Mississippi State University
Scholars Junction

Theses and Dissertations | Theses and Dissertations

12-13-2008

Multi-Classifiers And Decision Fusion For Robust Statistical Pattern Recognition With Applications To Hyperspectral Classification

Saurabh Prasad

Follow this and additional works at: https://scholarsjunction.msstate.edu/td

Recommended Citation

Prasad, Saurabh, "Multi-Classifiers And Decision Fusion For Robust Statistical Pattern Recognition With Applications To Hyperspectral Classification" (2008). *Theses and Dissertations*. 3221.
https://scholarsjunction.msstate.edu/td/3221

This Dissertation - Open Access is brought to you for free and open access by the Theses and Dissertations at Scholars Junction. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholars Junction. For more information, please contact scholcomm@msstate.libanswers.com.

MULTI-CLASSIFIERS AND DECISION FUSION FOR ROBUST STATISTICAL PATTERN RECOGNITION WITH APPLICATIONS TO HYPERSPECTRAL CLASSIFICATION

By

Saurabh Prasad

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Electrical Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

December 2008

Copyright by

Saurabh Prasad

2008

MULTI-CLASSIFIERS AND DECISION FUSION FOR ROBUST STATISTICAL PATTERN RECOGNITION WITH APPLICATIONS TO HYPERSPECTRAL CLASSIFICATION

By

Saurabh Prasad

Approved:

Lori M. Bruce Professor of Electrical and Computer Engineering (Director of Dissertation)	Nicolas H. Younan Professor of Electrical and Computer Engineering (Committee Member)
James E. Fowler Professor of Electrical and Engineering Graduate Program Director, Electrical and Computer Department (Committee Member)	Jenny Q. Du Associate Professor of Computer Electrical and Computer Engineering (Committee Member)

Sarah A. Rajala
Dean of the Bagley College of Engineering

Name: Saurabh Prasad

Date of Degree: December 12, 2008

Institution: Mississippi State University

Major Field: Electrical Engineering

Major Professor: Dr. Lori M. Bruce

Title of Study: MULTI-CLASSIFIERS AND DECISION FUSION FOR ROBUST STATISTICAL PATTERN RECOGNITION WITH APPLICATIONS TO HYPERSPECTRAL CLASSIFICATION

Pages in Study: 131

Candidate for Degree of Doctor of Philosophy

In this dissertation, a multi-classifier, decision fusion framework is proposed for robust classification of high dimensional data in small-sample-size conditions. Such datasets present two key challenges. (1) The high dimensional feature spaces compromise the classifiers' generalization ability in that the classifier tends to over-fit decision boundaries to the training data. This phenomenon is commonly known as the Hughes phenomenon in the pattern classification community. (2) The small-sample-size of the training data results in ill-conditioned estimates of its statistics. Most classifiers rely on accurate estimation of these statistics for modeling training data and labeling test data, and hence ill-conditioned statistical estimates result in poorer classification performance.

This dissertation tests the efficacy of the proposed algorithms to classify primarily remotely sensed hyperspectral data and secondarily diagnostic digital mammograms, since these applications naturally result in very high dimensional feature spaces and often do not have sufficiently large training datasets to support the dimensionality of the

feature space. Conventional approaches, such as Stepwise LDA (S-LDA) are sub-optimal, in that they utilize a small subset of the rich spectral information provided by hyperspectral data for classification. In contrast, the approach proposed in this dissertation utilizes the entire high dimensional feature space for classification by identifying a suitable partition of this space, employing a bank-of-classifiers to perform "local" classification over this partition, and then merging these local decisions using an appropriate decision fusion mechanism. Adaptive classifier weight assignment and nonlinear pre-processing (in kernel induced spaces) are also proposed within this framework to improve its robustness over a wide range of fidelity conditions. Experimental results demonstrate that the proposed framework results in significant improvements in classification accuracies (as high as a 12% increase) over conventional approaches.

DEDICATION

I dedicate this dissertation to my parents, Dr. Surendra and Dr. Usha Prasad, and my Brother, Sumedh Prasad.

ACKNOWLEDGEMENTS

It was a great pleasure and honor for me to work as a graduate research assistant at GeoSystems Research Institute, alongside so many dedicated colleagues and mentors. I am grateful to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever. First and foremost, I thank my dissertation advisor, Dr. Lori Bruce (Associate Dean at the Bagley College of Engineering, Mississippi State University) for her generous time and commitment, and for providing valuable guidance and encouragement over the course of my doctoral program. I thank my dissertation committee members, Dr. James Fowler, Dr. Nicolas Younan, and Dr. Jenny Du for providing valuable insight and guidance. I thank Dr. Lori Bruce and Dr. David Shaw (Director of the GeoSystems Research Institute at Mississippi State University) for ensuring continual financial and technological support during my doctoral program.

I am also grateful to Trey Breckenridge and his team of system and network administrators at the High Performance Computation Collaboratory at Mississippi State University for providing state-of-the-art computational resources and prompt assistance with cluster computing and other software and hardware issues. It would have taken me a considerably longer time to conduct experiments if I did not have access to these facilities and the support from his team.

