to distinguish various facial expressions.

6.3 Classification using Multiclass SVMs

The basic description of support vector machines (SVMs) can be phrased as a two class classification problem where data points are mapped into a high dimensional hyperspace so that they can be separated by a hyper plane [54]. A margin exists on each side of the hyper plane, which is distanced to the nearest set of data points of each class. A high margin indicates good separation and good generalization. The data points that sit on the margin are known as support vectors.

For facial expressions classification, the embedded KE vectors $g_j, j = 1,...,N$ are used as input to the SVMs system. All the classes are considered for the experiments, each one representing one of the basic facial expressions. The output of the SVMs system is a label that classifies the embedded KE under examination to one of the basic facial expressions.

The training data $(g_1, l_1),...,(g_1, l_N)$, where $g_j \in \Re^L$, are the KE vectors and $l_j \in \{1,...,6\}$, are the facial expression labels of the embedded KE. The training data are the facial expression labels of the embedded KE value. The multiclass SVMs problem solves only one optimization problem [55]. It constructs basic facial expressions rules, where the kth function $\boldsymbol{w}_k^T \phi(g_j) + b_k$ separates training vectors of the class *k* from the rest of the vectors, by minimizing the objective function:

$$\min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi}} \frac{1}{2} \sum_{k=1}^{6} \mathbf{w}_k^T \mathbf{w}_k + C \sum_{j=1}^{N} \sum_{k \neq l_j} \xi_j^k \tag{6.11}$$

Subject to the constraints

$$\begin{aligned} \mathbf{w}_{l_j}^T \phi(g_j) + b_{l_j} &\geq \mathbf{w}_k^T \phi(g_j) + b_k + 2 - \xi_j^k \\ \xi_j^k &\geq o, \quad \text{j=1,...,N,} \quad \text{k}\{1,...,6\}\backslash \text{l}_j \end{aligned} \tag{6.12}$$

where ϕ is the function that maps the deformation vectors to a higher dimensional space, where the data are supposed to be linearly or near linearly separable. C is the term that penalizes the training errors and $\boldsymbol{w}$ is the normal vector to the hyper plane. The vector $\boldsymbol{b} = [b_1 ... b_6]^T$ is the bias vector and $\xi = [\xi_1^1, ..., \xi_i^k, ..., \xi_N^6]^T$ is the slack variable vector. Then, the decision function is:

$$h(\boldsymbol{g}) = \arg\max_{k=1,...,6} (\mathbf{w}_k^T \phi(\boldsymbol{g}) + b_k) \tag{6.13}$$

Using this procedure, a test feature vector is classified to one of the basic facial expressions using (6.13). Once the multiclass SVMs system is trained, it can be used for testing, i.e., for recognizing facial expressions on new facial image sequences.

6.4 Results

The proposed algorithm for facial expression recognition is tested on Cohn-Kanade (CK), Japanese Female Facial Expression (JAFFE), Aleix Martinez and Robert Benavente (AR) and Static Facial Expressions in the Wild (SFEW) databases [80-84]. Sample images from CK, JAFFE, AR and SFEW databases are shown in Figure 6.5-6.8 respectively. The CK database contains 97 subjects, which posed in a lab situation for the six universal expressions and the neutral expression. The proposed fuzzy lattice technique is tested on database of 53 different subjects of CK database. In order to maximize the amount of training and testing data, the classifier accuracy is measured using the leave-one-subject-out cross-validation approach.

Figure 6.5 Sample Images from CK database

Figure 6.6 Sample Images from JAFFE database

Figure 6.7 Sample Images from AR database

Figure 6.8 Sample Images from the SFEW database

For this test, database is divided in 5 sets, which contain the sequences corresponding to 10 or 11 subjects (three sets with 11 subjects, two sets with 10 subjects). The sequences are used from a set as test sequences and the remaining sequences are used as training sequences. This test is repeated five times, each time leaving a different set out.

The JAFFE database contains 213 images of seven facial expressions (six basic expressions and neutral expression also) posed by 10 Japanese female models. The classifier accuracy has been measured using the leave-one-subject-out cross-validation approach i.e., every time, expression images of 9 out of 10 subjects are used as the training set and the images of the remaining subject are used as the testing set. The process is repeated for each subject.

AR database contains 4000 images corresponding to 126 subjects with 4 different facial expressions of each subject. The AR database is captured with different expressions, illumination conditions and occlusions (scarf and sunglasses). In the AR database, besides the lighting from the left and the right, lighting from both sides of each face is also adopted. The pictures are taken under strictly controlled conditions. No restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. have been imposed to participants. The experiment is conducted using the leave-one-subject-out cross-validation approach. In order to compute error rate with respect to certain facial expression, the image associated with it is used as test image. In order to recognize the test image, all images excluding the test one, are projected to reduced space. Then the test image is projected as well as recognition is performed.

The CK, JAFFE and AR databases have been captured in controlled lab environment. However, all three databases do not capture the conditions found in real world situations well. For experimental validation of the proposed algorithm in real-world environments, the database is needed that is captured in tough conditions. The Static Facial Expressions in the Wild (SFEW) is a dataset that has captured facial expressions in tough conditions. The word wild here refers to the challenging conditions, in which the facial expressions occur. The database covers unconstrained facial expressions, varied head poses, large age range, occlusions, varied focus, and different resolution of face and close to real world illumination. The SFEW database contains 700 images of seven facial expressions (angry, disgust, fear, happy, sad,

surprise and neutral) of 95 subjects. The Strictly Person Independent (SPI) Protocol for the SFEW database has been used for validation of the proposed algorithm that is divided into two sets. The sets are created in a strict person independent manner. Each set has seven subfolders corresponding to the seven expression categories. The images are divided on the basis of their expression labels in their respective expression folder. There are 346 images in set 1 and 354 images in set 2. For the evaluation of proposed algorithm, the experiment is conducted using twofold: first, the set one has been used for training and set two for testing and then vice-versa.

The expression analysis on data captured in lab-controlled and real world environment requires more sophisticated techniques at all stages of the approach, such as robust face localization/tracking, illumination and pose invariance. A unified model has been used for face detection, pose estimation and landmark localization using a mixture of trees with a shared pool of parts [85]. Every facial landmark has been modeled as a part and use global mixtures to capture topological changes due to viewpoint. The model is effective on standard face dataset as well as in the wild annotated dataset.

The multiclass SVMs has been used as classifier for computing recognition accuracy. The KE of fuzzy lattices has been computed using solution of the Schrödinger equations given by (6.7), (6.8) and (6.9). Three fuzzy lattices having maximum kinetic energy are selected. Finally, the KE parameter is embedded in three dimension space using (6.10). The embedded KE value is used as an input to a multiclass SVMs system for expression recognition. The training data are the facial expression of various face databases. The output of the SVMs system is a label that classifies the embedded KE value under examination to one of the basic facial expressions. Using this procedure, a test feature vector is classified to one of the basic facial expressions.

Table 6.1 Confusion matrix for fuzzy lattice technique on Cohn-Kanade database

Expressions	Anger (%)	Disgust (%)	Fear (%)	Sad (%)	Happy (%)	Surprise (%)	Neutral (%)
Anger	83.1	5.7	4.1	2.2	1.3	0	3.6
Disgust	7.0	85	1.8	1.8	0.6	0.8	3.0
Fear	5.6	3.2	83	4.3	1.1	1.2	1.6
Sad	0.3	0.2	0.9	97	0	0.3	1.3
Happy	1.0	0.8	0.2	0	96.5	0	1.5
Surprise	0	0.3	0.9	0	0	97	1.8
Neutral	5.8	7.4	2.3	4.1	1.1	1.3	78

Table 6.2 Confusion matrix for fuzzy lattice technique on JAFFE database

Expressions	Anger (%)	Disgust (%)	Fear (%)	Sad (%)	Happy (%)	Surprise (%)	Neutral (%)
Anger	87.9	4.1	1.7	0.4	0	0.1	5.8
Disgust	4.5	84.5	5.2	0.4	0.8	0.3	4.3
Fear	3.1	4.4	81.1	3.5	0.5	0.7	6.7
Sad	1.2	3.9	0.9	90.7	0.2	0.8	2.3
Happy	0.2	0.1	0.1	0.3	98.2	0.1	1.0
Surprise	1.1	0.3	0.5	0	0	97.4	0.7
Neutral	7.8	4.0	5.9	1.0	0.7	1.2	79.4

Table 6.3 Confusion matrix for fuzzy lattice technique on AR database

Expressions	Neutral (%)	Smile (%)	Anger (%)	Scream (%)
Neutral	89.1	3.3	2.0	5.6
Smile	1.4	90.9	4.2	3.5
Anger	6.6	3.8	87.6	2.0
Scream	1.5	1.5	2.6	94.4

Table 6.4 Confusion matrix for fuzzy lattice technique on SFEW database

Expressions	Anger (%)	Disgust (%)	Fear (%)	Sad (%)	Happy (%)	Surprise (%)	Neutral (%)
Anger	75.4	5.2	3.2	4.1	4.7	5.1	2.3
Disgust	6.9	71.2	4.9	9.9	1.3	2.2	3.6
Fear	7.8	7.6	63.8	3.0	3.5	2.7	11.6
Sad	7.1	9.1	1.9	71.2	1.5	1.7	7.5
Happy	1.3	2.2	2.9	2.4	81.3	4.2	5.7
Surprise	1.3	1.8	2.2	4.4	2.8	81.6	5.9
Neutral	3.0	8.1	1.5	5.8	1.8	1.3	78.5

The confusion matrices obtained with fuzzy lattice technique on CK, JAFFE, AR and SFEW databases are shown in Tables 6.1-6.4 respectively. The diagonal entries of the confusion matrix are the rates of facial expressions that are correctly classified, while the off-diagonal entries correspond to misclassification rates. An analysis of the confusion matrices for CK database (Table 6.1) and JAFFE database (Table 6.2) suggests that the best recognized categories are happy and surprise. The other expressions are highly confused with each other. An analysis of the confusion matrix for AR database (Table 6.3) suggests that expressions captured under different illumination have also been recognized accurately using proposed technique

The technique is based on detecting three lattices of the maximum features variations that corresponds to expressions. Thus, the effect of light variation is neutralized in the proposed technique. An analysis of the confusion matrix for SFEW database (Table 6.4) suggests that the most difficult to recognize is fear expression, which is highly confused with anger and disgust. It is also blended with neutral expression and hence a lower detection value. It is evident that proposed technique has high recognition accuracy on CK, JAFFE and AR but comparatively lower accuracy for SFEW database. This is due to the fact that SFEW database has been captured in uncontrolled environment conditions.

Table 6.5 Recognition accuracy of Cohn Kanade database

Techniques	PCA	LDA	Isomap	LLE	GVLVM	LBP	Gabor feature	Fuzzy Lattice
Dimensions	43	5	18	30	48	52	56	3
Recognition accuracy (%)	72.4	74.7	77.2	77.9	66.2	80.2	78.4	88.51

Table 6.6 Recognition accuracy of JAFFE database

Techniques	PCA	LDA	Isomap	LLE	GVLVM	LBP	Gabor feature	Fuzzy Lattice
Dimensions	43	5	18	30	48	52	56	3
Recognition accuracy (%)	74.1	75.6	79.9	80.7	68.0	81.6	79.1	88.46

Table 6.7 Recognition accuracy of AR database

Techniques	PCA	LDA	Isomap	LLE	GVLVM	LBP	Gabor feature	Fuzzy Lattice
Dimensions	45	5	18	30	48	52	56	3
Recognition accuracy (%)	71.1	74.6	77	78.2	64.30	79.1	78.2	90.50

Table 6.8 Recognition accuracy of SFEW database

Techniques	PCA	LDA	Isomap	LLE	GVLVM	LBP	Gabor feature	Fuzzy Lattice
Dimensions	43	5	18	30	48	52	56	3
Recognition accuracy (%)	51.8	52.2	54.0	54.9	47.0	18.7	59.2	74.1

The proposed fuzzy lattice technique is compared with PCA, LDA, Isomap, LLE, GVLVM, LBP and Gabor feature based techniques. The experimental results of recognition accuracy along with dimensions of embedded space on CK, JAFFE, AR and SFEW databases are shown in Tables 6.5-6.8 respectively. Facial expression

recognition rates changes with the dimensions of the embedding space. It is observed that the recognition rates increase with the dimensions of the embedding space at the beginning, but when the dimension of the embedding space reaches to a value as shown in Tables 6.5-6.8, the expression recognition rates reach nearly to their maximum values. Fuzzy lattice technique is based on extracting the non-linear features corresponding to contours of the face, which are appearing due to different expressions. It is observed that three contours having maximum energy are sufficient to represent the facial expression. When the dimensions of the embedding space increases to three, the recognition rates reaches nearly to its maximum value. All the tests for expression recognition have been performed using Multiclass SVM classifier. SVM makes binary decisions, so the multiclass classification here is accomplished by using the one-against-rest technique, which trains binary classifiers to discriminate one expression from all others, and outputs the class with the largest output of binary classification.

The results demonstrate that the techniques, which perform very well on the CK, JAFFE, AR datasets that have been developed on lab controlled data, are not robust when it applied to SFEW dataset captured in more real world like conditions. However, the proposed technique has significantly improved recognition accuracy on CK, JAFFE, AR datasets as well as SFEW dataset captured in uncontrolled environment conditions. The proposed technique is based on detecting three lattices of the maximum feature variations that corresponds to facial expressions. Thus, the effect of pose and light variation is neutralized in the proposed technique and the expressions captured under different pose, illumination and occlusion have also been recognized accurately. It is observed that the proposed technique consistently performed better than other techniques achieving the highest accuracy at reduced dimensions of three.

6.5 Conclusions

An application oriented technique has been developed that directly accommodates the local non-linear behavior of facial expressions. The message being conveyed or expression being expressed by a person can be known efficiently by

proposed technique. The experimental results show that the recognition accuracy of proposed technique is better than PCA, LDA, Isomap, LLE, GVLVM, LBP and Gabor feature based techniques. It is showing promising results on data captured in lab controlled conditions as well as real world like environment. It also has advantage of very low dimension of feature space. The technique can also be applied for lips recognition, hand writing recognition and image tracking.

Chapter-7

Fuzzy Lattice based Technique for Classification of PQ Events

Chapter-7

Fuzzy Lattice based Technique for Classification of PQ Events

In this chapter, proposed fuzzy lattice based technique is used for classification of PQ events. The proposed technique is different from others, in respect that, it is based on the concept of local non-linear relation. It uses non-linear fuzzy functions to extract the feature specific data. To extract any change during change in the patterns of PQ events, non-linear Gaussian functions have been used, which results in the formation of fuzzy lattices. The fuzzy lattices have been expressed in the form of Schrödinger equation, to find the KE used, corresponding to any change occurring in the PQ events. Finally, the KE value embedded in 2-D space has been used to distinguish various PQ events. The proposed technique efficiently distinguishes various PQ events in a single cycle and works perfectly in real time.

7.1 Introduction

Power quality (PQ) has become an important issue to electricity consumers due to the wide use of delicate electronic devices. Voltage disturbances, interrupting manufacturing processes and microcontrollers are some of the major causes of poor PQ. Voltage swell and sag can occur due to lightning, capacitor switching, motor starting, nearby circuit faults, or accidents, and can lead to power interruptions. Harmonic currents due to non-linear loads throughout the network also degrade the quality of services to the sensitive high-tech customers, such as India's IT parks in Bangalore, Hyderabad and many other places. The massive rapid transit system, Metro Railways in Delhi and few other places in India have facilitated the massive use of semiconductor technologies in the auto-traction systems, resulting in the increased level of harmonic distortion. The solution to the PQ related problems

requires continuous monitoring and the acquisition of large amount of data from the distribution system.

The need of an automated PQ detection and classification system to determine the cause of PQ disturbances is emphasized in [100]. Several signal processing and statistical analysis tools have been presented for the detection and classification of PQ events [61]. Signal processing is generally called upon when there is a need to extract specific information from the raw data, which typically in power systems are the voltage and current waveforms. The objectives of collecting data through measurements or simulations largely dictate, which signal processing technique is to be utilized. Using such steady state data, statistical signal processing can be used to predict performance or the health condition of voltage regulators on distribution circuits. The necessity of improved detection performance for continuous monitoring of electric signals has motivated the development of several techniques that show a good trade-off between computational complexity and performance.

The local linear constraint does not work to extract minor (non-linear) changes from the pattern of PQ events, and such techniques lost the contents while embedding [17] and [22]. The proposed work has classified the PQ events by considering local data set as non-linear and used the concept of fuzzy non-linear lattices to extract any change in the pattern of the PQ events. To extract any change occurring in the patterns of PQ events, non-linear Gaussian functions have been used, which results in formation of fuzzy lattices. Finally, the change in KE during the change in pattern of PQ events is extracted by expressing the fuzzy lattice equations in terms of Schrödinger equation. The paper is organized in the following manner. Section 7.2 describes the configuration of the model for generation of PQ events. The details of proposed technique based on fuzzy lattice for features extraction is described in Section 7.3. The simulation results for classification of PQ events are discussed in Section 7.4. Finally, the concluding remarks are given in Section 7.5.

7.2 Power Quality Events Generation

The PQ events have been generated in the Power System Laboratory at Delhi Technological University (Formerly Delhi College of Engineering). The events are generated according to IEEE std. 1159-1995 laid down in monitoring manual [105].