I also respectfully acknowledge the following individuals: Dr. C.H. Koger, Dr. Brien Henry, Dr. Abhinav Mathur and Louis Wasson for experimental hyperspectral data collection; Dr. John Ball for providing mammographic feature datasets; Paul Evangelista for providing Tamarisk presence and absence ground truth points in Colorado and Jeff Morisette for providing Hyperion imagery.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

Jain *et al* [1] define statistical pattern recognition as "The study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns." Supervised pattern classification entails the use of labeled training data for learning appropriate class conditional statistical models, which are later employed for making labeling decisions about unlabeled test data. In scenarios where labeled training data is not available, unsupervised classification is employed to identify clusters and patterns and assign them to unknown classes [1], [2]. Assigning class labels to patterns is a task employed in a wide variety of fields. Some examples include the use of such systems for fingerprint recognition [1], speech recognition [1], [3], speaker identification [4], Automatic Target Recognition (ATR) in images collected by remote sensing modalities and automatic detection of breast cancer by analyzing mammograms [5].

In the context of remote sensing applications, ATR systems employ statistical pattern recognition paradigms for identifying targets in images using spatial and spectral information. Hyperspectral target recognition uses the rich information available in spectral signatures of target and background pixels for identifying targets in an image.

Most hyperspectral sensors capture reflectance information at every pixel over hundreds of spectral bands. This results in a very high dimensional pattern recognition problem, thereby requiring an overwhelming amount of training data (ground truth) for accurate representation of class conditional distributions. Further, with an increase in dimensionality of a pattern recognition system, the generalization capacity of the recognizer decreases, thereby resulting in poorer recognition performance. This is well known as the Hughe's phenomenon in the remote sensing and pattern classification community. Conventional hyperspectral ATR systems project the high dimensional reflectance signatures onto a lower dimensional subspace using techniques such as Principal Components Analysis (PCA), Fisher's Linear Discriminant Analysis (LDA) and stepwise LDA etc., and then employ a single classifier for labeling tasks [6]. Although these dimensionality reduction schemes are successful in reducing the ground truth requirement for unbiased modeling by the classifier [6], [7], [8], these projections are not necessarily optimal from a pattern classification perspective [9]. For example, a PCA projection may discard useful discrimination information if it were oriented along directions of small global variance, an LDA projection will be inaccurate for multimodal class distributions, etc. Another factor that governs the efficacy of such dimensionality reduction techniques is the amount of training data required to learn the projections. For example, if the amount of training pixels is insufficient for a given feature space dimensionality, the sample scatter and covariance matrices are likely to be ill-conditioned, and transformations such as PCA and LDA may not yield optimal projections. Similarly, other techniques such as best-bands selection [10] are also likely to be sub-optimal for ATR and ground cover classification tasks, considering the fact that

they do not fully utilize the rich spectral information in hyperspectral (or multispectral) signatures for final classification.

There is a growing interest in using multiple data sources for robust ATR. Data fusion in this context typically exploits multiple, independent observations of a phenomenon and involves a feature level or a decision level fusion for various recognition and identification tasks [11], [12]. In this work, a divide-and-conquer approach is proposed that employs such data fusion techniques to exploit hyperspectral data, which otherwise typically suffers from the small-sample-size problem. The key problem that is addressed in this research is the design of a robust classification system, capable of performing recognition tasks such as ATR accurately under the small-sample-size formulation – i.e., when the size of the training data is much less than that required to support the dimensionality of the feature space.

1.2 Motivation behind the Proposed Work – Classification of Hyperspectral Imagery under Small Training Sample Size Conditions

Hyperspectral imagery is a three-dimensional cube where two dimensions are spatial and one dimension is spectral. Thus, each pixel is actually a vector comprised of a hyperspectral signature containing up to hundreds or thousands of spectral bands. Recording reflectance values over a wide region of the spectrum potentially increases the class separation capacity of the data as compared to gray scale imagery (where most of the class specific information is extracted from spatial relations between pixels) or multispectral imagery (where reflectance values at a few spectral bands are recorded). Availability of this rich spectral information has made it possible to design classification

systems that can perform ground cover classification and target recognition very accurately. However, this advantage of hyperspectral data is typically accompanied by the burden of requiring large training datasets. In order to facilitate accurate estimation of class conditional statistics of hyperspectral data and to avoid ill-conditioned formulations, it is necessary to have sufficient ground truth (labeled) training data available a-priori. This however is not guaranteed in a general remote sensing setup. In fact, in many hyperspectral applications (for example, the detection of isolated targets), the amount of ground truth pixels available to the analyst may be less than the dimensionality of the data. Another ramification of having a high dimensional feature space is over-fitting of decision boundaries by classifiers [2], and consequently, poor generalization capacity. In other words, in such high dimensional spaces, it is possible that a good classifier will learn the decision boundaries based on the training data remarkably well, but may not be able to generalize well to a test set that varies slightly in its statistical structure.

As a result of the problems associated with hyperspectral data outlined above, in the absence of a large training database, it is common for researchers to either (a) limit the number of spectral bands they use for analysis (for example, best-bands selection), or, (b) perform transform based dimensionality reduction (such as PCA, LDA, Stepwise LDA etc.) prior to classification. Conventionally, techniques such as best-bands selection, stepwise feature extraction (e.g., stepwise LDA) etc. are commonly employed in such scenarios, but as mentioned previously, these are sub-optimal in that they do not utilize the rich spectral information provided by hyperspectral signatures. The system proposed in this work employs a multi-classifier, decision fusion framework to exploit such hyperspectral data.

1.3 Alternate High Dimensional Classification Application – Robust Computer Aided Diagnosis (CAD) of Mammographic Masses

Despite mammography being the modality of choice for the detection of breast cancer, mammogram images are sometimes hard to read, because some breast cancers blend into breast tissue. Radiologists often employ Computer Aided Diagnosis (CAD) systems to facilitate greater accuracy in detection. Most end-to-end CAD systems follow a three step approach – (1) Image enhancement and segmentation, (2) Feature extraction, and, (3) Classification. While the state-of-the-art in image enhancement and segmentation can now very accurately identify regions of interest for feature extraction, the resulting feature spaces are typically very high dimensional. This adversely affects the performance of classification systems because a large feature space dimensionality necessitates a large training database to accurately model the statistics of benign and malignant features. As an alternate application, in this dissertation, the multi-classifier decision fusion framework that employs a divide-and-conquer approach for alleviating the affects of high dimensionality of feature vectors is tested with such a mammography dataset. The feature space is partitioned into multiple smaller sized groups, and a bank of classifiers (a multi-classifier system) is employed to perform classification in each group. Finally, a decision fusion system merges decisions from each classifier in the bank into a single decision per mammogram.

1.4 Contributions of this Work

This research seeks to design a system that is capable of performing classification tasks on high dimensional data when only a relatively small amount of training data is

available. Based on an intelligent partitioning scheme, the (one-dimensional) hyperspectral data is partitioned into smaller subspaces. After appropriate pre-processing, the data in each subspace is applied to a separate classifier (independent of other subspace classifiers). The local classifications resulting from this bank of classifiers are fused in an appropriate manner using a decision fusion system. This procedure partitions the single classification problem over the entire hyperspectral space into multiple classification problems, each over a subspace of a much smaller dimension. In the process, the system uses the entire spectral information for classifying pixels, while alleviating the problems associated with high dimensional data – ill-conditioning due to small-sample-size, and, over-fitting of decision boundaries due to high dimensionality.

The primary contributions of this dissertation are listed below.

1. Design appropriate partitioning schemes from a multi-classifier, decision fusion perspective, which will ensure acceptable local classification across all subspaces, and a robust decision fusion to fuse these local classifications.
2. Design an adaptive classifier weight assignment scheme for a multi-classifier decision fusion system, where weights are based on a-priori knowledge acquired from training data. This approach may prove critical for success in applications where non-uniform fidelity exists across subspaces.
3. Determine the sensitivity of various multi-classifier, decision fusion schemes, including the proposed methods, to different signal fidelity conditions, particularly for hyperspectral remote sensing applications.

4. Determine appropriate transform based projections at the subspace level that will improve class separation in the proposed framework. In particular, the following projections are studied.
    - Conventional dimensionality reduction techniques (such as LDA). These projections are known for their potential to improve class separation when the original class distributions are uni-modal.
    - Kernel based discriminant analysis projections (Such as Kernel LDA). These projections have been recently explored in the pattern classification community to improve class separation when original class distributions are multi-modal, or the class separation is non-linear.
5. Analyze the generalization ability of the proposed framework by applying it to alternate small-sample-size application classification tasks.

1.5 Outline of this Dissertation

The outline of this dissertation is as follows. Chapter II will review the relevant background information, current state-of-the-art in pattern classification systems, and the challenges faced in current classification paradigms under small-sample-size conditions. In Chapter III, the proposed multi-classifier and decision fusion framework is discussed in detail. The efficacy of various metrics for partitioning the hyperspectral space into contiguous subspaces is studied in this chapter. An adaptive confidence based classifier weight assignment is also proposed, that ensures robust classification performance when hyperspectral signatures possess fidelity that is non-uniform across the spectrum. In

chapter IV, a kernel discriminant analysis projection is proposed at the subspace level, to ensure reliable classification performance, even when the pixel mixing between target and background pixels is severe. In chapter V, to illustrate the aptness of the proposed multi-classifier, decision fusion framework to alternate applications, the framework is employed for three alternate practical classification tasks – (1) Invasive species classification using satellite hyperspectral imagery, (2) A multitemporal hyperspectral classification task, and, (3) A classification of malignant and benign masses using digital mammogram images in a CAD system. Results from these experiments will demonstrate that the MCDF framework can be extended to different high-dimensional, small-sample-size statistical pattern classification problems. Chapter VI concludes this dissertation with a summary of results and discussion of suggested future work in this direction.

REFERENCES

[1] A.K. Jain, R.P.W. Duin, Jianchang Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.22, no.1, pp.4-37, Jan 2000.

[2] R.O. Duda, P.E. Stark, D.G. Stork, *Pattern Classification*, Wiley Inter-science, October 2000.

[3] Pedro J Moreno. *Speech Recognition in Noisy Environments*, PhD Thesis*,* ECE Dept., Carnegie Mellon University, 1996.

[4] J.P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE* , vol.85, no.9, pp.1437-1462, Sep 1997.

[5] Kobatake, H., Yoshinaga, Y., and Murakami, M., "Automatic detection of malignant tumors on mammogram," *Proceedings of the International Conference on Image Processing.*, vol. 1, pp. 407–410, 1994.

[6] M.D. Farrell, R.M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," in *IEEE Geoscience and Remote Sensing Letters,* Vol. 2, No. 2, pp. 192-195, April 2005.