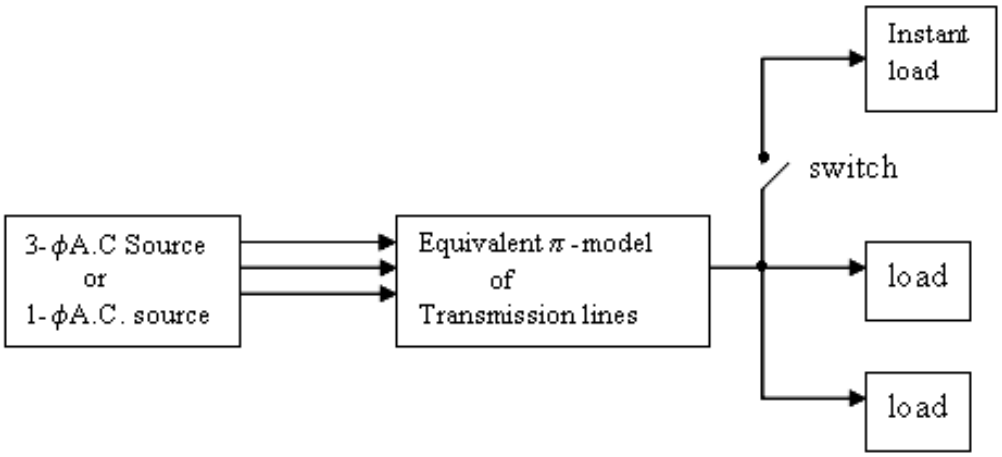
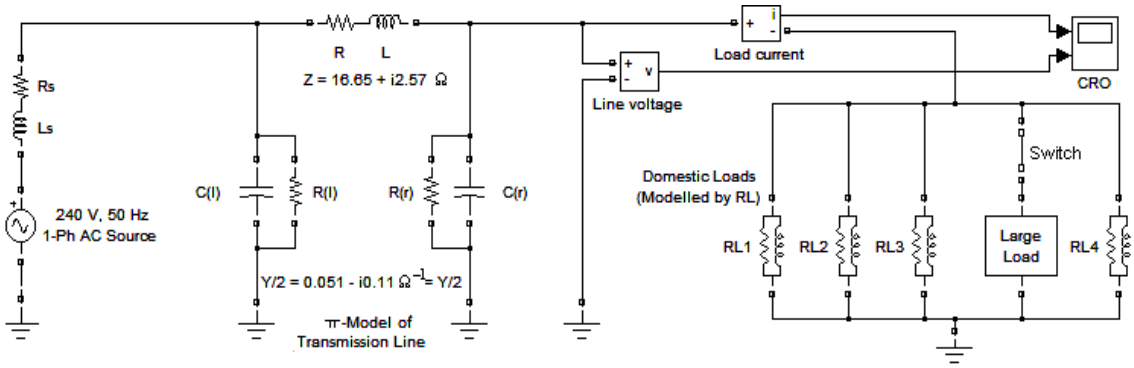
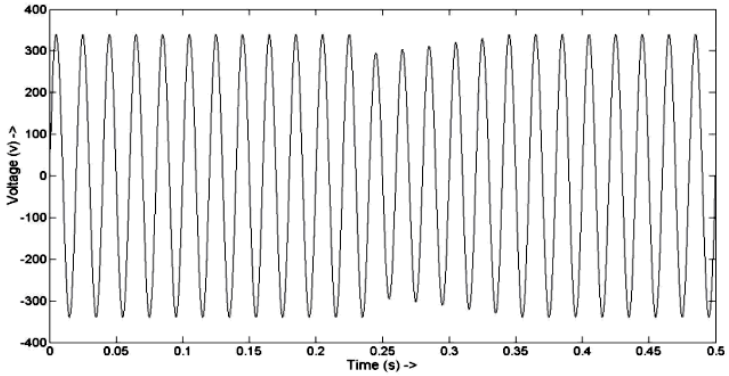


Figure 7.1 Configuration of the model for generation of PQ events

Figure 7.1 shows the configuration of the model used for generation of PQ events. The laboratory receives a 3-ϕ AC supply of 415 volts (Root mean square (RMS), line-to-line) with 50 Hz frequency. The PQ events generally occur due to sudden switching-on of a large load, such as induction motors (IMs); non-linear elements in the power system; circuit breakers; capacitor switching; lightning; or system faults. The events generated under different load conditions and system faults in the aforementioned laboratory are described below.



(a)

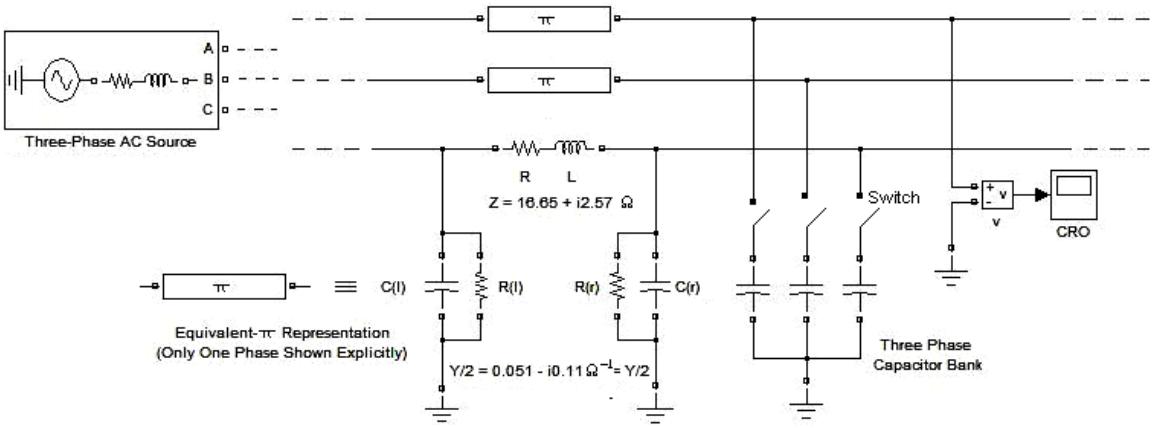


(b)

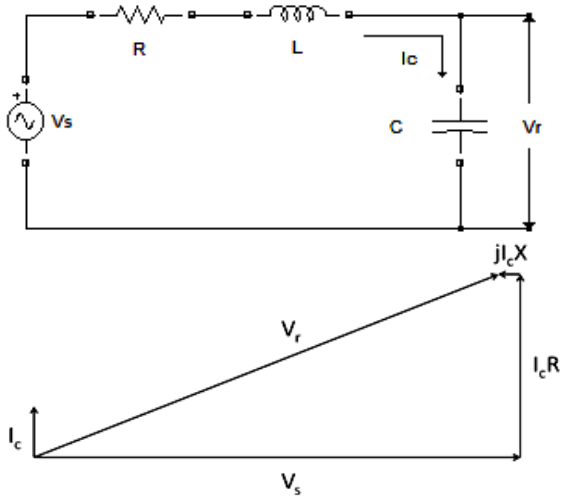
Figure 7.2 (a) Circuit representing occurrence of voltage sag (b) Output waveform of voltage sag

Voltage sag or dip refers to a fall in the voltage waveform at the receiver's end for a small interval of time. Voltage sags are caused due to the sudden switching-on of a large load, such as IM. Figure 7.2(a) represents a generalized circuit, which causes the occurrence of voltage sag due to the sudden switching-on of an arbitrary large load. The switch S represents the sudden switching operation. If the load is suddenly connected to the line then obtained output is shown in Figure 7.2(b).

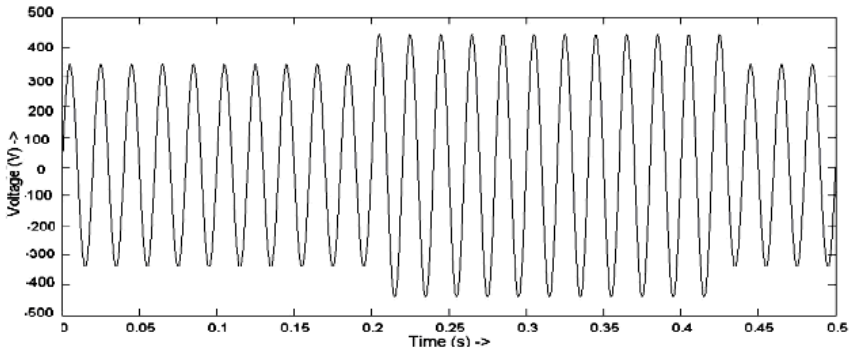
Voltage swell refers to a rise in the voltage from 1.0 p.u. (or 230V RMS) to a value above it. It generally occurs due to charging capacitors connected to the line, when the line is lightly loaded. All transmission lines contain some capacitance between the conductors of individual phases, in addition to a capacitance to earth.



(a)



(b)



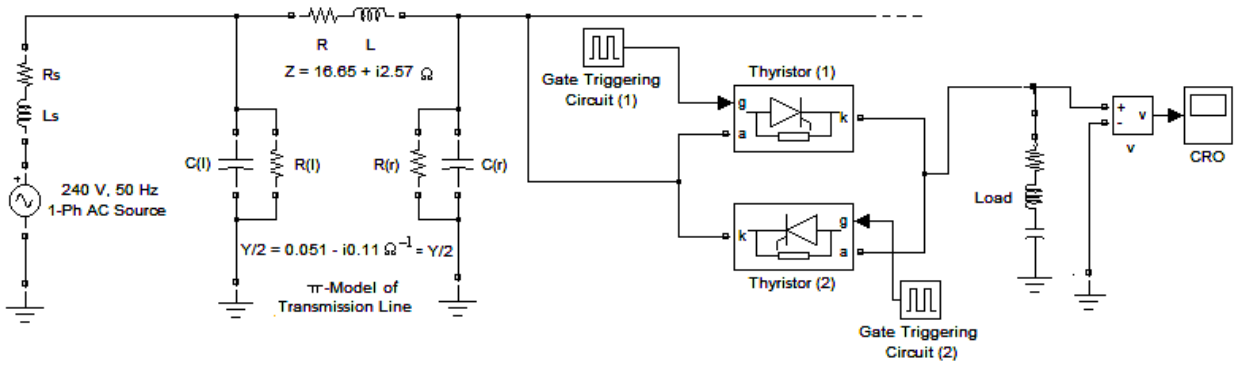
(c)

Figure 7.3 (a) Circuit representing occurrence of voltage swell (b) Phasor Diagram for Ferranti Effect (c) Output waveform of voltage swell

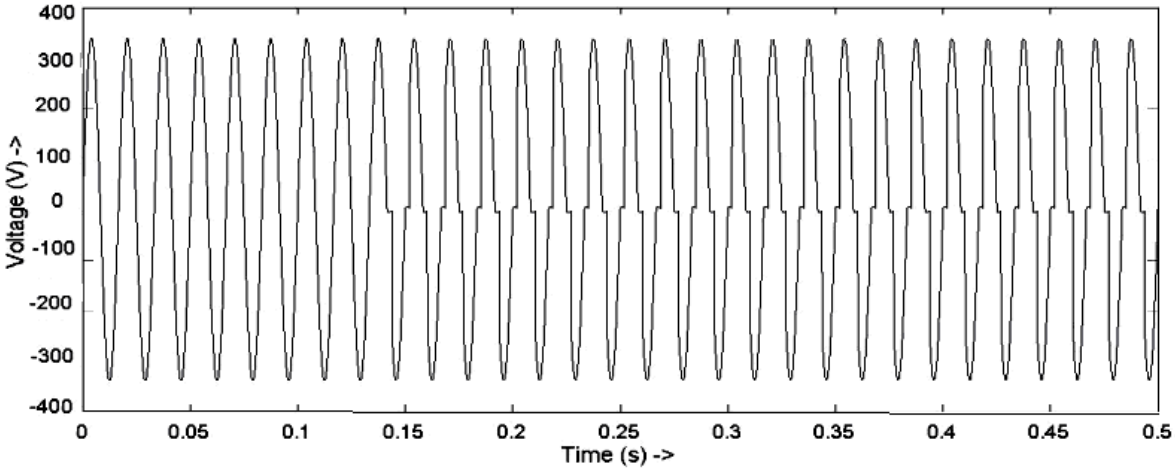
During periods of light loads, i.e., when the current flowing in the line is very low, the entire sending end voltage of the transmission line effectively appears across the capacitors, since the series impedance of the line causes a negligible voltage drop. This causes the capacitors to draw a charging current, which leads the input voltage. Thus, the net voltage at the receiving end, which is the phasor sum of the sending end voltage and the voltage drop across the series impedance due to this charging current, is greater than the sending end voltage, resulting in voltage swell. This is called Ferranti Effect.

The circuit used for recording voltage swell is shown in Figure 7.3(a). The phasor diagram displaying Ferranti Effect is shown in Figure 7.3(b). The voltage waveform shown in Figure 7.3(c), shows the occurrence of swell when the load on the line is suddenly decreased to a very low value.

Harmonic distortion means an undesirable change in the wave shape of the voltage waveform from the regular sinusoidal shape. Distortion also occurs due to the use of non-linear loads. Magnetic circuits, which are an indispensable part of all electromechanical devices like transformers, induction motors, DC motors etc, have non-linear characteristics, i.e., the relation between flux flowing through a magnetic path and the magnetic motive force (MMF) causing that flux to flow is highly non-linear. Thus, if one of the two quantities is a sinusoid, the other is non-sinusoidal.

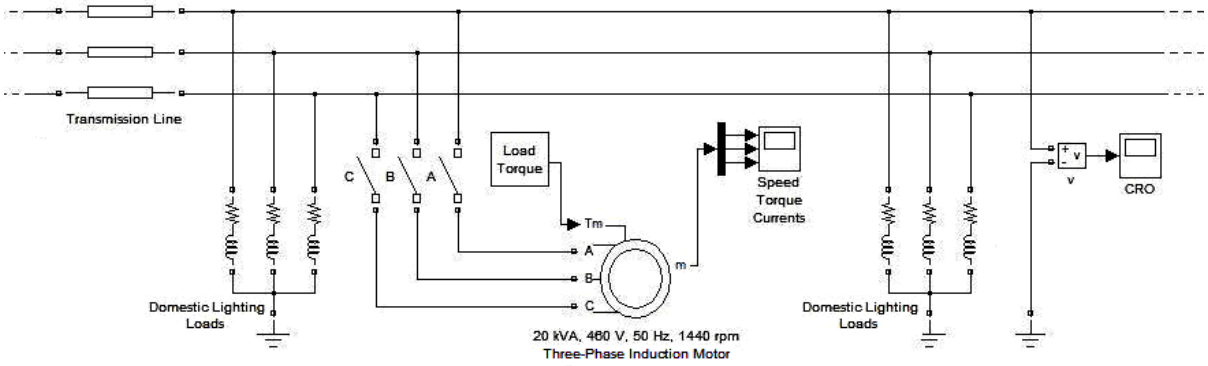


(a)

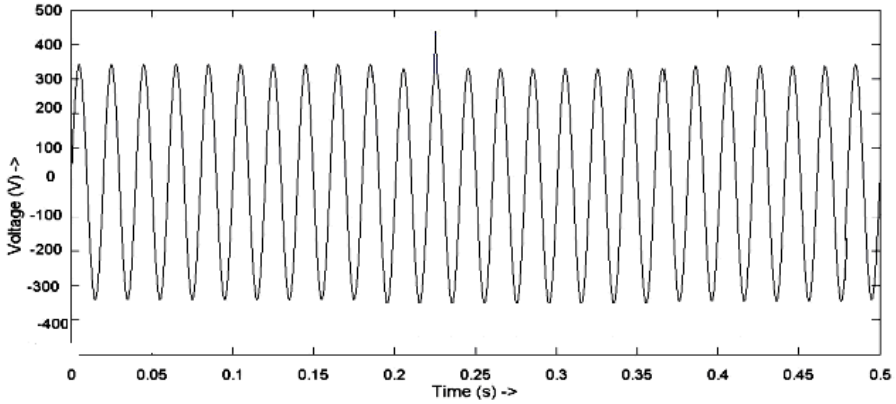


(b)

Figure 7.4 (a) Circuit representing occurrence of harmonic distortions (b) Output waveform of harmonic distortion



(a)



(b)

Figure 7.5 (a) Circuit representing occurrence of transient (b) Output waveform of transient

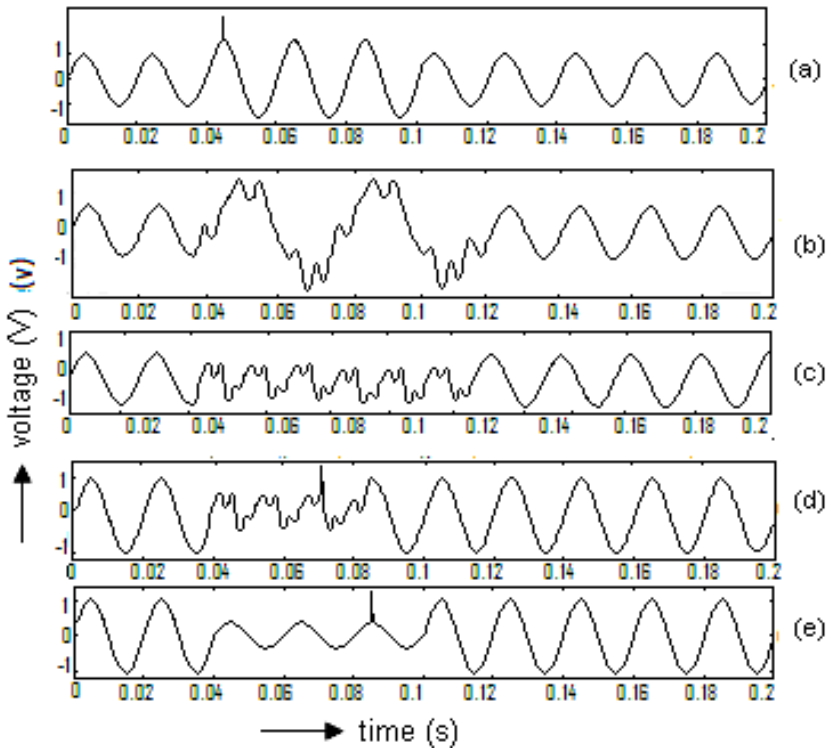


Figure 7.6: Multiple events (a) swell and transient (b) swell and harmonics (c) sag and harmonics (d) sag, transient and harmonics (e) sag and transient

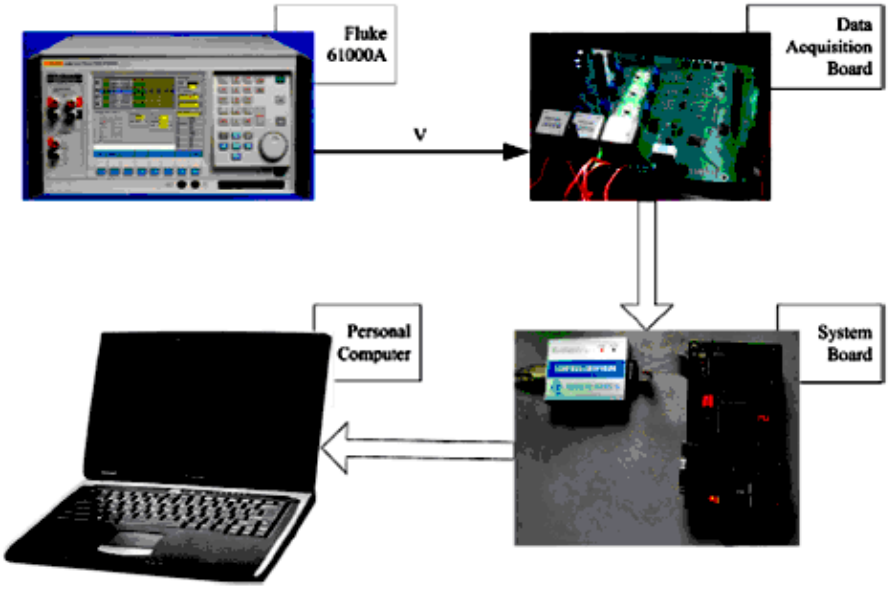


Figure 7.7 Photo of setup for generation of PQ events

AC voltage controllers are thyristor based devices that convert fixed AC voltage to variable AC without a change in the frequency. The circuit of a 1-ϕ full-wave AC voltage controller used for speed control of a 1-ϕ induction motor is shown in Figure 7.4(a). The firing angles for thyristors are controlled in such a way so as to obtain the desired RMS voltage at the input terminals of the motor. The waveform of the voltage at the input of the motor is shown in Figure 7.4(b).