[7] A. Agatheeswaran, "Analysis of the Effects of JPEG2000 Compression on Texture Features Extracted from Digital Mammograms, Master of Science", *MS Thesis*, Mississippi State University, 2004.

[8] D.L. Swets and J. Went, "Using Discriminating Eigenfeatures for Image Retrieval," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 18, No. 8, pp. 831-836, August 1996.

[9] S. Prasad, L.M. Bruce, "Limitations of Subspace LDA in Hyperspectral Target Recognition Applications," *Proc. IEEE International Geoscience and Remote Sensing Symposium*, July 2007.

[10] R. Pu, P. Gong, "Band Selection from Hyperspectral Data for Conifer Species Identification," Proceedings of Geoinformatics'00 Conference, Monterey Bay, pp 139-146, June 21-23.

[11] P. Watanachaturaporn, P.K. Varshney, M.K. Arora, "Multisource fusion for land cover classification using support vector machines", *Proc. 8'th International Conference on Information Fusion*, pp. 614 – 621, July 2005.

[12] B. Jeon, D.A. Landgrebe, "Decision fusion approach for multitemporal classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 37, no. 3, pp 1227-1233, May 1999.

CHAPTER II

BACKGROUND AND CHALLENGES

2.1 Conventional Single-Classifier based Pattern Recognition Systems

Traditional pattern recognition systems are typically based on supervised or unsupervised classification algorithms. Supervised pattern classification entails the use of labeled training data for learning appropriate class conditional statistical models, which are then employed for making labeling decisions about unlabeled test data. In scenarios where labeled training data is not available, unsupervised classification is employed to identify clusters and patterns and assign them to unknown classes [1], [2]. In this dissertation, supervised classifiers are employed and studied, since they are expected to perform better than unsupervised classifiers, and the tasks that are studied in this dissertation allow for labeled training data to be available a-priori. Assigning class labels to patterns is a task employed in a wide variety of fields. Some examples include the use of such systems for fingerprint recognition [1], speech recognition [1], [3], speaker identification [4], automatic target recognition in images collected by remote sensing modalities and automatic detection of breast cancer by analyzing mammograms [5]. In this chapter, a description of some supervised classification techniques will be provided. A description of various dimensionality reduction techniques, and their advantages and disadvantages will also be presented. Finally, relevant background information for

hyperspectral image analysis systems and CAD systems for digital mammography will be provided.

2.1.1 Supervised Classification Techniques

As the name suggests, supervised classification techniques are "supervised" with the aid of class labels of training data provided to the classifier. Based on these class labels, a supervised classifier builds a statistical model for every class. For each test data sample that comes through, it compares it to each of the available statistical models, to find the "nearest" match, and assigns the label of the statistical model with which the sample matches well. Supervised classifiers themselves can be either parametric, or non-parametric in nature. Parametric classifiers, such as a maximum-likelihood classifier, parameterize the statistical model for every class with a finite number of parameters. In the case of a maximum-likelihood classifier, which will be described momentarily, the parameters could be the class means and covariance matrices. Another commonly employed parametric classifier, particularly for time-series analysis is a Hidden Markov Model (HMM) classifier. Non-parametric classifiers on the other hand do not attempt to parameterize the class-conditional distributions. Instead, they rely on other techniques, such as a histogram approximation for representing class-conditional statistics. In this dissertation, quadratic maximum-likelihood classifiers will be employed for class labeling.

Quadratic maximum-likelihood classifiers employed in this dissertation assume a normal (Gaussian) class conditional distribution [2] for every class. Assuming equal

priors (a-priori class probabilities), the class membership function for such a classifier is given by [2]

$$M(w_i \mid \vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x} - \vec{\mu}_i) - \frac{1}{2}\ln\left|\Sigma_i\right|, \tag{2.1}$$

where the distribution for the i'th class is given by $p(w_i \mid \vec{x}) \sim N(\mu_i, \Sigma_i)$.

Assuming unequal priors [2], the corresponding class membership function in (2.1) is modified to incorporate a-priori information as

$$M(w_i \mid \vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x} - \vec{\mu}_i) - \frac{1}{2}\ln\left|\Sigma_i\right| + \ln P(w_i) \ , \tag{2.2}$$

where, $P(w_i)$ is the prior probability of class i estimated from the training data. For a two-class problem, the resulting decision surfaces, as represented by the class membership functions in (2.1) and (2.2) are hyperquadrics [2].

The choice of this classifier in this dissertation is motivated by the fact that after a LDA (or KDA) based pre-processing, most feature spaces will exhibit Gaussian distributions in the transformed space (by virtue of the central-limit theorem), and hence a normal density function becomes a fair assumption. Further, the choice of a normal density function in the model makes parameterization easy – mean vectors and covariance matrices are sufficient to represent class statistics.

2.2 Dimensionality Reduction Techniques for Pattern Classification

As mentioned previously, the high dimensionality of hyperspectral data comes with both advantages and disadvantages. The rich spectral information is likely to be beneficial for most target recognition and ground cover classification tasks. However, lack of sufficient training data creates a possibility for the class conditional statistical

models to be ill-conditioned, and hence, can have a detrimental effect on the recognition or classification problem at hand. Further, the extremely large dimensionality of hyperspectral data coupled with limited training data also often results in poor generalization capability, where the decision boundaries learned by the statistical classifier over-fit the training data. To mitigate these consequences of hyperspectral data, most researchers project the hyperspectral data onto lower dimensional spaces before proceeding with the classification task.

Common choices of dimensionality reduction algorithms in the pattern classification community are best-bands selection, PCA, LDA and variations such as subspace LDA, stepwise LDA etc. [6], [7], [8], [10]. PCA and its variants are sub-optimal in a general classification setup [9]. LDA transformations and their variants require labeled training data to estimate the transformation. Further, LDA transformations are likely to break down when the class conditional distributions are multi-modal in nature. Techniques such as best-bands selection on the other hand do not utilize the rich spectral information available in hyperspectral signatures, and hence, by design are sub-optimal. The following sub-sections will briefly review three commonly employed dimensionality reduction mechanisms – PCA, LDA and best-bands selection, since these projections (and some of their variations, such as stepwise LDA) will be employed as baseline transformations – as a comparison with the recognition performance of the proposed system. An intuitive mathematical argument will also be presented to show the detrimental effects of PCA on class separation.

2.2.1 *Principal Components Analysis (PCA)*

PCA seeks to find a linear transformation $\vec{y} = W^T \vec{x}$, where $\vec{x} \in \Re^m$, $\vec{y} \in \Re^n$ and $m > n$, such that the variance of the data is maximized in the projected space. Mathematically, PCA is a diagonalizing transformation that diagonalizes the covariance matrix of the global data set. It is also an unsupervised transformation in the sense that it does not require labeled training data for finding the transformation. While *m* is the dimensionality of the original feature space, *n* is the desired dimension of the projected space, and is usually determined as the number of significant eigenvalues in the spectral decomposition of the global covariance matrix. Assume that the covariance matrix of $\{\vec{x}\}$ (let us denote it as $\sum_x$) is diagonalizable, and the spectral decomposition is given by

$$\sum_x = U \Lambda U^T , \tag{2.3}$$

where Λ is a diagonal matrix with eigenvalues on the diagonal, and U is the corresponding eigenvector matrix. It is easy to show [2], [13] that the optimal transformation in a mean squared error sense, W^T that maximizes the overall spread of the data is given by the eigenvector matrix after removing eigenvectors corresponding to small eigenvalues from it. A detailed discussion and explanation of the PCA algorithm can be found in many standard image processing books [2], [13]. Although it is a very powerful tool in signal analysis and coding, by design, PCA does not maximize class separation in the projected space. A mathematical argument describing the ineptness of PCA in pattern classification tasks is presented in section 2.3.

2.2.2 *Linear Discriminant Analysis (LDA)*

LDA seeks to find a linear transformation $\vec{y} = W^T \vec{x}$, where $\vec{x} \in \Re^m$, $\vec{y} \in \Re^n$ and $n \leq c-1$, (c is the number of classes), such that the within class scatter is minimized and the between class scatter is maximized [13]. The transformation W^T is determined by maximizing Fisher's ratio,

$$J_1(W) = \left|W^T S_b W\right| \Big/ \left|W^T S_w W\right|, \tag{2.4}$$

which can be solved as a generalized eigenvalue problem. The solution is given by the eigenvectors of the following eigenvalue problem.

$$S_w^{-1} S_b W = \Lambda W, \tag{2.5}$$

where S_b is the between-class scatter matrix and S_w is the within-class scatter matrix, defined as

$$S_b = \sum_{i=1}^{c} n_i (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T, \tag{2.6}$$

$$S_w = \sum_{i=1}^{c} \sum_{\vec{x} \in C_i} (\vec{x} - \vec{m}_i)(\vec{x} - \vec{m}_i)^T \tag{2.7}$$

Note that $S_T = S_w + S_w$ is the total scatter matrix, which is related to the global covariance matrix by a scaling factor. The rank of S_b is not greater than c-1, where c is the number of classes. This results in a transformation matrix W which projects data from an m-dimensional feature space, to an n-dimensional projected space, $n \leq c-1$ [13]. Further, by seeking a solution that maximizes Fisher's ratio, the projected data contains class clusters that possess compact within-class structure and a well separated between-

class structure. This ensures that all classes are well separated in the feature space, thereby enhancing recognition and classification performance.

2.3 Limitations of PCA in Pattern Classification Tasks

In recent work, Prasad and Bruce [9] showed by means of a mathematical argument and experimental evidence that PCA projections can be detrimental to pattern classification tasks. Before discussing the discrimination power in the projected space, a suitable optimality criterion quantifying class separation needs to be defined. For linear transformations of the feature space, $\vec{y} = W^T \vec{x}$, a common choice for quantifying class separation is Fisher's ratio

$$J_1(W) = \left| W^T S_b W \right| \Big/ \left| W^T S_w W \right|. \tag{2.8}$$

The choice of this ratio as an optimality criterion stems from the need of feature extraction algorithms for pattern classification tasks to minimize the within class scatter while maximizing the between class scatter. Lu *et al* have suggested the following modification of the Fisher's ratio in [14], which, can be proved to be equivalent to the original Fisher's ratio, in terms of the maximizing solution

$$J_2(W) = \left| W^T S_b W \right| \Big/ \left| W^T (S_w + S_b) W \right|. \tag{2.9}$$