Transient is commonly known as switching surges or voltage spike. They can be caused by circuit breakers out of adjustment, capacitor switching, lightning, or system faults. They are characterized by a sudden non power change in frequency, high amplitude, fast rise and decay times, and high energy concern. Figure 7.5(a) shows the circuit used for recording transient caused due to the sudden switching on of a large load, such as 3-ϕ IM connected to the line through a switch, in parallel with household lighting load. Figure 7.5(b) shows its voltage waveform.

A multiple event refers to that cycle of the signal, which has more than one type of event. The occurrence of sag and transient in a single cycle of a signal is an example of multiple PQ events. In other words, there are cases when two events almost occur at the same time called multiple events. When capacitive load with switch is used, initially transient and then swell occurs as shown in Figure 7.6(a). Harmonics with swell occurs in case of capacitive and inductive load as shown in Figure 7.6(b). In case of resistive and inductive load, harmonics with sag occurs as shown in Figure 7.6(c). In case of fast trip breaker after a short circuit, sag, harmonics and transient occurs as shown in Figure 7.6(d). In case of heavy load switching, transient with sag occurs as shown in Figure 7.6(e).

The photo of setup for generation of PQ events is shown in Figure 7.7. These generated events are recorded through data acquisition board (DAQ) NI USB-9215 and a computer, which performed all the required signal processing. The sampling rate of the DAQ was set to 50kS/s. The recorded events in computer are classified by implementing the proposed algorithm using MATLAB. All the data (waveform) of events are generated and recorded in Power System laboratory at Delhi Technological University. The events are not generated from MATLAB simulink. Only the circuits corresponding to real time generated events are drawn in MATLAB simulink for

better symbolic representation of components and equipments used in laboratory. The technique used to accomplish the task of classification of PQ events is described in the next section.

7.3 Fuzzy Lattice based Technique for Events Classification

In this chapter, non-linear dimensionality reduction technique has been used to extract any substantial change occurring in the patterns of the PQ events. To describe the non-linear relation between two nearest neighborhood elements, the Gaussian vector as membership function has been used. It represents the non-linearity and can easily interpolate the non-linear relation between elements. In addition, Gaussian membership functions have been carried out because their shapes are easily specified and Gaussian curves are intuitive and easy to manipulate. The non-linear relationship among elements provides more realistic view of the quantized event. Therefore, it has been used to detect any changes in the patterns of the PQ events.

7.3.1 Block Diagram

The Gaussian functions with various mean and variance have been used to represent non-linear relation between elements of PQ events, which results in formation of the fuzzy lattices. The information generated due to any change in power quality pattern represents KE involved corresponding to change occurring in the events. Schrödinger equation is an important tool to ascertain the extent of the non-linearity. Thus, analysis of the fuzzy lattices to obtain the KE parameter has been carried out using solution of Schrödinger's equation. It contains all the dynamic information about a signal. Two fuzzy lattices having maximum KE are selected for top two features of the events as they contain most of the important information of data. Finally, the KE parameter embedded in 2-D space has been used to distinguish various PQ events.

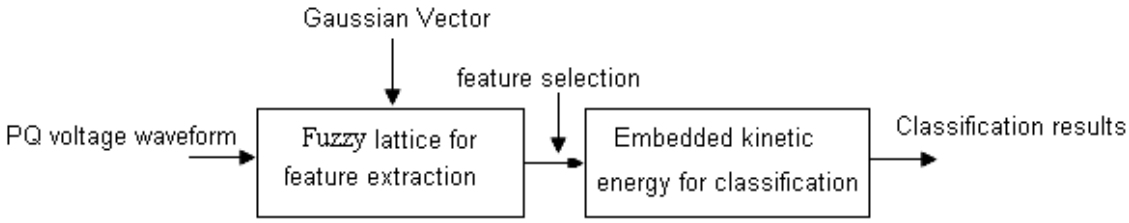


Figure 7.8 Block diagram of fuzzy lattice technique for PQ Events classification

The block diagram of the proposed fuzzy lattice based technique for classification of PQ events is shown in Figure 7.8. The fuzzy lattice based technique is used for extracting useful features of the PQ events. The use of fuzzy lattice appears to be interesting where the non-linearity is involved. The change in the features of the pattern results in equivalent KE. These fuzzy lattices have been represented in terms of Schrödinger equation to find the equivalent KE, which is obtained by solving second order differential equation (Schrödinger's equation). Thus, the values of KE depend upon amount of the change occurring in the patterns of events. The two fuzzy lattices having maximum KE are selected as top two dimensions of the events. Finally, the KE parameter embedded in 2-D space has been used to distinguish various PQ events. The mathematical details of the proposed technique (fuzzy lattice based technique) for classification of PQ events are described in the section that follows:

7.3.2 Mathematical Analysis

The proposed non-linear dimensionality reduction technique using fuzzy lattices is based on reducing most of the redundant data not related to features. To describe non-linear relation between the two nearest neighborhood elements, the Gaussian vector as membership function has been used. It represents non-linearity and can easily interpolate the non-linear relation between two nearest neighborhood elements. The non-linear relationship among elements provides more realistic view of the quantized event. Therefore, it has been used to detect any change in the patterns of the PQ events. The Gaussian function that has been used as fuzzy membership function to represent the non-linearity can be represented as follows:

$$\mu_{ij}(y) = e^{-(y-m_i)^j / 2\sigma_i^2} \tag{7.1}$$

Where m_i and σ_i are the centre and width of the membership function $\mu_{ij}, i = 1,...,r,\ j = 1,...,s$ and y is the amplitude at time x of the PQ event.

All the values of the PQ event belong to nearest neighborhood values with some membership degree of the Gaussian function. The Gaussian membership functions with different mean and power have been used to check the non-linear relation between two nearest neighborhood elements of the PQ event. It results in formation of the fuzzy lattices [78, 79].

These fuzzy lattices are in turn Gaussian vectors as they hold the property described as follows:

If the vector $G^T = (G_1,...,G_n)$ is Gaussian, all its components are thus Gaussian random variable. If the components $G_k,\ k = 1,...,n$ of a random vector G are Gaussian and independent, the vector G is thus also Gaussian. If G_k are Gaussians then $\cup G_k,\ \forall k$ are Gaussian: $G_k \cap G_l \neq \varnothing,\ k \neq l$

The fuzzy lattice that has been used can be described as given below:

$$L = D \in R \,|\, D = e^{-(y-m_i)^j / 2\sigma_i^2} \tag{7.2}$$

The set L is a fuzzy lattice and binary relation ≤ is defined as follows:

$$\sup(D_1, D_2) = \max(D_1, D_2) \in L \tag{7.3}$$

$$and \quad \inf(D_1, D_2) = \min(D_1, D_2) \in L \tag{7.4}$$

From definition, it is clear that lattice asymptotically increases. But, in proposed technique lattice structure can increases or decreases depending on relation between PQ elements. Due to this fuzzy relation, it is called fuzzy lattice. In fuzzy lattice, whenever any change occurs in the patterns of the PQ events, its lattice deforms accordingly. In order to select the useful information, the value of neighboring elements should lie in the range of Gaussian membership function. It results in

formation of number of fuzzy lattices and some isolated elements. These isolated elements have been discarded, which results in removal of unimportant features. Each fuzzy lattice corresponds to dimension (feature) of the PQ event that can be described as follows:

$$D_q = e^{-(y-m_i)^j / 2\sigma_i^2} \tag{7.5}$$

All the dimensions of the event are independent, non-redundant, and non-overlapping and so they are orthogonal to each other. Whenever any change occurs in the power quality patterns, there is corresponding change in KE used. The two fuzzy lattices having maximum KE are selected as top two dimensions.

Let $S = \left\{ L_q : q = 1,..., Q \right\}$ be the set of all the existing lattices.

The set S is partitioned as follows:

$$D_1 = Max\left\{ K.E.\left\{ L_q : L_q \in S \right\} \right\} \tag{7.6}$$

$$D_2 = Max\left\{ K.E.\left\{ S \setminus D_1 \right\} \right\} \tag{7.7}$$

Equations (7.6) and (7.7) represent the selection criteria of choosing the best two fuzzy lattices required. Schrödinger's equation is an important tool to ascertain the extent of the non-linearity. Thus, the analysis of the fuzzy lattices to obtain the KE parameter has been carried out using solution of Schrödinger equation. It contains all the dynamic information of a signal. Therefore, to obtain the non-linearity corresponding to the particular PQ event, the Schrödinger equation is solved at every point of control. The information generated due to change occurs in the PQ pattern represents KE used. This KE is obtained by solving the second order differential equation (Schrödinger's equation).

Differentiating (7.5) with respect to *y*,

$$\frac{\partial D}{\partial y} = \frac{-j}{2\sigma_i^2}(y-m_i)^{j-1} e^{-(y-m_i)^j / 2\sigma_i^2} \tag{7.8}$$

$$\frac{\partial^2 D}{\partial y^2} = \frac{-j}{2\sigma_i^2}\left[\frac{-j}{2\sigma_i^2}(y-m_i)^{2j-2} + (j-1)(y-m_i)^{j-2}\right] e^{-(y-m_i)^j/2\sigma_i^2} \quad (7.9)$$

$\frac{\partial^2 D}{\partial x^2}$, can be obtained on the similar lines

$$\frac{\partial^2 D}{\partial x^2} = \frac{-j}{2\sigma_i^2}\left[\frac{-j}{2\sigma_i^2}(x-m_i)^{2j-2} + (j-1)(x-m_i)^{j-2}\right] e^{-(x-m_i)^j/2\sigma_i^2} \quad (7.10)$$

KE in *x* and *y* directions are computed separately using (7.9) and (7.10). Finally, embedding of the KE computed in *x* and y direction is obtained as follows:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) D = K.E \quad (7.11)$$

The value of embedded kinetic energy has been used for distinguishing various PQ events. The steps of whole process have been described in algorithm given as follows:

7.3.3 Algorithm

Step 1: Read in power quality waveform

Step 2: Creation of non-linear associative membership set using (7.1)

Step 3: Fuzzy Lattice formation using non-linear membership function applied on the PQ events using (7.2)

Step 4: Dividing top two dimensions into two orthogonal components

Step 5: Computing the 2-D Kinetic Energy parameter separately using (7.9) and (7.10)

Step 6: Embedding of the extracted KE parameter in 2-D space using (7.11)

7.4 Results

To prove the effectiveness and efficiency of any algorithm used for detection and classification PQ events, it is important to implement the algorithm on events that occur in practical circumstances. The proposed algorithm is tested on real time generated PQ events that are recorded in the Power System Laboratory at Delhi Technological University (Formerly Delhi College of Engineering) as mentioned in Section-7.2. Five hundred samples of PQ events have been used for testing of proposed algorithm. These events have been generated and recorded in 10 different conditions and 50 events have been recorded in each condition. The sag is recorded for 10 different load conditions; swell for 10 different values of capacitive load with switch; transient is recorded for switching with induction load at different values and harmonics are recorded for different non-linear loads. Similarly the multiple events (sag, transient and harmonic, swell and transient, sag and transient, swell and harmonic, sag and harmonic) are recorded as mentioned in Section 7.2.

Table 7.1 Range of the embedded kinetic energy for PQ Events

S.No.	PQ Events	Range of embedded KE
1	Sag, transient and harmonic	1982-2125
2	Swell and transient	1825-1981
3	Sag and transient	1654-1824
4	Swell and harmonic	1455-1650
5	Sag and harmonic	1200-1454
6	Transient	1001-1199
7	Swell	600-999
8	Harmonic	421-599
9	Sag	298-420

Table 7.2 Confusion matrix for fuzzy lattice technique on PQ events

PQ Events	sag, transient and harmonics	swell and transient	sag and transient	swell and harmonics	sag and harmonics	transient	swell	harmonics	sag
sag, transient and harmonics	498	2	-	-	-	-	-	-	-
swell and transient	1	498	1	-	-	-	-	-	-
sag and transient	-	1	499	-	-				
swell and harmonic	-	-	-	498	2	-	-	-	-
sag and harmonic	-	-	-	2	497	1	-	-	-
transient	-	-	-	-	01	499			
swell	-	-	-	-	-		498	02	
harmonic	-	-	-	-	-		01	499	
sag	-	-	-	-	-			01	499

Table 7.3 Classification accuracy of PQ events

Technique	Classification accuracy (%)
S-transform	96.17
Wavelet-based neural network	95.24
Fuzzy lattice	99.67

The proposed technique for classification of real time generated PQ events based on embedded KE has been implemented using MATLAB. The results of the range of embedded kinetic energy for five hundred samples of each PQ events: sag, swell, transient, harmonic distortions and multiple events are depicted in Table 7.1. The information generated due to any change in events represents KE involved corresponding to the change occurring in the events. This, KE is obtained by solving the second order differential equation. Thus, the values of KE depend upon amount of the change occurring in pattern of events. Like the total change during occurrence of multiple events (sag, transient and harmonics) is maximum in compare to other events, so KE is also maximum. Similarly, values of KE for other events depending on change in pattern of PQ event are depicted in Table 7.1. The problem of finding the KE at different instances for the same type of PQ event has also been considered.

In Table 7.2, confusion matrix shows the performance of the classifier based on proposed algorithm. It demonstrates the number of times the proposed technique distinguishes the events correctly. For example, when 500 samples of transient event are tested by proposed algorithm, 499 times classified correctly as transient and once misclassified as harmonic. The five hundred samples of each event have been tested by proposed technique for classification and the diagonal of the matrix shows the correct recognition of the events.

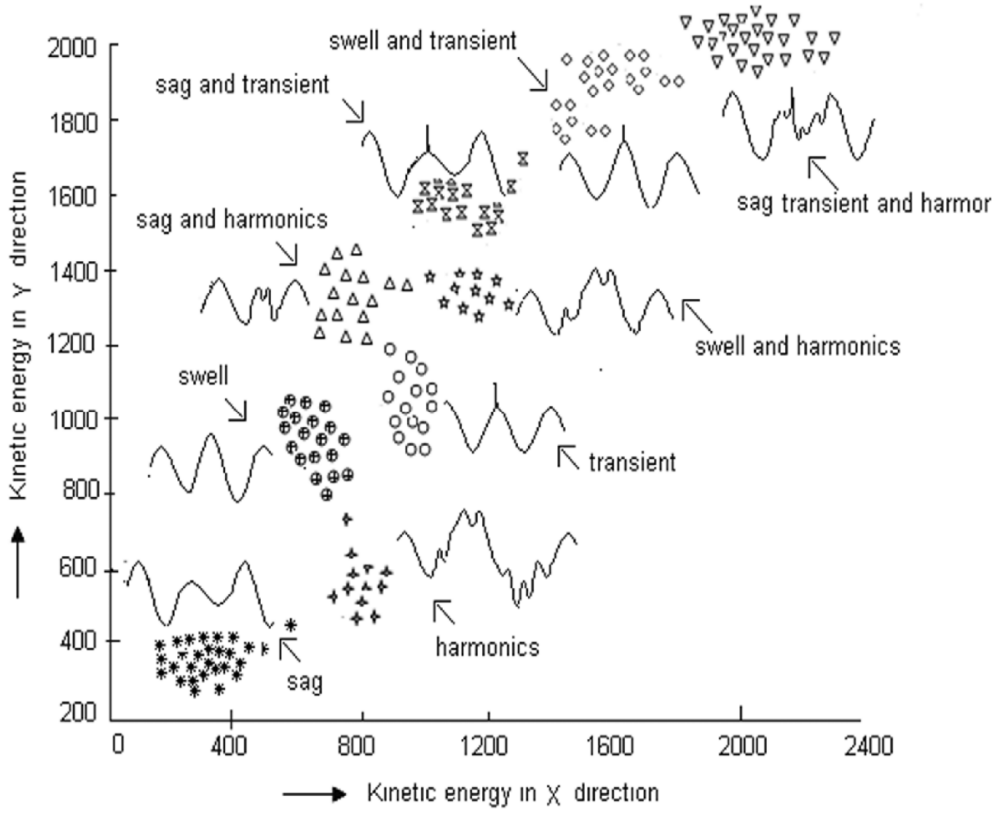


Figure 7.9 Results for PQ events Classification based on embedded KE

For classification of events, the computed KE is embedded in a 2-D energy space which is shown in Figure 7.9 where each point corresponds to an event. From Figure 7.9, it is observed that the different events are separated from each other and similar ones are grouped together. Therefore, the fuzzy lattice based technique works well to discover the non-linear pattern of the events.

The proposed algorithm is compared with the S-transform and wavelet-based neutral network algorithms. The experimental results of classification accuracy of PQ events are tabulated in Table 7.3. The comparative results clearly demonstrate that the proposed technique works well to discover the non-linearity of the various PQ events and outperforms other algorithms such as S-transform and wavelet-based neutral network. The proposed non-linear dimensionality reduction technique is based on detecting two lattices of features corresponding to maximum change.

7.5 Conclusions

The technique is originally a dimension reduction technique and is very different as compared to the wavelet based techniques. The embedded KE has been used to distinguish various types of PQ events like sag, swell, transient, harmonic distortion and multiple events. It is demonstrated that the algorithm can efficiently distinguishes the PQ events based on variation of values for embedded KE The value of embedded KE shows a substantial change whenever there is any change in PQ event. As a result, the fuzzy lattice based algorithm can efficiently distinguish the real time generated PQ events in a single cycle while the earlier techniques based upon wavelets distinguishes the PQ events in the few cycles of the power signal.

Chapter-8

Data Compression using Statistically Matched Wavelet

Chapter 8

Data Compression using Statistically Matched Wavelet

In this chapter, the performance of the Discrete Wavelet Transform is improved using statistically matched wavelet as the mother wavelet for compression of power quality (PQ) events. The matched wavelet is designed based on the characteristic of PQ events. The concept of Fractional Brownian motion (FB-m) has been used to design the matched wavelet. The results are simulated using MATLAB and are compared with the mostly used Daubechies wavelet.

8.1 Introduction

In the recent past, wavelets have been using extensively for extracting distinct features of various types of PQ events [107]. A number of wavelets have been applied for detection of PQ events, like, Daubechies, Multi-wavelets, Dyadic and Symlets [115]. The wavelet multi-resolution analysis has been adopted by a number of researchers where an algorithm based on the energies of the decomposed signals has been proposed to distinguish different classes of PQ events [101]. Wavelet transform do not have a unique basis like Fourier Transform, which is one of the reasons that wavelets are finding applications in diverse fields. Since the basis is not unique, it is desired to find a wavelet that can provide the best representation of the signal. One of the exciting advantages of wavelet over Fourier analysis is the flexibility they afford in the shape and form of the analyzer, which cuts up and studies the signal of interest [106]. However, with flexibility comes the difficult task of choosing or designing the appropriate wavelet or wavelets for a given application. The proposed technique overcomes this difficulty and presents a new approach for detection of PQ events using wavelet statistically matched to characteristics of PQ events. Statistically matched wavelet provides better result of compression as the events can be detected up to six level of decomposition while mostly used Daubechies wavelet cannot allow event detection beyond four level of decomposition [108], [110], [111].

The PQ events detection and classification system have been presented using higher order cumulants with quadratic classifiers [101]. A technique for the PQ events classification using support vector machine has been proposed [102]. The covariance based behavior of several features, determined from the voltage waveform within a time window for PQ event detection and classification, has been analyzed [103]. All these techniques can detect the fault disturbances but the number of samples required is large which results in complex algorithm and does not work in the real time.

The real time generated events are detected after compressing of events using statistically matched wavelet. The statistically matched wavelet is designed based on the characteristic of events using the concept of Fractional Brownian motion (FB-m). The proposed technique is compared with Daubechies wavelet to show its superiority in compression of the PQ events. To classify the detected events, Iterative Closest Point (ICP) algorithm is used which classifies detected event even in presence of outlier points and Gaussian noise. The technique is applied to classify the various PQ events like transient, sag, swell and harmonics.

The chapter is organized in the following manner. Section 8.2 describes the proposed system applied for compression of PQ events. Section 8.2.1 discusses the concept of Fractional Brownian Motion to design the wavelet matched to PQ event. Section 8.2.2 explains the method for estimating self similarity index (H) of signal. Using estimated H, method to design statistically matched wavelet is described in Section 8.2.3. Based on statistically matched wavelet, the procedure used to design perfect reconstruction filter-bank is given in Section 8.2.4. These designed filters are applied for detection of events by compressing the PQ data. Performance parameter used to measure the quality of data compression is described in Section 8.3. Section 8.4 discusses the simulation results for compression of PQ events. Section 8.5 describes the technique for classification of PQ events. Section 8.6 discusses the simulation results for classification of detected event. Finally, concluding remarks are given in Section 8.7.

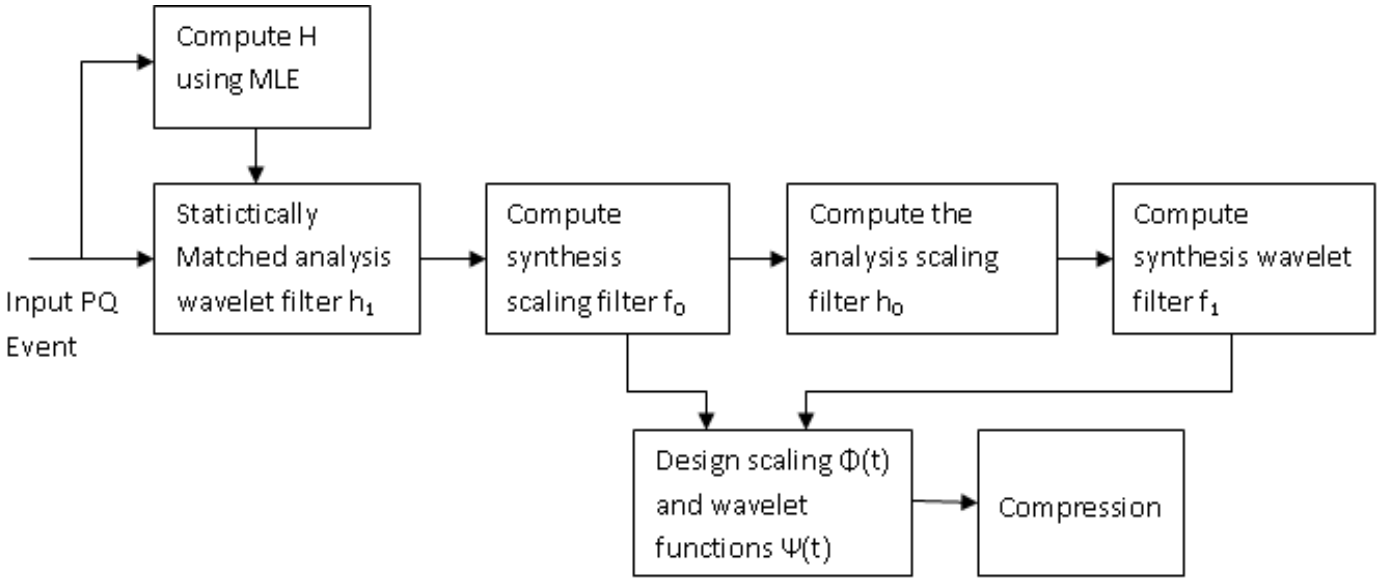


Figure 8.1 Proposed System of data compression using matched wavelet

8.1 Proposed System

In the proposed method, wavelet is designed that is matched to a given signal in the statistically sense. Furthermore, the methods are presented to design perfect reconstruction filter-bank.

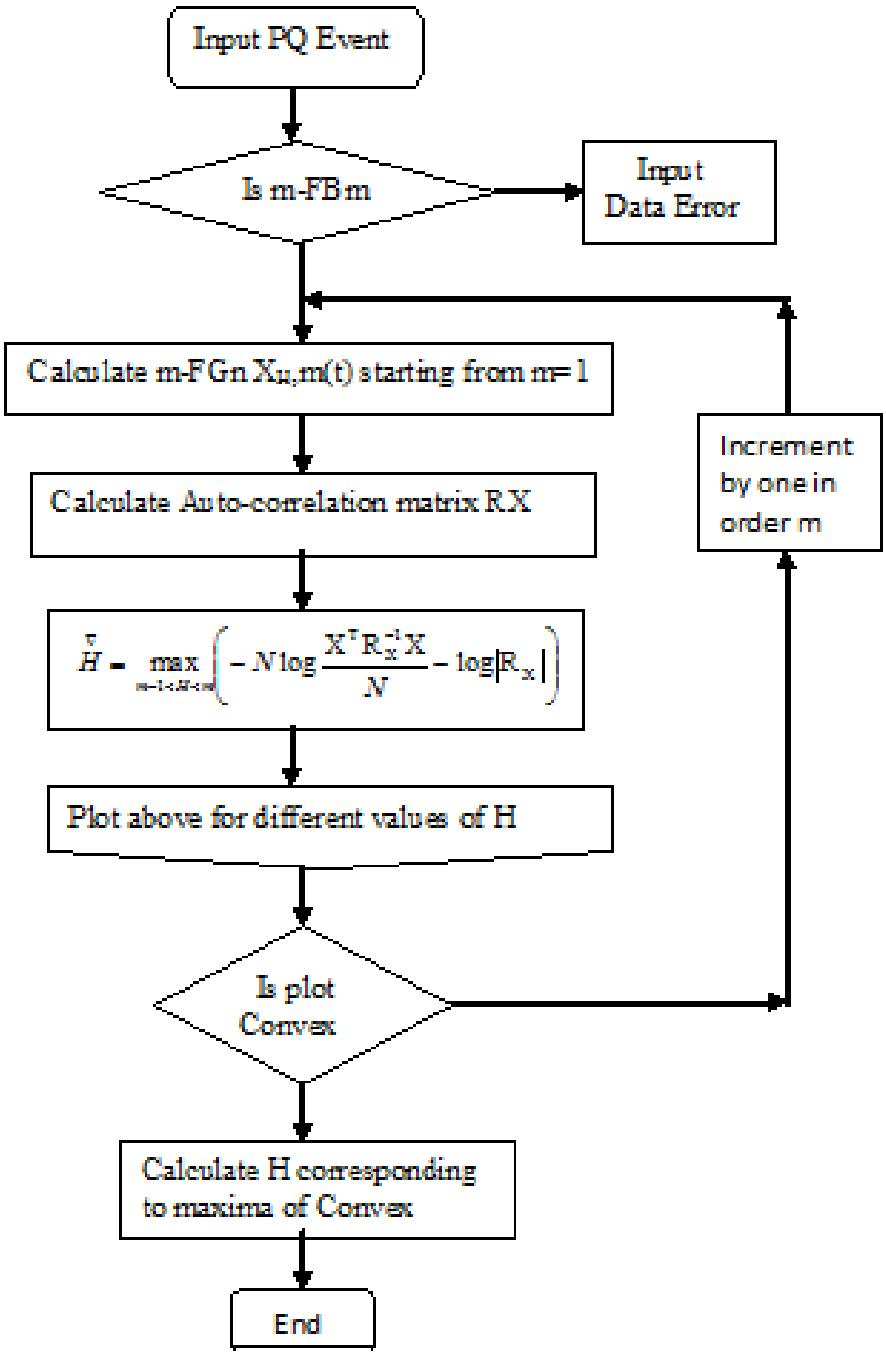


Figure 8.2 Estimation of H parameter for matched wavelet

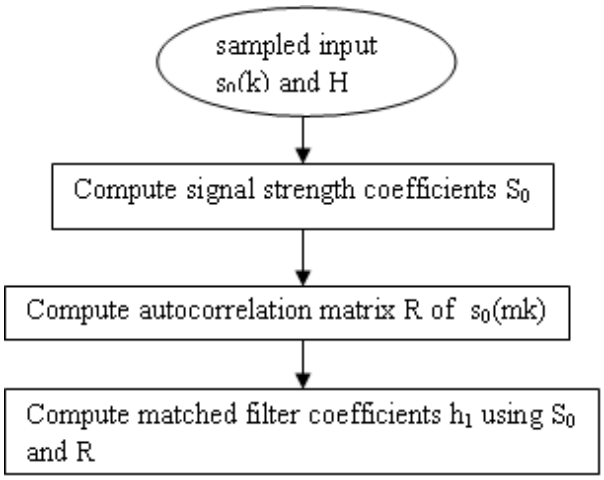


Figure 8.3 Estimation of Analysis wavelet filter coefficients

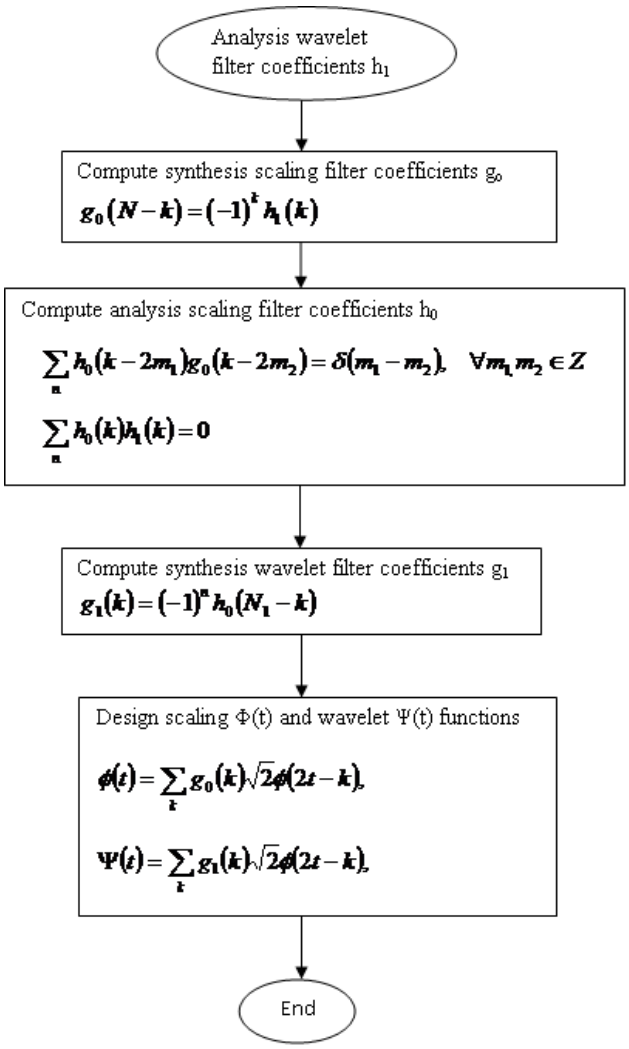


Figure 8.4 Design scaling and wavelet functions of Bi-orthogonal wavelet

Estimated wavelet for different signals is compared with standard bi-orthogonal wavelets for the application of compression. The proposed system of designing wavelet that is matched to a given signal in the statistically sense is shown in Figure 8.1. The FBm is used for designing of statistically matched wavelet. In first step, the estimation of self similarity index (H parameter) for PQ events is required for the designing of matched wavelet. Maximum Likelihood Estimation (MLE) has been used to estimate H parameter. This is described in Flowchart of Figure 8.2. In second step, using estimated H, analysis wavelet filter h_1 is designed from the given event which is elaborated in Flowchart of Figure 8.3. In third step, Synthesis scaling filter g_0 is computed from h1. In fourth step, analysis scaling filter h_0 is computed using g_0 and then synthesis wavelet filter g1 is computed using h_0. Furthermore, the scaling and wavelet functions have been designed from the scaling and wavelet filter using 2-scale recursive relations. The filters are constructed using the properties of perfect reconstruction bi-orthogonal filter bank. The procedure of designing scaling and wavelet functions are described by Flowchart of Figure 8.4. The designed filters are applied for detection of PQ events by compressing the PQ data. The details of each flowchart are described in subsequent sections.

8.1.1 Fractional Brownian motion

The signals which are collected for the purpose of designing matched wavelet must be self similar. It is well known that a number of natural and man-made phenomenon exhibit self similar characteristics. Also known as fractal processes, these waveforms arises in natural landscape, ocean waves, and distribution of earthquakes. They have found profound applications in various engineering fields like image analysis, audio compression, and characterization of texture in bone radiographs etc. These processes are in general non stationary and they exhibit self-similarity in the statistically sense. A class of these signals is called 1/fB processes, which have measured power spectral density that decays by a factor of 1/fB. Since 1/fB processes simultaneously exhibit statistically scale invariance and time invariance thus wavelet-like bases having both scaling and shifting properties can represent these signals well. In a real world power system, voltages and especially

currents are practically always distorted from pure sinusoid thus they are random in nature. The exact behaviour of random signals cannot be predicted but if the signal is stationary its behaviour can be measured by its statistical properties like mean, variance and auto correlation. The entire PQ patterns are not stationary, as the time of generation of pattern is not known and ensemble average of PQ events is not same. Behavior of such signals has been detected so far by standard wavelets. But no work has been reported for designing of wavelet matched to the characteristic of PQ events. In the proposed technique, the concept of Fractional Brownian motion (FBm) denoted by $B^H(t)$ is used for designing of such wavelet [110], [113]. So far, it has been shown that FBm based methods are more appropriate to deal with non stationary signals. FBm is a continuous-time Gaussian process with the properties of stationary increments and self-similarity. FBm has stationary increments means the difference as given by (8.1) is independent of time.

$$B^H(t) - B^H(s) \approx B^H(t-s) \tag{8.1}$$

And self similarity means a self-similar object is exactly or approximately similar to a part of itself i.e. its statistical properties are scale invariant that can be expressed as follows:

$$B^H(at) \approx a^H B^H(t) \tag{8.2}$$

where the random process $B^H(t)$ is self similar with the similarity index H for any scale parameter $a > 0$**.** The equality in above equation holds in the statistically sense only. The estimation of self similarity index 'H' of signal is done by MLE [116] which is described in the next section.

8.1.2 Estimation of H Parameter

The Maximum likelihood estimator (MLE) can be used for estimation of self similarity index (H). In MLE, variance of this estimator nearly achieves the minimum bound. If the input process is m-FBm, then its mth-order incremental process results in an mth-order Fractional Gaussian Noise (m-FGn) stationary process. MLE is performed using a discrete m-FGn vector X and is denoted as follows:

$$\hat{H} = \max_{m-1<H<m} \left(-N \log \frac{\mathbf{X}^{\mathbf{T}} \mathbf{R}_{\mathbf{X}}^{-1} \mathbf{X}}{N} - \log \left| \mathbf{R}_{\mathbf{X}} \right| \right) \tag{8.3}$$

where $\mathbf{R_x}$ autocorrelation matrix of a discrete m-FGn process and is given by:

$$r_{H,m}^{m} = \frac{\sigma_H^2}{2} (-1)^m \sum (-1)^j \binom{2m}{m+j} |n+j|^{2H} \tag{8.4}$$

where $\sigma_H^2 = Var\{B_H(1)\} = 1/gamma(2H+1)|\sin(H)|$

8.1.2.1 **Algorithm:** Estimation of H by MLE

Step 1: Form the mth-order incremental process X (i.e., discrete m-FGn) from the given input signal starting from m=1using (8.5)

$$X_{H,m}(t) = \sum_{j=0}^{m} (1)^{m-j} \binom{m}{j} B_{H,m}(t+jl) \tag{8.5}$$

Step 2: For a = M, the process given in (8.2) can be written as follows:

$$B^H(Mt) \approx M^H B^H(t) \tag{8.6}$$

Step 3: The corresponding autocorrelation function of discrete input process is as:

$$r_a(Mk_1, Mk_2) = M^{2H-1} \sigma_H^2 \left(|k_1|^{2H} - |k_1 - k_2|^{2H} + |k_2|^{2H} \right) \tag{8.7}$$

Step 4: Compute autocorrelation matrix of the resulting m-FGn process using (8.4)

Step 5: Plot the graph of bracketed term in (8.3) for various values of H. If the graph is convex upward, the value of H corresponding to minima in the graph is the correct value of H.

Step 6: If the graph is linear, increment m, and repeat steps 1, 2 and 3.

Using the value of H, FB-m designs the wavelet based on characteristic of PQ event.

8.1.3 Design of Statistically Matched Wavelet

Using the characteristic of the input signal, it is desired to find a wavelet that can provide the best representation of the input signal. Similar to the two-band case,

one usually uses a multi-resolution analysis (MRA) with a scaling factor of M to construct M-band wavelets [112]. M-band wavelets have better results for compression in comparison to the 2-band wavelet.