The theoretical argument presented here covers two scenarios – (1) When the within-class and total scatter matrices are full ranked, and, (2) When the within-class and total scatter matrices are rank-deficient. Scenario (1) is likely to occur when there are enough training data vectors relative to the dimensionality of the feature space, and, the feature space itself does not contain overly redundant features (redundant features typically make the scatter matrices rank deficient.) For the purpose of analysis, we can

use the total scatter matrix, S_T instead of the total covariance matrix, since both are related by a normalization constant. Let S_b, S_w and S_T be the scatter matrices in the original feature space, as defined in section 2.2. A PCA projection will then solve the following eigenvalue problem

$$(S_w + S_b)W = \Lambda W \ . \tag{2.10}$$

On the other hand, LDA specifically solves for the maximization of the optimality criterion using the generalized eigenvalue approach [2], [13]. Using the second form of Fisher's ratio, J_2, LDA solves the following eigenvalue problem

$$(S_w + S_b)^{-1} S_b W = \Lambda W \ , \tag{2.11}$$

which is known to maximize class separation. It follows that PCA will maximize the optimality criterion only when the solution of (2.8) is same as the solution of (2.9). It is obvious that a common solution will not exist for any arbitrary S_b, S_w and S_T. An intuitive way to picture this is the following. Let the spectral decomposition of S_T be

$$S_T = U \Lambda U^T \ . \tag{2.12}$$

In a PCA projection, we choose eigenvectors corresponding to large eigenvalues of S_T for projection (let us say, the first n are retained), and the projection matrix is given by $\widetilde{U}^T$, which denotes a matrix containing the principal directions for projection. Let the corresponding diagonal matrix of eigenvalues be $\widetilde{\Lambda}$. When $\widetilde{U}^T$ is used as the projection matrix, in the projected space, the modified Fisher's ratio, J_2, becomes

$$J_2(\widetilde{U}) = \frac{\left|\widetilde{U}^T S_b \widetilde{U}\right|}{\left|\widetilde{U}^T (S_w + S_b) \widetilde{U}\right|} = \frac{\left|\widetilde{U}^T S_b \widetilde{U}\right|}{\left|\widetilde{U}^T S_T \widetilde{U}\right|} = \frac{\left|\widetilde{U}^T S_b \widetilde{U}\right|}{\left|\widetilde{\Lambda}\right|} = \frac{\left|\widetilde{U}^T S_b \widetilde{U}\right|}{\prod_{i=1}^{n} \lambda_i} . \tag{2.13}$$

Here, n is the number of principal components retained in the PCA projection. Clearly, the modified Fisher's ratio is not guaranteed to increase relative to the original space by this projection because

- the value of the numerator, $\left|\tilde{U}^T S_b \tilde{U}\right|$ is not guaranteed to increase since $\tilde{U}$ represents principal directions of S_T, not S_b.
- the value of the denominator, $\left|\tilde{\Lambda}\right|$ is actually greater than the value of $|\Lambda|$ in the original space, because small eigenvalues (typically, numerically close to zero) in Λ were discarded to create $\tilde{\Lambda}$.

From these arguments, it is clear that PCA is not an optimal transformation for feature extraction stages of pattern recognition systems. Further, it also follows that any transformation that employs PCA as an intermediate transformation (e.g., Subspace LDA [9]) is also likely to be sub-optimal from a pattern classification perspective. Any reported improvements in classification performance due to PCA projections are likely to be a consequence of the characteristic of the dataset that was employed for the study (e.g., if the directions of large global variance were indeed the directions of good class separation.)

The discussion above deals with scenario (1) - the case where the within-class and total scatter matrices are well-conditioned. Scenario (2), which deals with situations where PCA is applied as a tool to discard the null space of S_T, as with subspace LDA, is studied next. Note that by definition, S_b has a rank of at most $c-1$, where, c is the number of classes. On the other hand, S_w (and hence, S_T) may either be full ranked or rank deficient, depending on the amount of training data and redundancy of features. Zheng *et*

al pointed out in [15] that when S_w is rank deficient, the transformation that maximizes the optimality criterion in an ideal sense would project the data onto a subspace $N(S_T)^{\perp} \cap N(S_w)$, where $N(S_T)^{\perp}$ is the orthogonal complement of the null space of S_T and $N(S_w)$ is the null space of S_w. One way to visualize this is to realize that an ideal transformation will shrink the within class scatter (by projecting to the null space of S_w) in the non-null space of S_T.

In the following discussion, scatter matrices in the transformed space are denoted with a tilda, $\widetilde{S}$. It is common practice in PCA transformations to project the data in directions such that the significant eigenvalues of the overall covariance matrix (or total scatter matrix) are retained. In situations where S_T is rank deficient, consider a simple PCA projection that discards the null space of S_T. Techniques such as subspace LDA employ PCA with this goal. Such a transformation ensures that after projection, $N(\widetilde{S}_T) = \{\Phi\}$, the null set. If we restrict $\widetilde{S}_b$ and $\widetilde{S}_w$ to be positive semi definite,

$$N(\widetilde{S}_T) = \left\{N(\widetilde{S}_b) \cap N(\widetilde{S}_w)\right\}. \tag{2.14}$$