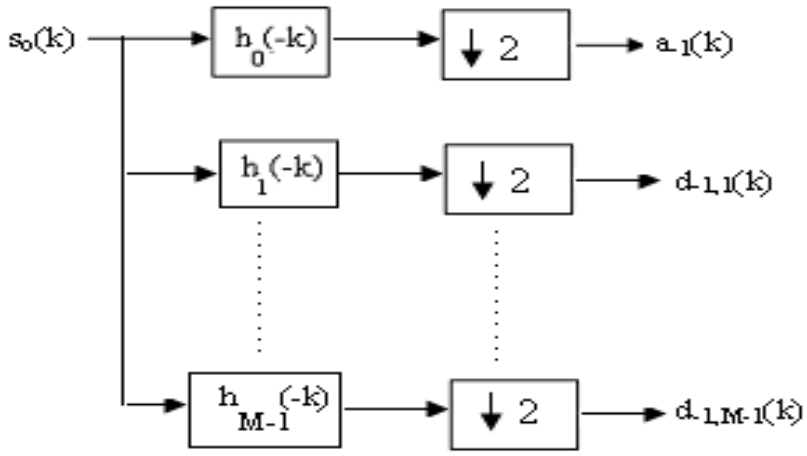


Figure 8.5 M -band Wavelet

Figure 8.5 shows the analytical part of the M-band filter-bank. Here, $s_0(k)$ is applied as input that is the sampled version of given continuous time signal $s(t)$. Here, h_0 is the low pass filter, $h_1, h_2, ..., h_{M-2}$ are band pass filters, and h_{M-1} is the high pass filter such that $a_{-1}(k)$ represents the approximation coefficients at scale $j = -1$, and $d_{-1,1}(k), d_{-1,2}(k), ..., d_{-1,M-1}(k)$ represents the finer information in wavelet subspaces at scale $j = -1$. Let us assume that the length of filter h_{M-1} is $N=3$ then, $d_{-1,M-1}(k)$ can be written in terms of filter weights as given below:

$$d_{-1,M-1}(k) = [h_{M-i}(0)s_o(Mk) + h_{M-i}(1)s_o(Mk+1) + h_{M-i}(2)s_o(Mk+2)] \quad (8.8)$$

The signal $d_{-1,M-1}(k)$ provides the detail or high pass information. Therefore, we would like to express this signal as smoothening error signal. Now, if the centre weight $h_{M-1}(1)$ of the filter is set to unity, then (8.8) can be rewritten as given below:

$$d_{-1,M-1}(k)s_0(Mk+1) = -\left\{-\left[h_{M-i}(0)s_o(Mk) + h_{M-i}(2)s_o(Mk+2)\right]\right\}$$

$$= s_0(Mk+1) - \hat{s}_0(Mk+1) = e(k) \quad (8.9)$$

where $\hat{s}_0(Mk+1) = \left\{-\left[h_{M-i}(0)s_o(Mk) + h_{M-i}(2)s_0(Mk+2)\right]\right\}$

This equation plays a key role in the estimation of the matched wavelet. With the centre weight fixed to unity, $\hat{s}_O(Mk+1)$ is the smoother estimate of $s_0(Mk+1)$ from the past as well as from future samples. Thus, $d_{-1,M-1}(k)$ is the error in estimating $s_0(Mk+1)$ from its hood and hence represents finer information. Since $d_{-1,M-1}(k)$ represents error signal between the actual value and its estimated value, the mean square value of this error signal has to be minimized. Here, resulting filter h_{M-1} is observed to be a high pass filter. Using (8.9), $d_{-1,M-1}(k)$ can also be computed as follows:

$$d_{-1,M-1}(k) = s_0(Mk+J_1) - \mathbf{F_0^T S_0} \tag{8.10}$$

where J_1= index of centre weight of filter h_{M-1},

$$\boldsymbol{F_0} = [h_{M-i}(0)h_{M-i}(1).....h_{M-i}(J-1) \; .h_{M-i}(J+1).....h_{M-i}(N-1)]^T$$

$$S_0 = -\left[s_0(Mk)\,s_0(Mk+1).....s_0(Mk+J_1-1).....+s_0(Mk+J_1+1).....s_0(Mk+N-1)\right]^T$$

Mean square error from (8.10) is given as follows:

$$\therefore E\left[e^2(k)\right] = E\left[s_0^2(Mk+J_1)\right] - 2E\left[s_0(Mk+J_1)F_0^T S_0\right] + E\left[F_0^T S_0 S_0^T F_0\right] \tag{8.11}$$

To minimize E [e^2 (k)], the derivative of E[e^2 (k)], with respect to F_0 is equal to zero.

$$\frac{\partial E\left[e^2(k)\right]}{\partial \boldsymbol{F_0}} = -2E\left[s_0(Mk+J_1)\mathbf{S_0^T}\right] + 2\boldsymbol{R_0 F_0} = 0$$

$$\Rightarrow E\left[s_0(Mk+J_1)\boldsymbol{S_0^T}\right] = \boldsymbol{R_0 F}_0 \tag{8.12}$$

Therefore, if statistics of the input signal are known, then using (8.12), filter h_{M-1} can be computed.

8.2.3.1 **Algorithm:** To design statistically matched wavelet

Step 1: Assign the estimated value of H for the known signal as computed in Section 3 and sampled values of input signal $s_0(k)$.

Step2: Calculate the signal strength coefficient as described below:

$$S_0 = -[s_0(Mk)s_0(Mk+1).....s_0(Mk+J_1-1)...s_0(Mk+J_1+1)...s_0(Mk+N-1)]^T \quad (8.13)$$

Step 3: Compute the autocorrelation matrix R of $s_0(Mk)$ with fixed length N.

$$r_{B,m}^{H}(Mk_1, Mk_2) = M^{2H-1}\sigma_H^m(-1)^m \left\{ |k_1-k_2|^{2H} - \sum_{j=0}^{m-1}(-1)^j \binom{2H}{j} \left[\left(\frac{k_1}{k_2}\right)^j |k_2|^{2H} + \left(\frac{k_2}{k_1}\right)^j |k_1|^{2H} \right] \right\} \quad (8.14)$$

where $\sigma_H^m = \sigma_H^2$

Step 4: After the computation of $\boldsymbol{S_0}$ and autocorrelation matrix $\boldsymbol{R_0}$, statistically designed filter coefficients $\boldsymbol{F_0}$ can be computed using (8.12) which can be given as follows:

$$\boldsymbol{F_0} = \boldsymbol{R_0}^{-1} * E[s_0(Mk+J_1)\boldsymbol{S_0}^T]$$

Step 5: The filter coefficients can be assigned as follows:

$$\boldsymbol{F_0} = [h_{M-i}(0)h_{M-i}(1).....h_{M-i}(J_1-1)h_{M-i}(J_1+1).....h_{M-i}(N-1)]^T$$

Step 6: The resulting filter h_{M-1} is the statistically matched wavelet.

Next synthesis scaling filter, analysis scaling filter, and synthesis wavelet filter are computed using the properties of perfect reconstruction bi-orthogonal filter bank.

8.1.4 Design of Perfect Reconstruction Filter-bank

Here, bi-orthogonal wavelet is used for designing of perfect reconstruction filter-bank. The bi-orthogonal wavelet is a wavelet where the associated wavelet transform is invertible but not necessarily orthogonal. Designing bi-orthogonal wavelets allows more degrees of freedoms than orthogonal wavelets. This family of wavelet exhibits the property of linear phase which is needed for signal reconstruction. One additional degree of freedom is the possibility to construct

symmetric wavelet functions. To illustrate the perfect reconstruction property, Figure 8.6 shows the 2-band perfect reconstruction bi-orthogonal filter-bank, which contains two decompositions filters h_0 and h_1 and two reconstructions filters g_0 and g_1.

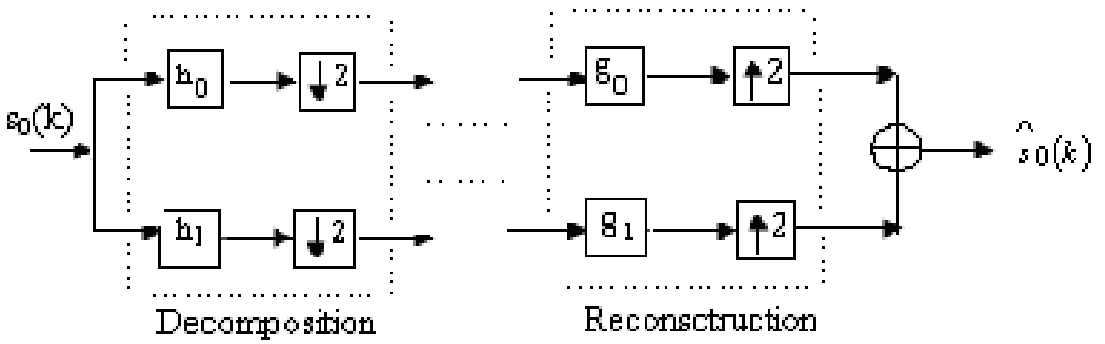


Figure 8.6 2-Band perfect reconstruction filter-bank

For a perfect reconstruction bi-orthogonal filter-bank, the scaling filter g_0 and its dual h_0, wavelet filter g_1 and its dual h_1 are related as follows:

$$h_1(k) = (-1)^n g_0(N_1 - k) \tag{8.15}$$

$$g_1(k) = (-1)^n h_0(N_1 - k) \tag{8.16}$$

In the bi-orthogonal case, there are two scaling functions, which may generate different multi-resolution analyses, and accordingly two different wavelet functions. So the numbers of coefficients in the scaling sequences may differ. The scaling sequences must satisfy the following bi-orthogonality conditions.

$$\sum_n h_0(k-2m_1) m g_0(k-2m_2) = \delta(m_1 - m_2), \quad \forall m_1 m_2 \in Z \tag{8.17}$$

$$\sum_n h_0(k) h_1(k) = 0 \tag{8.18}$$

For the two-band wavelet system, the scaling function and wavelet function are defined by the two-scale difference equation as follows:

$$\varphi(t) = \sum_k g_0(k) \sqrt{2} \varphi(2t - k) \tag{8.19}$$

$$\Psi(t) = \sum_k g_1(k)\sqrt{2}\varphi(2t-k) \tag{8.20}$$

In case of M-band, (8.19) and (8.20) can be modified as follows:

$$\varphi(t) = \sum_k g_0(k)\sqrt{M}\varphi(2t-k) \tag{8.21}$$

$$\Psi(t) = \sum_k g_1(k)\sqrt{M}\varphi(2t-k) \tag{8.22}$$

8.2.4.1 **Algorithm:** To design bi-orthogonal filter-bank

Step 1: Estimate the statistically matched wavelet h_1 of the order N_1 from given signal as described in Section 8.2.4.

Step 2: If it is desired to design wavelet filter of order N2 > N1, then append extra zeros before and after such that its order is N2.

Step 3: Use (8.15) to compute the synthesis-scaling filter g_0.

Step 4: Compute the analysis scaling filter h_0 using (8.17) and (8.18). Here h_0 is computed only for those values of m1 and m2 for which the vectors $f_0(k-2m)$ overlap with $h_0(k)$.

Step 5: Use (8.16) to compute the synthesis wavelet filter g_1.

Step 6: Design the scaling and wavelet functions from the scaling and wavelet filter using 2–scale recursive relations given by (8.19) and (8.20)

8.2 Performance Measurement

Power quality patterns are detected after compressing the data using statistical matched wavelet. To assess the quality of compression, the signal to noise ratio (SNR) in decibels is used as the performance index.

$$SNR\,(db) = 10\log_{10}\left(\frac{\sum_{k=0}^{T-1}\left|s(k)\right|^2}{\frac{1}{T}\sum_{k=0}^{T-1}\left|s(k)-\hat{s}(k)\right|^2}\right) \tag{8.23}$$

where T is total number of samples in input signal, s(k) is the original signal and $\hat{s}_0(k)$ is the reconstructed signal. The parameter SNR provides measurement of similarity and mismatch between signal and matched wavelet. Higher SNR means higher the quality of reconstructed signal.

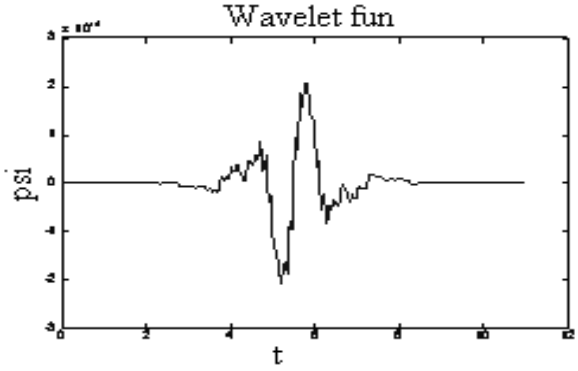
8.3 Results of Statistically Matched Wavelet

The PQ events are generated according to IEEE laid down in monitoring manual [105]. These events are detected by compressing the PQ event data using statistically matched wavelet. To estimate the H parameter for designing of statistically matched wavelet consider the segment of transient PQ event waveform. Now estimate the H parameter using (8.3) for the purpose of matching scaling function with the applied PQ event. For designing of statistically matched wavelet h_0, h_1, g_0 and g_1 are calculated as procedure explained in Section 5 and results are tabulated in Table 8.1.

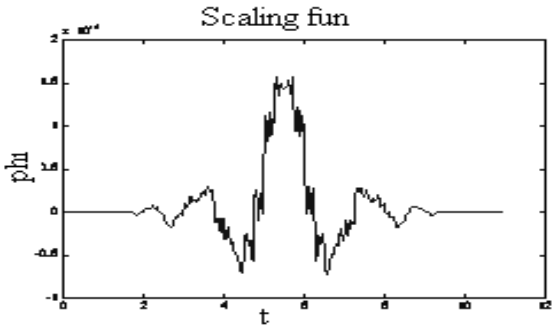
Table 8.1 Analysis and synthesis filter coefficients for transient event

Event	Value of H	Filter Length	Filter Coefficients
Transient	3.1	N=13	h_0(-n)=[0 0 -0.518 -0.1354 -0.2243 -0.2243 -0.2243 -0.2243 -0.2243 -0.1888 -0.0150 0 0] h_1(-n)=[0 0 0 0 0.2602 0.0702 0.1000 -0.3657 0.0908 0 0 0 0] f_0(-n)=[0 0 0 0 -0.5181 0.1637 -1.0000 -0.7383 -0.1500 0 0 0 0] f_1(-n)=[0 0 0.0150 -0.1888 0.2243 -0.2243 0.2243 -0.2243 0.2243 -0.1354 0.0518 0 0]

From the recursive relation using (8.19) and (8.20), the wavelet function and scaling function for the transient event are designed as shown in Figure 8.7.



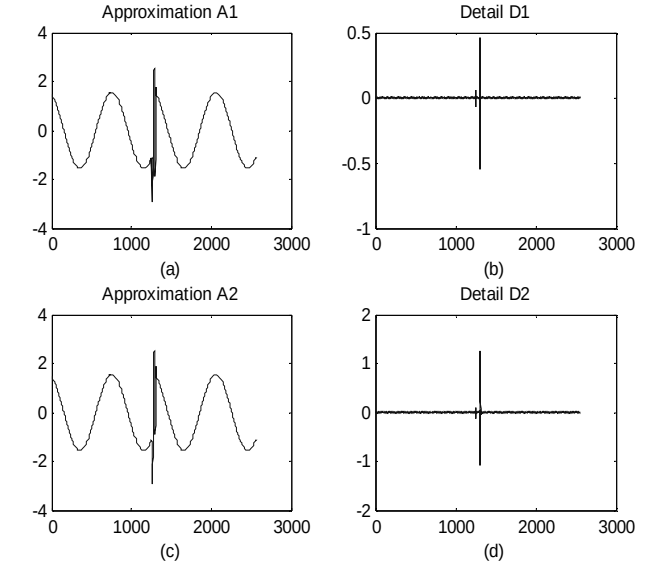
(a)

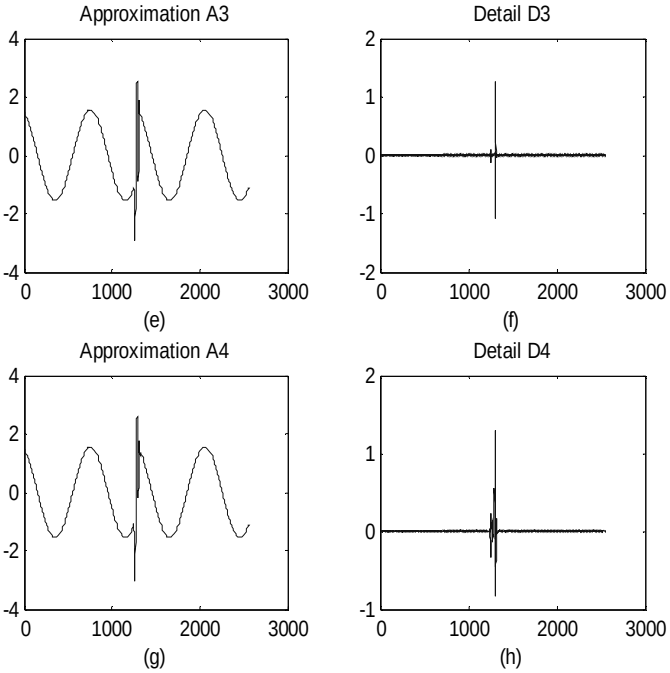


(b)

Figure 8.7 (a) Statistically matched wavelet (b) scaling function

Designed filters are used for application of compression as a result the event is detected. Figure 8.8 shows the six level of decomposition for transient event. It is analysed that statistical matched wavelet provides compression up to six level of decomposition as compared to Daubechies four level. The results are shown for transient event as the objective signal can be any type of signal to be detected.


Approximation A1
Detail D1
(a)
(b)
Approximation A2
Detail D2
(c)
(d)


Approximation A3
Detail D3
(e)
(f)
Approximation A4
Detail D4
(g)
(h)

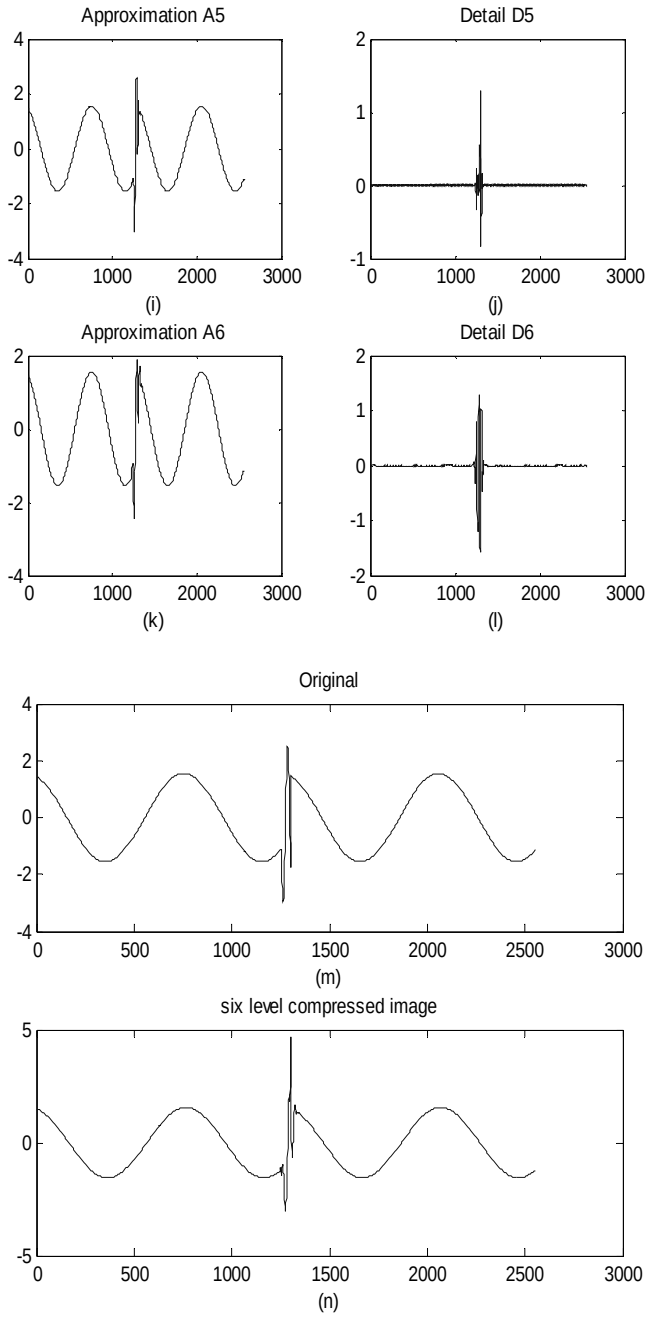


Figure 8.8 Decomposition of transient event using statistically matched wavelet (a)–(b) depicts first level decomposition having low pass information A1 and high pass information D1 (c)–(d) depicts second level decomposition having low pass information A2 and high pass information D2 (e)–(f) depicts third level decomposition having low pass information A3 and high pass information D3 (g)–(h) depicts forth level decomposition having low pass information A4 and high pass information D4 (i)–(j) depicts fifth level decomposition having low pass information A5 and high pass information D5 (k)–(l) depicts sixth level

Table 8.2 SNR of Transient Event

	1st level of decomposition	2nd level of decomposition	3rd level of decomposition	4th level of decomposition	5th level of decomposition	6th level of decomposition
Db1	33.7175	33.1647	30.2881	24.5309	22.7936	21.3397
Db2	38.7732	34.4003	30.1748	25.7375	21.9717	21.3112
Db4	44.3172	36.9045	32.0737	29.4073	23.9872	22.2787
Db6	40.5654	35.6775	31.5412	27.9025	22.4232	21.6418
Matched wavelet	48.3684	43.8273	38.1789	35.6121	33.1050	28.4110

decomposition having low pass information A6 and high pass information D6 (m) original event (n) reconstructed event after six level of decomposition.

The parameter SNR is used in order to measure the quality of proposed method. For the transient event, Table 8.2 shows that the value of SNR is optimal for statistically matched wavelet in compare to Daubechies wavelet for all level of decomposition.

Therefore, in statistically matched wavelet, numbers of samples are reduced by 1/64 as compare to 1/16 in Daubechies wavelet. So detection by statistically matched wavelet takes less time to detect events and work well in real time. The detected events are classified using classifier based on Iterative Closest Point algorithm which is described in next section.

8.1 Iterative Closest Point Algorithm

In power engineering, problems of PQ are not limited only to detection and localization of disturbances. More important is the ability to classify various types of PQ disturbance [107], [109]. This section explains the classification process based on Iterative Closest Point (ICP) algorithm followed to support a particular event.

ICP is an algorithm employed to match two clouds of points. The ICP is designed to fit points in a target signal to points in a control signal. It is important that an initial estimate is made regarding where the overlay of the signals should be. The base component of the algorithm calculates the smallest distance between each point in the target signal to a point in the control signal. These calculated points are then used to form a translation vector that is applied over all points in the target signal to adjust them towards the control signal. This process is repeated numerous times, with the end result being a target signal with points that are within a specified squared error distance of their corresponding points in the control signal. The algorithm is very simple and it iteratively estimates the transformation between two raw scans.

8.1.1 Algorithm: ICP

Step 1: Establish correspondence between pairs of features in the two structures that are to be aligned based on proximity,

Step 2: Estimate the rigid transformation that best maps the first member of the pair onto the second member.

Step 3: Apply that transformation to all features in the first structure.

Step 4: These three steps are then reapplied until convergence is concluded. The ICP algorithm always converges monotonically to a local minimum with respect to the mean square error distance as objective function.

8.2 Results of Classification

The PQ event when tested with ICP algorithm has given the results in the corresponding ranges of the translation values as depicted in Table 8.3. Based on resulting translation vector, the event is recognized.

Table 8.3 Range of translation vector

S No.	Event	Max. value	Min value	Range
1	Transient	0.0979	0.0541	0.0438
2	Sag	0.0079	0.0053	0.0026
3	Swell	0.0199	0.0172	0.0027
4	Harmonic	0.0375	0.0318	0.0057

Table 8.4 Classification of PQ events

	Transient	Sag	Swell	Harmonics
Transient	96	02	01	01
Sag	0	97	03	0
Swell	0	8	92	0
Harmonic	02	04	04	0

In Table 8.4, confusion matrix is showing the performance of ICP algorithm based classifier. It demonstrates the number of times the ICP method could successfully distinguish among the various PQ events. The diagonal of the matrix shows the correct recognition of the PQ events. The ICP works very well with signals of high resolution and gives values of translation vector to be approximately zero in case of matching patterns, or horizontally and vertically shifted patterns. In the case of tilted or unaligned patterns, the method is found to give varied results.

8.3 Conclusions

The effectiveness of designed statistically matched wavelet has been tested for detection of real time generated PQ events. The statistically matched wavelet provides compression up to six level of decomposition as compared to Daubechies four level of decomposition. Thus, number of samples to detect PQ events is reduced using statistically matched wavelet as compared to Daubechies wavelets. The proposed technique takes very less time to detect PQ events and works efficiently in real time. For classification of detected events, ICP algorithm has been used. The results carried out on 100 samples show that proposed classifier based on ICP algorithm gives better results of classification. It classifies events correctly in presence of outlier points and Gaussian noise.

Chapter-9

Morphmap for Non-Linear Dimensionality Reduction

Chapter 9

Morphmap for Non-Linear Dimensionality Reduction

The proposed method is an improvement over the well known non-linear dimensionality reduction techniques such as Isomap and LLE. In the proposed method neighborhood graph construction is topology based instead of a constant epsilon or k nearest neighborhood method. Isomap and LLE methods require a user base input or calculation of k, which is itself a computational burden. In the proposed method, neighborhood graph construction is done by stacking image in third dimension using Morphological mapping (Morphmap). The Cohn-Kanade (CK), Japanese Female facial expression (JAFFE) and Aleix Martinez and Robert Benavente (AR) face database have been used to evaluate the performance of the proposed method. The results of proposed method show the improvement over computationally expensive Geodesic distances method used in Isomap.

9.1 Introduction

Human being has a very intelligent and fast system. For example while moving in a bus, it comes across a very large number of scenes but does not get exhausted. The recognition using computers is fast with reduced dimensions. While studying various dimensionality reduction techniques, it has been noticed that though the techniques are getting faster but the qualitative information content in the low dimensional space is going down. Any dimensionality reduction technique cannot work practically till it is computationally efficient and at the same time the features inherent in the scene are not lost.

Isomap preserves global properties, like the Geodesic distances, approximated as shortest paths on the proximity graph. Semidefinite embedding algorithm maximizes the variance in the data set while keeping the local distances unchanged, thereby approximately preserving Geodesic distances in the manifold [41]. Finally,

the Isomap algorithm considers both local and global invariants [17]. Short Geodesic distances are assumed to be equal to Euclidean distances, and longer ones are approximated as shortest paths length on the proximity graph, using standard graph search methods like Dijkstra's algorithm [88]. Isomap then uses multidimensional scaling, attempting to find an m-dimensional Euclidean representation of the data, such that the Euclidean distances between points are as close as possible to the corresponding Geodesic ones. All such methods are based upon the neighborhood information. Such methods fail when the data is spread over the image having a multiple number of clusters.

An important point regarding the embedding of the points in the lower dimensions is the preservation of the small changes as well as the large changes in the overall manifolds of the objects under observations. For this, the method of computation of minimum residual variance is proposed which requires the calculation of variance for all values of k for selecting optimum value of k. The proposed method gives an alternative to the computationally expensive Geodesic distance approach. This approach also leads to the birth of the concept of group vectors, a cluster of vectors depicting the valid connections of any selected point with neighboring points lying in the next traversed stacked layer.

The paper is organized in the following manner. The details of proposed non-linear dimensionality reduction method using Morphmap are described in Section 9.2. The multiclass SVMs used for classification is presented in Section 9.3. Section 9.4 describes the simulation results for face expression recognition. Finally, the concluding remarks are given in Section 9.5.

9.2 Proposed Morphmap Method

In proposed method, the neighborhood graph construction is topology based instead of k nearest neighbors or constant epsilon (ε) method, both the methods require a user base input. In earlier methods, the calculation of k or ε is a computational burden. They do not take the topology of intensity variation of object in consideration.

In proposed method, the depth information present in the image has been stacked in third dimension for each significant object present in the image. The concept of Morphological mapping has been used for visualizing an image in three dimensions: two spatial coordinates versus gray levels (by stacking image in third dimension). The representation of the depth has been taken into account within small area in these objects by applying the intensity attenuation function used in graphical modeling.

$$F(d) = {}^{1}\!\Big/\!_{(a_0 + a_1 d + a_2 d^2)} \tag{9.1}$$

where a_0, a_1 and a_2 are variants used to depict shading information with respect to light source. Usually in case of large distances a_1 is negligible and in case of small distances a_2 is negligible.

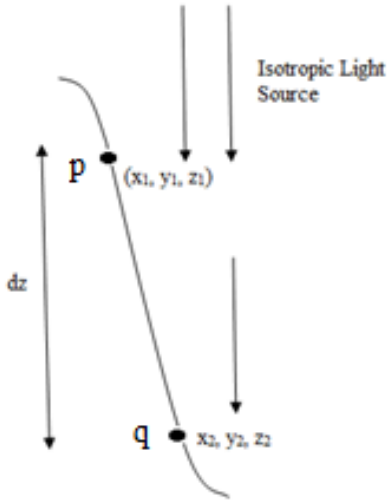


Figure 9.1 Separations of neighborhood pixels

Reciprocating the (9.1),

$$(a_0 + a_1 d + a_2 d^2) = {}^{1}\!\Big/\!_{F(d)} \tag{9.2}$$

Using (9.2) for pixels p and q in Figure 9.1, (9.3) and (9.4) are obtained as given below:

$$(a_0 + a_1 z_1 + a_2 z_2^2) = {}^{1}\!\Big/\!_{F(z_1)} \tag{9.3}$$

$$(a_0 + a_1 z_1 + a_2 z_2^2) = {}^{1}\!/_{F(z_2)} \tag{9.4}$$

Since pixels p and q are extremely close, factor a_2 can be neglected. Thus subtracting (9.3) and (9.4), (9.5) is obtained as follows:

$${}^{1}\!/_{F(z_2)} - {}^{1}\!/_{F(z_1)} = a_1.(z_2 - z_1) \tag{9.5}$$

Addition of the extra factor a_1 in the planar Euclidean distance between two pixels makes the total Euclidean distance as follows:

$$\|p - q\| = ((x_2 - x_1)^2 + (y_2 - y_1)^2 + a_1^2 (z_2 - z_1)^2)^{1/2} \tag{9.6}$$

This leads to the requirement of an algorithm for further connection of points not connected to each other. As can be seen in the surface intensity plot of the face shown in Figure 9.2, regions or layers of the intensity are created based on intensity threshold crossing in each object of interest. For example, in this case the only objects of interest in the image are faces. The intensity threshold is iteratively performed on the faces starting from the global intensity minima to the edges detected by watershed transform. The edges of these layers as traversing up the stack are of common intensity. The applications of the proposed method could be done within group vectors of two adjacent layers of the stack in a fully connected way as shown in Figure 9.2.

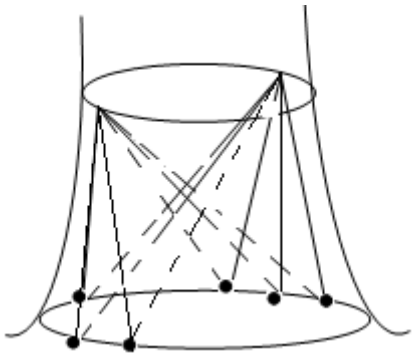


Figure 9.2 Connected edge pixels in the two adjacency layers

The connection of edge pixels in the two adjacency layers are shown in Figure 9.2. The solid lines show the valid vectors in group and the dashed lines denote the erroneous vectors. This is because the dashed lines are attempting to connect points

by traversing through the object. Now consider the two adjacent layers of the object to be consecutive cross-sections of a right circular cone as shown in Figure 9.3.

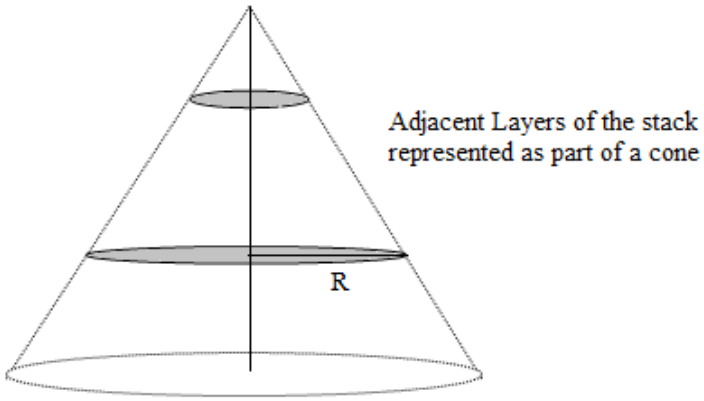


Figure 9.3 Adjacent layers of the stack represented as part of the cone

Radii of any pixel in both regions can be found by two dimensional Euclidean distances between any edge pixel and centroid of region. Now if any edge pixel has radius as R which is the Euclidean distance from the centroid, the angle of divergence is given as follows:

$$d = \left(2.R \big/ n.R\right) = 2 \big/ n \tag{9.7}$$

where n is the number of edge pixels in a circle of radius R. The pixels in the a consecutive layers can be connected using the intensity attenuation function given by (9.1). The pixels in the layer with the larger area which come directly under the angle of divergence of an edge pixel in the smaller layer can form a part of group vector and such an approach leads to benefit of removal of erroneous vectors. Further, the approach tends to optimize the connections between the two layers which lead to further drastic reduction in the computation required for finding the distance matrix. The idea of using group vectors is unique to extract the advantage of properly seeking relation between points of interest based on the concept of layers. Pixels in non adjacent layers are connected using shortest path algorithm. The network of connected pixels is then passed through MDS in distance matrix form. MDS has the capability to

convert high dimensional data present in the distance matrix into a two-dimensional embedding while retaining the distance relations among the points.

9.2.1 Algorithm: Morphmap

Step 1: **Construct neighborhood graph**

(i) The Morphological mapping is used for visualizing an image in three dimensions.

(ii) Number of edge pixels in a circle of radius r connected to consecutive layers using intensity attenuation function given by (9.1).

(iii) The pixels in the layer with the larger area which come under angle of divergence given by (9.7) of an edge pixel in the smaller area can form a part of group vectors with that pixel and other vectors are removed.

Step 2: **Compute shortest paths**

Pixels in non adjacent layers are connected using shortest path algorithm.

Step 3: Construct the low dimensional embedding

The network of connected pixels is then passed through MDS in distance matrix form.

Step 4: Recognition

Recognition is performed by multiclass SVM classifier.

In next section, multiclass SVMs is described to distinguish various facial expressions.

9.3 Classification using Multiclass SVMs

The basic description of support vector machines (SVMs) can be phrased as a two class classification problem where data points are mapped into a high dimensional hyperspace so that they can be separated by a hyper plane [54]. A margin exists on each side of the hyper plane which is distanced to the nearest set of data points of each class. A high margin indicates good separation and good generalization. The data points that sit on the margin are known as support vectors.

For facial expressions classification, the embedded KE vectors $g_j, j=1,...,N$ are used as input to the SVMs system. All the classes are considered for the experiments, each one representing one of the basic facial expressions. The output of the SVMs system is a label that classifies the embedded KE under examination to one of the basic facial expressions.

The training data $(g_1,l_1),...,(g_1,l_N)$ where $g_j \in \Re^L$ are the KE vectors and $l_j \in \{1,...,6\}$ are the facial expression labels of the embedded KE. The training data are the facial expression labels of the embedded KE value. The multiclass SVMs problem solves only one optimization problem [55]. It constructs basic facial expressions rules, where the kth function $\boldsymbol{w}_k^T \phi(g_j) + b_k$ separates training vectors of the class *k* from the rest of the vectors, by minimizing the objective function:

$$\min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi}} \frac{1}{2} \sum_{k=1}^{6} \mathbf{w}_k^T \mathbf{w}_k + C \sum_{j=1}^{N} \sum_{k \neq l_j} \xi_j^k \tag{9.8}$$

Subject to the constraints

$$\begin{aligned} \mathbf{w}_{l_j}^T \phi(g_j) + b_{l_j} &\geq \mathbf{w}_k^T \phi(g_j) + b_k + 2 - \xi_j^k \\ \xi_j^k &\geq o, \quad \text{j=1,...,N,} \quad \text{k}\{1,...,6\}\backslash l_j \end{aligned} \tag{9.9}$$

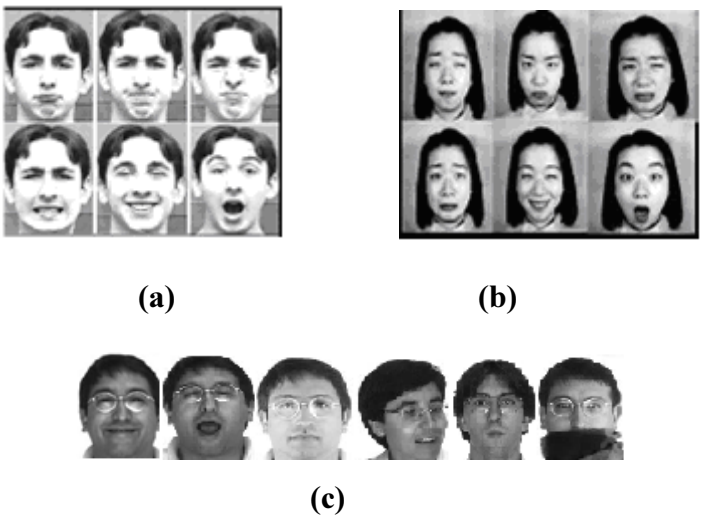
where ϕ is the function that maps the deformation vectors to a higher dimensional space, where the data are supposed to be linearly or near linearly separable. C is the term that penalizes the training errors and **w** is the normal vector to the hyper plane. The vector $\boldsymbol{b} = [b_1 ... b_6]^T$ is the bias vector and $\xi = [\xi_1^1,...,\xi_i^k,...,\xi_N^6]^T$ is the slack variable vector. Then, the decision function is defined as follows:

$$h(\boldsymbol{g}) = \underset{k=1,...,6}{\arg\max} \, (\mathbf{w}_k^T \phi(\boldsymbol{g}) + b_k) \tag{9.10}$$

Using this procedure, a test feature vector is classified to one of the basic facial expressions using (9.10). Once the multiclass SVMs system is trained, it can be used for testing, i.e., for recognizing facial expressions on new facial image sequences.