Hence, after the PCA projection,

$$N(\widetilde{S}_b) \cap N(\widetilde{S}_w) = \{\Phi\}. \tag{2.15}$$

Recall that the desired projection space is $N(\widetilde{S}_T)^{\perp} \cap N(\widetilde{S}_w)$. However, if $N(\widetilde{S}_T) = \{\Phi\}$ (after a PCA projection), it only implies that the intersection of the null spaces of $\widetilde{S}_b$ and $\widetilde{S}_w$ is a null set. This does not guarantee $N(\widetilde{S}_w)^{\perp} = \{\Phi\}$; i.e. it does not guarantee retention of the null space of S_w. However, the most discriminative projected space requires both:

$$N(\widetilde{S}_T) = \{\Phi\} \text{ and } N(\widetilde{S}_w)^{\perp} = \{\Phi\}. \quad (2.16)$$

Hence, even in a situation where PCA is used as a preprocessing step (to resolve singularity issues in S_T) before another feature reduction step (e.g., as in subspace LDA), discarding the null space of S_T is not necessarily the optimal strategy.

2.4 Limitations of LDA in Pattern Classification Tasks

LDA, though beneficial as a supervised dimensionality reduction technique, is not designed for tasks where the class-conditional statistics are multi-modal. This however is not a serious impairment for a wide array of classification tasks, since for these tasks, uni-modal class conditional distribution functions make for good approximations to the actual distributions. However, since the generalized eigenvalue formulation for LDA involves an inverse of the within-class scatter matrix, S_w, in situations where S_w is rank-deficient, such as when limited training data (relative to the dimensionality of the original feature space) is available to estimate it, it is not possible to find a reliable solution to the corresponding eigenvalue problem.

One solution that some researchers have previously proposed to recondition the formulation is a regularization technique, where, a small energy identity matrix is added to S_w before estimating its inverse. The resulting formulation is known as regularized LDA (R-LDA). Another approach that has had some level of success with face recognition tasks is the Subspace LDA approach, which is discussed in the previous section. In recent work, Prasad and Bruce [9] demonstrated mathematically and experimentally that this approach is sub-optimal at best, and does not help in classification of hyperspectral data. Yet another approach that alleviates the high

dimensionality, small-sample-size problem is the stepwise LDA (S-LDA) approach, also known as Discriminant Analysis Feature Extraction (DAFE). S-LDA with forward selection and backward rejection is now commonly employed to mitigate affects of small-sample-size on LDA transformations. The forward selection procedure starts by calculating a certain metric (in this work, A_z, area under the Receiver Operator Characteristics / ROC curve is chosen) for each feature. The A_z values are sorted in descending order. The feature with the highest A_z gets placed into a feature vector, and the ROC area A_{z_BEST} is set to A_{z1}. The second best feature is then appended to the feature vector, and A_{z2} is computed. The second best feature is only retained if $A_{z2} > A_{z\_BEST}$. This process is repeated until all the individual features are examined, or until the bound on the maximum number of features in the feature vector is reached. This bound should not be larger than would be supported by the available training data size insofar as learning the LDA transformation is concerned.

Next, backward rejection is performed. Assume at this stage that there are b features selected in the feature vector, and the best ROC area is A_{z_BEST}. If $b = 1$, then no features may be removed, and the process halts. If $b > 1$, then the first feature is removed, and the ROC area A'_{z1} is calculated. If $A'_{z1} > A_{z\_BEST}$, then the first feature vector is removed, and A_{z_BEST} is set to A'_{z1}. This process continues until all features have been examined.

This forward selection and backward rejection approach results in a determination of the "best" feature subset, upon which if LDA is applied, the class separation in the resulting space will be high. Recent studies involving the DAFE algorithm have shown A_z [26] to work well in the forward selection, backward rejection task. Although this

algorithm allows us to draw on the benefits of the LDA transformation for high dimensional feature spaces, it is still sub-optimal, in that the selection and rejection procedures outlined above do not perform an exhaustive search in the feature space to find the optimal 'combinations' of features.

2.5 Hyperspectral Best-Bands Selection

Hyperspectral best-bands selection (also referred to as band selection) is a dimensionality reduction technique employed by many researchers to extract the "most useful" bands of a hyperspectral dataset for classification and recognition tasks [16]. This approach entails the use of an appropriate performance metric as a criterion for selecting bands out of all available bands of hyperspectral signatures for training a classifier. The performance metric employed could be supervised (i.e., make use of class label information), or unsupervised. In either case, this approach is sub-optimal by design even though it can reduce the dimensionality of high dimensional hyperspectral data substantially, since it discards a majority of the spectral information before classification.

In this work, band selection (BNDS) will be used as another baseline method, against which recognition performance of the proposed system will be compared. Due to its popularity and previously documented "efficacy" [16], entropy will be used as the metric for band selection in this dissertation. The corresponding BNDS algorithm is simple – Use training data to estimate entropy of each feature in the feature vector; rank order features in descending order of entropy; select the top n features as the best-bands / best-features. Entropy based BNDS is believed to be an effective dimensionality reduction method because it selects the n most-informative features as the reduced

dimensional feature space. In chapter 3, experimental results will provide a quantitative comparison of the efficacy of these approaches (R-LDA, S-LDA and BNDS), as compared to the proposed divide-and-conquer approach.