9.4 Results

The proposed method for FER has been tested on the CK [80-81], JAFFE [82] and AR databases [83]. The CK database contains 97 university students with all six expressions (Happiness, Sad, Surprise, Fear, Anger and Disgust). The JAFFE database contains 213 images of seven facial expressions (six basic expressions and neutral expression also) posed by 10 Japanese female models.



(a) **(b)**

(c)

Figure 9.4 Sample images from (a) CK database (b) JAFFE database (c) AR database

The AR database contains 116 face images, 26 images are available for each person. The AR database is captured with different expressions, illumination conditions and occlusions (scarf and sunglasses). In the AR database, besides the lighting from the left and the right, lighting from both sides of each face is also adopted. Sample Images from Cohn-Kanade, JAFFE and AR databases are displayed in Figures 9.4 (a), (b) and (c) respectively. In all the databases, single image per expression of each subject is selected as training image.

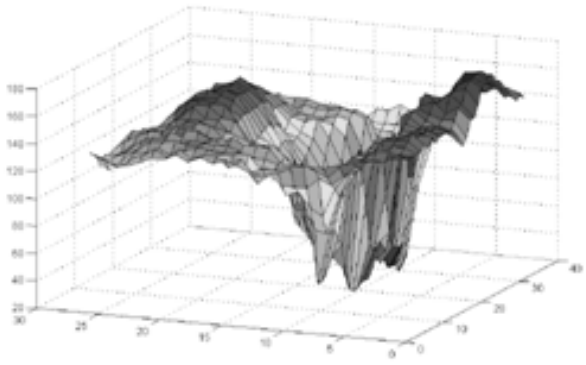
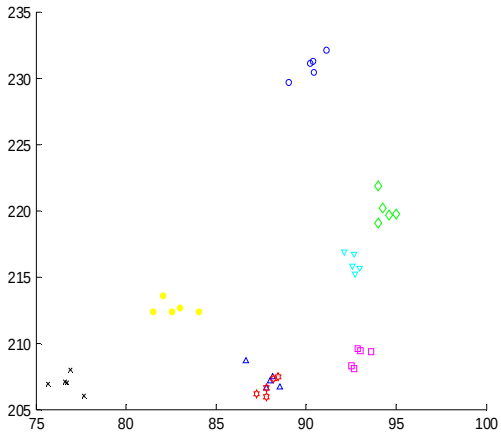


Figure 9.5 Morphological plot of a face

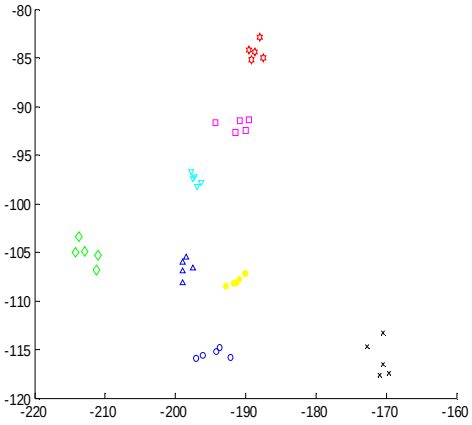
The proposed method takes topology of intensity variation of each object in consideration by selecting each layer on intensity threshold criterion. The morphological plot of a face has been shown in Figure 9.5. The pixels in non adjacent layers are connected using shortest path algorithm. The network of connected pixels is then passed through MDS in distance matrix form. The MDS seeks to preserve the intrinsic geometry of the data as captured in the morphological manifold distance between all pair of data points [42].

The 2-D embedding of CK and AR face images are shown in Figure 9.6 (a, b) respectively, where each point corresponds to a face. From Figures 9.6 (a, b), it is observed that the face images of same person are grouped together and of different persons are separated from each other. At the same time, different face expressions of a person are separated from each other and similar ones grouped together. Similarly, the 2-D embedding of JAFFE face database is shown in Figure 6 (c) where, each point corresponds to a face. From Figure 9.6 (c), it is observed that the different face expressions are separated from each other and similar ones grouped together.

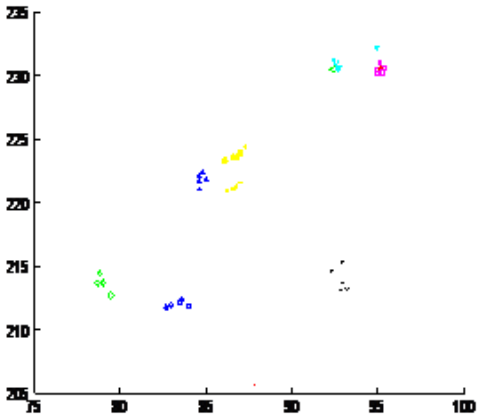
Morphmap highlights the natural clusters of the faces and show separate clusters between dissimilar faces. They make similar face of the same individual lie in the vicinity of the face image space and make dissimilar faces from different individuals appear far away in their reduced embedding spaces.



(a)



(b)



(c)

Figure 9.6 Two-dimensional embedding by Morphmap of (a) CK database (b) AR database (c) JAFFE database

The multiclass SVMs has been used as classifier for computing recognition accuracy. The output of the SVMs system is a label that classifies the one of the basic facial expressions under examination. Using this procedure, a test feature vector is classified to one of the basic facial expressions using (9.10). It is observed that expressions captured under different illumination have also been recognized using proposed method.

Table 9.1 Recognition Accuracy of CK Database

Method	Recognition accuracy (%)
LLE	77.9
Isomap	77.2
Morph map	79.2

Table 9.2 Recognition Accuracy of JAFFE Database

Method	Recognition accuracy (%)
LLE	80.7
Isomap	79.9
Morph map	81.6

Table 9.3 Recognition Accuracy of AR Database

Method	Recognition accuracy (%)
LLE	78.2
Isomap	77.0
Morph map	78.8

The proposed method has been compared with LLE and Isomap. The results of recognition accuracy for CK, JAFFE and AR databases are depicted in Tables 9.1, 9.2 and 9.3 respectively. The recognition accuracy of Morphmap based method is better than LLE and Isomap. The computational complexity of Morphmap is much less as compared with other LLE and Isomap. The comparative results clearly demonstrate that the proposed method works well to discover the non-linearity of the various facial

expressions. The results of facial expression recognition are better than the LLE and Isomap methods and it is also much efficient in terms of computation cost.

9.5 Conclusions

The proposed Morphmap based method has improved the computationally expensive Geodesic distance method used in Isomap. The proposed method may be applied, when non-linear geometry complicates the use of Isomap, LLE or related. The proposed method help in better understanding of how the brain comes to represent the dynamic appearance of objects and suggest a central role for Geodesic transformations on non-linear manifolds.

List of Publications of PHD Work

List of Publications of PhD Work

International Journal

1. Rajiv Kapoor, Rashmi Gupta, "Non-Linear Dimensionality Reduction using Fuzzy Lattices," ***IET Computer Vision***, pp. 1–8, doi: 10.1049/iet-cvi.2012.0097, 2013.
2. Rajiv Kapoor, Rashmi Gupta, "Classification of PQ disturbances using non-linear dimensionality reduction," *Int. Journal of Electrical Engineering,* **Springer**, vol. 95, no. 2, pp. 147-156, 2013.
3. Rashmi Gupta, Rajiv Kapoor, "*Comparison of Graph Based Methods for Non-Linear Dimensionality Reduction,*" Int. J. of Signal and Imaging System Engineering, Special Issue on Feature Extraction and Selection for Images Recognition in Large Databases**, Inderscience**, vol. 5, no.2, pp. 101-109, 2012.
4. Rashmi Gupta, Rajiv Kapoor, "*Morphmap for Non-linear Dimensionality Reduction Technique,"* Int. J. of Pattern Analysis and Applications, **Springer**, Communicated.
5. Rashmi Gupta, Rajiv Kapoor, "Extensions and Analysis of Local Non-Linear Techniques," ***Int. J. of Computer Applications*****,** U.K., vol. 51, no. 13, pp. 1-6, 2011.
6. Rajiv Kapoor, Rashmi Gupta, "Statistically Matched Wavelet Based Method for Detection of PQ Events," I*nternational Journal of Electronics*, **Taylor and Fransis,** vol. 98, no.1, pp.109-127, 2011.
7. Rajiv Kapoor, Rashmi Gupta, "Fuzzy Lattice based Technique for Classification of Power Quality Disturbances,*" Int. Transactions on Electrical Energy Systems*, **Wiley-Blackwell**, vol. 22, no. 8, pp. 1053-1064, 2012.
8. Rashmi Gupta, Rajiv Kapoor, "Constraint Isomap: A Global framework for Non-Linear Dimensionality Reduction," ***Journal of Pattern Recognition and Research*****,** First revision.

9. Rashmi Gupta, Rajiv Kapoor, "Data Compression by Discrete Wavelet Transform using Matched Wavelet," *Int. J. Signal and Imaging Systems Engineering*, **Inderscience**, Accepted.
10. Rashmi Gupta, Pooja Pandey, Rajiv Kapoor, "Spatial Distance Preservation Based Methods for Non-Linear Dimensionality Reduction," ***Int. J. of Computer Applications*****,** U.K., vol. 69, no. 20, 2013.
11. Rashmi Gupta, Rajiv Kapoor "Unsupervised Locally Linear Embedding for Dimension Reduction," *Int. J. of Recent Trends in Engineering and Technology* (Letter), **ACEEE,** Finland, vol. 4, no. 1, 2010.
12. Rashmi Gupta, Rajiv Kapoor, "Implementation of Dimension Reduction Technique for Face Recognition," *Int. Journal of Recent Trends in Engineering,* **ACEEE,** Finland, vol. 3, no.2, pp 144-146, 2010.
13. Rashmi Gupta, Varun Bhardwaj, Rajiv Kapoor, "Signature Recognition using LDA and MFA," *Int. Journal of Signal Processing, Elsevier***,** Communicated**.**
14. Manish Sharma, Rashmi Gupta, Deepak Kumar, Rajiv Kapoor, "Efficacious approach for satellite Image Classification," *J. of Electrical and Electronics Engineering Research*, **Academic journals**, vol. 3, no. 8, pp. 143-150, 2011
15. Rashmi Gupta, Rajiv Kapoor, Akash Bhatia, Nalin Handa, Gurpreet Singh "Classification of Texture Images on Entropy Basis," *International Journal of Advances in Communication Engineering*, IJACE, vol. 1, no.2, pp.61-65, 2009.

International/National Conference

1. Rashmi Gupta, Rajiv Kapoor "Image Compression using 2-Dimensional Transforms" *4th IEEE International Conference on Advanced Computing and Communication Technologies* (ICACCT-2010), Oct. 30, 2010, pp.355-360.
2. Manish Sharma, Rashmi Gupta, Deepak Kumar, Rajiv Kapoor "Fuzzy Logic System for Image Classification" *Fifteenth Annual Conference of Gwalior Academy of Mathematical Sciences (GAMS),* Dec 12-14, 2010, New Delhi.

References

References

[1] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[2] W.S. Torgerson, "Multidimensional scaling I: Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[3] S.T. Roweis, "EM algorithms for PCA and SPCA," *In Advances in Neural Information Processing Systems*, vol. 10, pp. 626–632, The MIT press, 1998.

[4] N.D. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[5] L.I. Smith, "*A tutorial on Principal Components Analysis,*" February 26, 2002.

[6] M.A. Turk and A.P. Pentland, "Face recognition using eigenfaces," *In Proc. of the Computer Vision and Pattern Recognition*, pp. 586–591, 1991.

[7] P.N. Belhumeur, J.P. Hepanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE. Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[8] M. Welling, "Fisher Linear Discriminant Analysis," *Department of Computer Science,* University of Oronto, Toronto, M5S 3G5 Canada.

[9] A. Bansal, K. Mehta, and S. Arora, "Face Recognition using PCA and LDA Algorithms," *Second Int. Conf. on Advanced Computing and Communication Technologies*, 2012.

[10] C.J.C. Burges, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, chapter Geometric Methods for Feature Selection and Dimensional Reduction: A Guided Tour, Kluwer: Academic Publishers, 2005.

[11] L.K. Saul, K.Q. Weinberger, J.H., Ham, F. Sha, and D.D Lee, "Spectral methods for dimensionality reduction," *In Semisupervised Learning,* Cambridge, MA: The MIT Press, 2006.

[12] G. Baudat and F. Anouar, “Generalized discriminant analysis using a kernel approach,” *Neural Computation*, vol. 12, pp. 2385-2404, 2000.

[13] X. He and P. Niyogi, “Locality Preserving Projections,” *Advances in Neural Information Processing Systems16*, Vancouver, British Columbia, Canada, 2003.

[14] J.B. Tenenbaum, V. de Silva, and J.C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[15] M. Balasubramanian and E.L. Schwartz, “The Isomap algorithm and topological stability,” *Science,* vol. 295, no. 5552, pp.7-7, 2002.

[16] J.A. Lee and M. Verleysen, “Nonlinear dimensionality reduction of data manifolds with essential loops,” *Neurocomputing*, vol. 67, pp. 29–53, 2005.

[17] H. Choi and S. Choi, “Robust kernel Isomap,” *Pattern Recognition*, vol. 40, no. 3, pp. 853–862, 2007.

[18] B. Sch¨olkopf, A.J. Smola, and K.-R. M¨uller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[19] J. Shawe-Taylor and N. Christianini, “Kernel Methods for Pattern Analysis,” *Cambridge University Press*, Cambridge, UK, 2004.

[20] K.I. Kim, K. Jung, and H.J. Kim, “Face recognition using kernel principal component analysis,” *IEEE Signal Processing Letters*, vol. 9, no. 2, pp. 40–42, 2002.

[21] K.Q. Weinberger, F. Sha, and L.K. Saul, “Learning a kernel matrix for nonlinear dimensionality reduction,” *In Proc. of the 21 International Conference on Machine Learning*, 2004.

[22] S.T. Roweis and L.K. Saul, “Nonlinear dimensionality reduction by Locally Linear Embedding,” *Science,* vol. 290, no. 5500, pp. 2323–2326, 2000.

[23] E. Kokiopoulou and Y. Saad, “Orthogonal Neighborhood Preserving Projections: A projection-based dimensionality reduction technique,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 29, no. 12, pp. 2134–2156, 2007.

[24] X. He, D. Cai, S. Yan, and H.-J. Zhang, “Neighborhood preserving

embedding," *In Proc. of the 10th IEEE Int. Conf. on Computer Vision*, pp. 1208–1213, 2005.

[25] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik, "Dimensionality reduction: A comparative review," *MICC,* Maastricht University, Netherlands, 2008.

[26] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and spectral techniques for embedding and clustering," *In Advances in Neural Information Processing Systems*, vol. 14, pp. 585–591, Cambridge, MA, USA, 2002.

[27] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," *In Proceedings of the 10th IEEE Int. Conf. On Computer Vision*, pp. 1208–1213, 2005.

[28] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Machine Learning*, vol. 56, no. 3, pp.209–239, 2004.

[29] X. He and P. Niyogi, "Locality preserving projections," *In Advances in Neural Information Processing Systems16*, 2003.

[30] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[31] D.L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," *Proc. of the National Academy of Sciences*, vol. 102, no. 21, pp. 7426–7431, 2005.

[32] N. Patwari and A.O. Hero, "Manifold learning algorithms for localization in wireless sensor networks," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 857–860, 2004.

[33] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via Local Tangent Space Alignment," *SIAM Journal of Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[34] T. Zhang, J. Yang, D. Zhao, and X. Ge, "Linear local tangent space alignment and application to face recognition," *Neurocomputing*, vol. 70, pp.1547–1533, 2007.

[35] X. Geng, D. Zhan, and Z. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Trans. Syst., Man, Cybern., Part-B, Cybern.,* vol. 35, no. 6, pp. 1098–1107, 2005.

[36] D. Ridder, O. Kouropteva, O. Okun, M. Pietikäinen, and R. P. W. Duin, "Supervised locally linear embedding," *in Proc. Int. Conf. Artif. Neural Netw.,* Istanbul, Turkey, pp. 333–341, 2003.

[37] D. Sun and D. Q. Zhang, "Bagging constraint score for feature selection with pairwise constraints," *Pattern Recognit.,* vol. 43, no. 6, pp. 2106–2118, 2010.

[38] F. Wang, "Semi-supervised metric learning by maximizing constraint margin," *IEEE Trans. Syst., Man, Cybern., Part-B, Cybern.,* vol. 41, no. 4, pp. 931–939, 2011.

[39] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Trans. Neural Network*, vol. 20, no. 4, pp. 729–735, 2009.

[40] Y. Bengio, J. Paiement, P.Vincent, O. Dellallaeu, N.L Roux, and M. Quimet, "Out-of sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering," *Neural Information Processing Systems*, 2003.

[41] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[42] M. D. Lee, "Determining the Dimensionality of Multidimensional Scaling Models for Cognitive Modeling," *Journal Mathematical Psychology*, vol. 45, no. 1, pp. 149-166, 2001.

[43] M. W. Huang, Z. W Wang, and Z. L.A. Ying, "Novel Method of Facial Expression Recognition based on GPLVM Plus SVM," *Proc. Int. Conf. Signal process.,* Beijing, pp. 916-919, 2010.

[44] D. Cai, X. F. He, J. W. Han, and H. J. Zhang, "Orthogonal Laplacian Faces for Face Recognition," *IEEE Trans. Image Process*., vol. 15, no. 11, pp. 3608–3614, 2006.

[45] S. S. Dacheng Tao and K.P. Chan, "Evolutionary Cross-Domain Discriminative Hessian Eigenmaps," *IEEE Trans. Image Process.,* vol. 19, no. 4, pp.1075-1086, 2010.

[46] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” *IEEE Int. Conf. Automatic Face and Gesture Recognition*, Nara, Japan, pp. 200–205, 1998.

[47] X. Feng, M. Pietik¨ainen, and A. Hadid, “Facial expression recognition with local binary patterns and linear programming,” *Pattern Recognition and Image Analysis*, vol. 15, no. 2, pp. 546–548, 2005.

[48] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[49] Y. R. Yeh, S.Y. Huang, and Y. J. Lee, “Nonlinear Dimension Reduction with Kernel Sliced Inverse Regression,” *IEEE Trans. knowledge and data engineering*, vol. 21, no. 11, pp. 1590-1603, 2009.

[50] P. Khurd, and C. Davatzikos, “On Analyzing Diffusion Tensor Images by Identifying Manifold Structure using Isomaps*,” IEEE Trans. Medical Imaging*, vol. 26, no. 6, pp. 772-778, 2007.

[51] Li. He, J. M. Buenaposada, and L. Baumela, “An empirical comparison of graph-based dimensionality reduction algorithms on facial expressions recognition tasks,” *Int. Conf. Pattern Recognition,* Tampa, Florida, pp. 1-4, 2008.

[52] Y. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 23, no. 2, pp. 97–115, 2001.

[53] I. Buciu, C. Kotropoulos, and I. Pitas, “ICA and Gabor representation for facial expression recognition,” *IEEE Int. Conf. Image Process.,* vol. 2, no. 3, pp. 855-858, 2003.

[54] I. Kotsia, and I. Pitas, “Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines,” *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172-187, 2007.

[55] C. W. Hsu and C. J. Lin, “A Comparison of Methods for Multiclass Support Vector Machines,” *IEEE Trans. on Neural Networks*, vol. 13, no. 2, pp. 415-425, 2002.

[56] B. Allen and L.R. Rabiner, “A unified approach to short time Fourier analysis and synthesis,” *IEEE Proc.*, vol. 65, no. 11, pp. 1558-1564, 1977.

[57] R. Altes, “Detection estimation and classification with spectrograms,” *Journal of the Acoustical Society of America*, vol. 67, no. 4, pp. 1232-1246, 1980.

[58] Y. H. Gu and M. H .J. Bollen, “Time-frequency and time-scale domain analysis of voltage disturbances,” *IEEE Trans. on Power Delivery*, vol. 15, no. 4, pp. 1279-1284, 2000.

[59] B. Perunicic, M. Mallini, Z. Wang, Y. Liu, and G.T. Heydt, “Power quality disturbance detection and classification using wavelets and artificial neural networks,” in *Proc. 8th Int. Conf. Harmonics and Quality of Power*, pp. 77–82, 1998.

[60] S. Santoso, W. M. Grady, E .J. Powers, J. Lamoree, and S. C. Bhatt, “Characterization of disturbance power quality event with Fourier and wavelet transforms,” *IEEE Trans. on Power Delivery,* vol. 15, no. 1, pp.247-254, 2000.

[61] W. R. Anis Ibrahim and M. M. Morcos, “Artificial intelligence and advanced mathematical tools for power quality applications: a survey,” *IEEE Trans. on Power Delivery*, vol. 17, no.2, pp. 668-673, 2002.

[62] T. X. Zhu, S. K. Tso, and K. L. Lo, “Wavelet-based fuzzy reasoning approach to power-quality disturbance recognition,” *IEEE Trans. on Power Delivery*, vol. 19, no. 4, pp. 1928-1935, 2004.

[63] C. H. Lin and C. H. Wang, “Adaptive Wavelet Networks for Power Quality Detection and Discrimination in a Power System,*” IEEE Trans. on Power Delivery,* vol. 21, no.3, pp.1106-1113, 2006.

[64] S. Shukla, S. Mishra, and B. Singh, “Empirical-Mode Decomposition With Hilbert Transform for Power-Quality Assessment,” *IEEE Trans. on Power Del.,* vol. 24, no. 4, pp. 2159-2165, 2009.

[65] H. He and J. A. Starzyk, “A Self-Organizing Learning Array System for Power Quality Classification Based on Wavelet Transform,” *IEEE Trans. on Power Delivery*, vol. 21, no.1, pp. 286-295, 2006.

[66] M. B. I. Reaz, F. Choong, M. S. Sulaiman, F. M. Yasin, and M. Kamada “Expert System for Power Quality Disturbance Classifier,” *IEEE Trans. on Power Delivery,* vol. 22, no. 3, pp. 1979-1999, 2007.

[67] A. M. Gargoom, N. Ertugrul, and W. L. Soong, "Automatic Classification and Characterization of Power Quality Events," *IEEE Trans. on Power Delivery*, vol. 23, no. 4, pp. 2417-2425, 2008.

[68] U. D. Dwivedi and S. N. Singh, "Enhanced Detection of Power-Quality Events Using Intra and Interscale Dependencies of Wavelet Coefficients," *IEEE Trans. on Power Del.,* vol. 25, no. 1, pp. 358–366, 2010.

[69] P. K. Dash, B. K. Panigrahi, and G. Panda, "Power quality analysis using S-transform," *IEEE Trans. on Power Delivery*, vol. 18, no. 2, pp. 406–411, 2004.

[70] M. V. Chilukuri and P. K. Dash, "Multi-resolution S-transform-based fuzzy recognition system for power quality events," *IEEE Trans. on Power Delivery*, vol. 19, no. 1, pp. 323 – 330, 2004.

[71] S. Mishra, C. N. Bhendeand, and B. K. Panigrahi, "Detection and Classification of Power Quality Disturbances Using S-Transform and Probabilistic Neural Network," *IEEE Trans. on Power Del.*, vol. 23, no. 1, pp. 280-287, 2008.

[72] F. Zhao and R. Yang, "Power Quality Disturbance Recognition Using S-Transform," *IEEE Trans. on Power Delivery*, vol. 22, no. 2, pp. 944-950, 2007.

[73] W.N. Anderson and T.D. Morley, "Eigenvalues of the Laplacian of a graph," *Linear and Multilinear Algebra*, vol. 18, pp. 141–145, 1985.

[74] R.W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, pp. 345, 1962.

[75] A. Saxena, A. Gupta, and A. Mukerjee, "Non-linear dimensionality reduction by locally linear Isomaps," *Lecture Notes in Computer Science*, vol. 3316, pp. 1038–1043, 2004.

[76] J. Yang, D. Zhang, A. Frangi, and J. Yang, "Two-dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 26, no.1, pp. 131-137, 2004.

[77] H. Kong, E. Teoh, J. Wang, and R. Venkateswarlu, "Two Dimensional Fisher Discriminant Analysis: Forget about Small Sample Size Problem," *Proc. IEEE Conf. Acoust. Speech Signal Process.*, pp. 761-764, 2005.

[78] E. A. Edmonds, “Lattice Fuzzy Logics,” *Int. J. of Man-Machine Studies*, vol. 13, no. 4, pp. 455-465, 1980.

[79] V. Petridis and V.G. Kaburlasos, “Learning in the Framework of Fuzzy Lattices,” *IEEE Trans. Fuzzy Systems*, vol. 7, no. 4, pp. 422-440, 1999.

[80] T. Kanade, J. Cohn, and Y. Tian, “Comprehensive Database for Facial Expression Analysis,” *Proc. IEEE Int. Conf. Face and Gesture Recognition*, pp. 46-53, 2000.

[81] http://vasc.ri.cmu.edu//idb/html/face/facial_expression/index.html

[82] http://www.kasrl.org/jaffe.html

[83] A.M. Martinez and R. Benavente, The AR Face Database, CVC Technical Report #24, 1998.

[84] A. Dhall, R. Goecke, S. Lucey, and T Gedeon, “Static Facial Expression Analysis in Tough Conditions: Data, Evaluation Protocol and Benchmark,” *IEEE Int. Conf. Computer Vision*, Workshop BEFIT, Barcelona, Spain, pp. 6-13, 2011.

[85] X. Zhu and D. Ramanan, “Face detection, pose estimation and landmark localization in the wild,” *Computer Vision and Pattern Recognition*, Providence, Rhode Island, pp.1-8, 2012.

[86] X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang, “Face recognition using laplacianfaces,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[87] S. Lafon, and A.B. Lee, “Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 28, no. 9, pp. 1393–1403, 2006.

[88] E.W. Dijkstra, “A note on two problems in connection with graphs,” *Numerische Mathematik, 1*, pp. 269–271, 1959.

[89] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern Classification,” 2nd ed., New Jersey, USA, *Wiley, Interscience,* 2000.

[90] W.Y. Yang, S.X. Liu, T.S. Jin, and X. M. Xu, “An Optimization Criterion for Generalized Marginal Fisher Analysis on Under sampled Problems,” *Int. Journal of Automation and Computing,* vol. 8, no. 2, pp. 193-200, 2011.

[91] D. Kalenova, “Personal Authentication Using Signature Recognition,” *Department of Information Technology, Laboratory of Information Processing, Lappeenranta* University of Technology, 2003.

[92] J.W. Sammon, “A nonlinear mapping algorithm for data structure analysis,” *IEEE Trans. on Computers*, vol. C-18, no. 5, pp. 401-409, 1969.

[93] J. Sun, M.Crowe, C. Fyfe, “Incorporating visualization quality measures to curvilinear component analysis,” *Information Sciences,* vol. 223, pp. 75-101, 2013.

[94] R.N. Shepard, “The analysis of proximities: Multidimensional scaling with an unknown distance function,” *Psychometrika*, vol. 27, no. 2, pp. 125-140, 1962.

[95] P. Dermatines and J. Herault, “Curvilinear component analysis. A self-organizing neutral network for nonlinear mapping of data sets,” *IEEE Trans. on Neutral Networks,* vol. 8, no. 1, pp. 148-154, 1997.

[96] The Olivetti and Oracle Research Laboratory Face Database of Faces. [Online]. Available: http://www.cam-orl.co.uk/facedatabase.html

[97] The Brendan’s Face Database of Faces. [Online]. Available: http://www.cs.nyu.edu/~roweis/data.html

[98] A.M. Martinez and R. Benavente, The AR Face Database, CVC Technical Report #24, June 1998.

[99] A.K. Jain, R.P.W. Duin, and J. Mao, “Statistical Pattern Recognition: A Review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, 2000.

[100] G T. Heydt, Electric Power Quality Stars in a Circle, 1994.

[101] Ö. N. Gerek and D. G. Ece, “Power Quality Event Analysis Using Higher Order Cumulants and Quadratic Classifiers,” *IEEE Trans. on Power Delivery,* vol. 21, no. 2, pp. 883-889, 2006.

[102] P. Janik and T. Lobos “Automated Classification of Power-Quality Disturbances Using SVM and RBF Networks,” *IEEE Trans. on Power Delivery,* vol. 21, no. 3, pp. 1663-1669, 2006.

[103] Ö. N. Gerek, D. G. Ece, and A. Barkana, “Covariance Analysis of Voltage Waveform Signature for Power-Quality Event Classification,” *IEEE Trans. on Power Delivery,* vol. 21, no. 4, pp. 2022-2031, 2006.

[104] P. G. V. Axelberg, I. Y. H. Gu, and M. H. J. Bollen, "Support Vector Machine for Classification of Voltage Disturbances," *IEEE Trans. on Power Delivery,* vol. 22, no. 3, pp. 1297-1303, 2007.

[105] IEEE Std. 1159, IEEE Recommended Practice for Monitoring Electric Power Quality, *IEEE Inc*. New York, pp. 1-59, 1995.

[106] I. Daubechies, *Ten Lectures on wavelet*, Philadelphia, PA: SIAM, 1992.

[107] S. Dahiya, A. Kumar, R. Kapoor, and M. Kumar, "Detection and classification of power quality events using multiwavelets," *International Journal of Energy Technology and Policy (IJETP),* vol. 5 no. 6, pp. 673-683, 2007.

[108] N. P. Subramaniam, B. Bagan, and K. Bagan, "Analysis of High Impedence Transients and Improved Data Compression Using Wavelet Transform," *Serbian Journal of Electrical Engineering,* vol. 3, no. 1, pp. 19-31, 2006.

[109] C. Sharmeela, M. R. Mohan, G. Uma, and J. Baskaran, "A Novel Detection and Classification Algorithm for Power Quality Disturbances using Wavelets," *American Journal of Applied Sciences*, vol. 3, no. 10, pp. 2049-2053, 2006.

[110] A. Gupta, S. D. Joshi, and S. Prasad, "A New Approach for Estimation of statistically Matched Wavelet," *IEEE Trans. on* Signal Proc., vol. 53, no. 5, pp. 1778-1793, 2005.

[111] M. V. Ribeiro, C. A. Duque, and J. M. T. Romano, "An improved method for signal processing and compression in power quality evaluation," *IEEE Trans. on Power Del.*, vol. 19, no. 2, pp. 464–471, 2004.

[112] P. L., Shui, Z. Bao, and X. Zhang, "M-band compactly supported orthogonal symmetric interpolating scaling functions," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1704–1713, 2001.

[113] E. Perrin, R. Harba, C. B. Joseph, I. Iribarren, and A. Bonami, "nthorder fractional Brownian motion and fractional Gaussian noises," *IEEE Trans. Signal Proc.*, vol. 49, no. 5, pp. 1049–1059, 2001.

[114] J. Arrillaga, M. Bollen, and N. R. Watson, "Power quality following deregulation," *Proc. IEEE*, vol. 88, no. 2, 246–261, 2000.

[115] Tewfik., D. Sinha, and P. Jorgensen, "On the optimal choice of a wavelet for signal representation," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 747–765, 1992.

[116] T. Lundahl, W. J. Ohley, S. M., Kay, and R. Siffert, “Fractional Brownian motion: A maximum likelihood estimator and its application to image texture,” *IEEE Trans. Med. Imag.*, MI-5, pp. 152–161, 1986.

[117] W.N. Anderson and T.D. Morley, “Eigenvalues of the Laplacian of a graph,” *Linear and Multilinear Algebra*, vol. 18, 1985, pp. 141–145.

Designing a Global Framework for Non-Linear Dimensionality Reduction

Thesis Submitted in the fulfillment of the requirements for the award of the degree of

Doctor of Philosophy
in
Electronics and Communications Engineering

By
Rashmi Gupta

Under the Supervision of
Prof. Rajiv Kapoor

Faculty of Technology
Delhi College of Engineering
University of Delhi

September 2013

Chapter-10

Conclusions and Scope of Future Research

Chapter 10

Conclusions and Scope of Future Research

In this chapter, the main results from the different chapters have been clubbed together with general conclusions. Finally, some exploratory extensions of this work are proposed.

10.1 Conclusions

In this thesis, firstly linear dimensionality reduction techniques have been discussed and an improved MFA is introduced which is performing significantly better than LDA for classification of data. The LDA works on the assumption that the data of each class has a Gaussian distribution while MFA is applicable for all types of data. Moreover, the MFA with suitable threshold value has improved the recognition accuracy and detection of forged signatures.

Secondly, the stress function is improved for efficiently handle non-linear manifolds. Sammon's NLM can efficiently handle non-linear manifolds, at least if they are not too heavily doped. As a main drawback, Sammon's NLM lacks the ability to generalize the mapping to new points. Another shortcoming of NLM is its optimization procedure, which may be slow or inefficient for some data sets. CCA is much more flexible because the user can choose and parameterize the weighting function. From a computational point of view, the optimization procedure of CCA works much better than the quasi-Newton rule of Sammon's NLM. On the other hand, the interpretation of CCA error criteria is difficult, since weighing function is changing when CCA is running.

The existing non-linear methods are efficient at visualizing artificial data sets and powerful to handle non-linear data. However, they fail to identity inter or intraclass types of neighborhoods and unable to handle discriminatory information. To address these issues, constraint Isomap is proposed that provides geometrical as well as discriminatory information of data and powerful for handling multiple-class real problems. Clustering results obtained by constraint-Isomap are better than existing

non-linear dimensionality reduction methods. From the experimental results, it is observed that constraint Isomap is delivering clear separation on the manifold embedding of multiple classes.

A novel framework for non-linear dimensionality reduction has been proposed to extract prominent features of a person using fuzzy lattices. The developed method directly accommodates the local non-linear behaviour of any object. The message being conveyed or expression being expressed by a person can be known efficiently by proposed method. The experimental results show that the recognition accuracy of proposed method is better than PCA, LDA, Isomap, LLE, GVLVM, LBP and Gabor feature based methods. It is showing promising results on data captured in lab controlled conditions as well as real world like environment. It also has advantage of less numbers of dimensions of feature space. The method can also be applied for lips recognition, hand writing recognition and image tracking. Compared to earlier framework for analyzing high dimensional data that lie on or near a low dimensional manifold, the proposed method has interesting property of representing any object with small set of features.

The fuzzy lattice based technique has also been tested to distinguish various types of PQ events. It is demonstrated that the algorithm can efficiently distinguish the PQ events based on variation of embedded KE value. The value of embedded KE shows a substantial change whenever there is any change in PQ event. As a result, the fuzzy lattice based algorithm can efficiently distinguish the real time generated PQ events in a single cycle. While the earlier techniques based upon wavelets, distinguishes the PQ events in the few cycles of the power signal.

The wavelet statistically matched to the signal has been developed and applied for application of data compression. The effectiveness of designed statistically matched wavelet has been tested for detection of real time generated PQ events. The statistically matched wavelet provides compression up to six level of decomposition as compared to Daubechies four level of decomposition. Thus, number of samples to detect PQ events is reduced using statistically matched wavelet as compared to Daubechies wavelets. The proposed technique takes very less time to detect PQ events and works efficiently in real time. For classification of detected events, ICP algorithm has been used. The results carried out on 100 samples show that proposed

classifier based on ICP algorithm shows better results of classification. It classifies events correctly in presence of outlier points and Gaussian noise.

Another proposed Morphmap based method has improved the computationally expensive Geodesic distance method used in Isomap. The proposed method may be applied when non-linear geometry complicates the use of Isomap, LLE or related. The proposed method help in better understanding of how the brain comes to represent the dynamic appearance of objects and suggest a central role for Geodesic transformations on non-linear manifolds.

10.2 Scope of Future Research

Nonetheless, there remain many issues to explore, and I hope this work inspires further research. In future, development of non-linear dimensionality reduction techniques can be carried out that do not suffer from the presence of trivial optimal solutions and do not rely on hood graphs to model the local structure of the data manifold. The other concern in the development of novel techniques for dimensionality reduction is their optimization, which should be computationally and numerically feasible in practice thus allowing it to be applied to a wider section of datasets.

However, many of the commonly used non-linear dimensionality reduction techniques, such as Locally Linear Embedding or Laplacian Eigenmaps, do not produce conformal maps. Post-processing techniques formulated as instances of semi-definite programming problems can be applied to the output of either Locally Linear Embedding or Laplacian eigenmaps to produce a conformal map. However, the effectiveness of this approach is limited by the computational complexity of SDP solvers. In future, an alternative post processing algorithm can be developed that produces a conformal map but does not require a solution to SDP problem and thus, more computationally efficient.

.