2.6 Hyperspectral Image Analysis Background

As mentioned previously, hyperspectral imagery is a three-dimensional cube where two dimensions are spatial and one dimension is spectral. Thus, each pixel is actually a vector comprised of a hyperspectral signature containing up to hundreds or thousands of spectral bands. This dense sampling of reflectance values over a wide region of the spectrum potentially increases the class separation capacity of the data as compared to gray scale imagery (where most of the class specific information is extracted from grey level statistics and spatial relations between pixels) or multispectral imagery (where reflectance values at a few spectral bands are recorded). Availability of this rich spectral information has made it possible to design classification systems that can perform ground cover classification and target recognition very accurately. Hyperspectral reflectance signatures of various plant species can be seen in Figure. 2.1.

Note the sharp “red-edge”, prominent at the transition from the red region of the spectrum to the Near-Infrared (NIR) region. This sharp transition is a typical characteristic of most vegetation signatures, and is in-fact used for distinguishing vegetation species from non-vegetation species in an image. Also note that although the various species look similar in the visible portion of the spectrum, considerable differences in reflectance values can be observed at larger wavelengths (e.g., in the NIR region.) Hence, it is safe to infer that having reflectance values recorded over a wide region of the spectrum can indeed provide a better recognition performance.

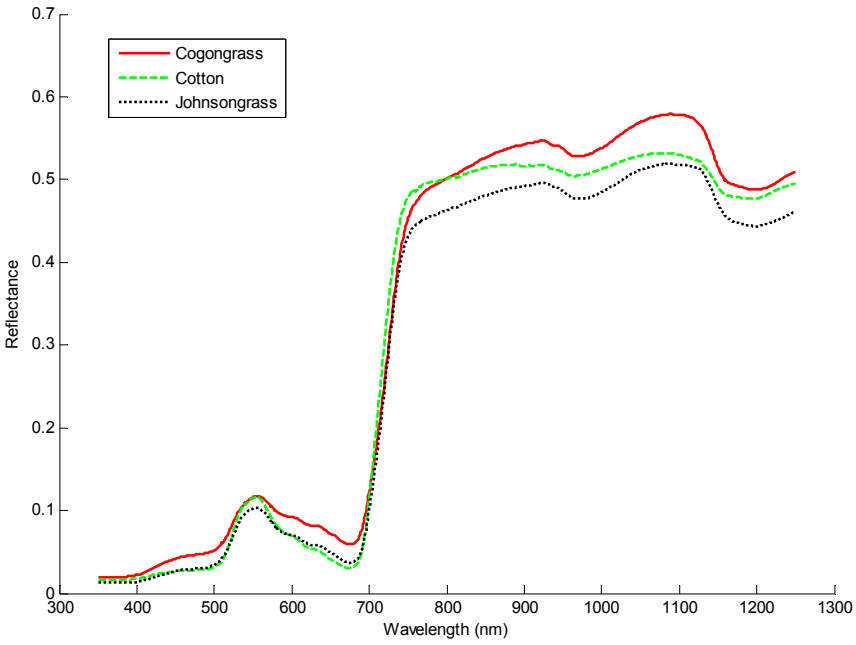


Figure 2.1 Hyperspectral signatures of various plant species.

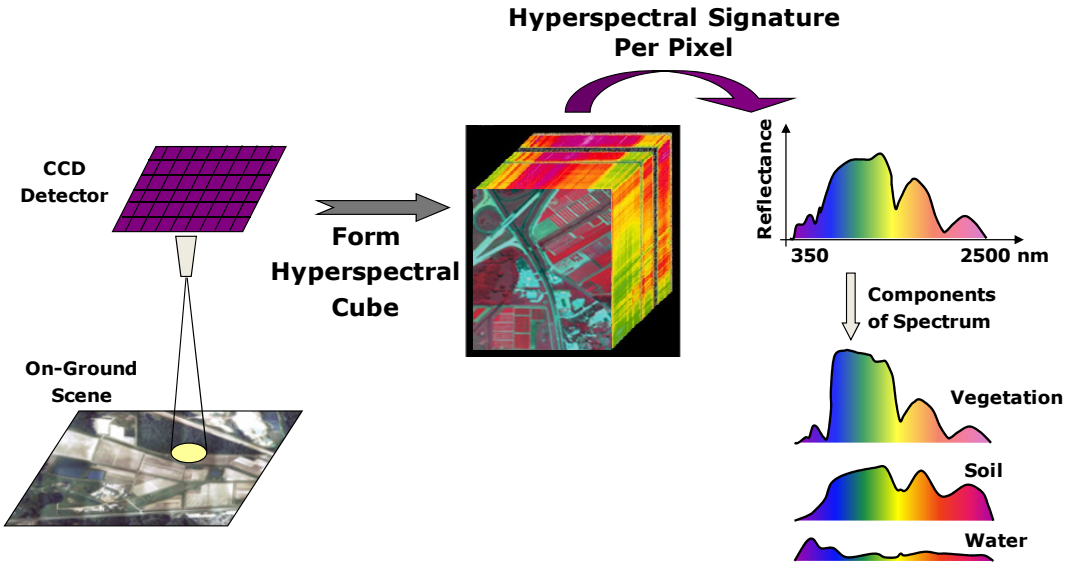


Figure 2.2 Overview of hyperspectral remote sensing systems.

Image courtesy Dr. Lori Mann Bruce, Mississippi State University.

Figure 2.2 illustrates a typical hyperspectral remote sensing system. Despite its advantages, using hyperspectral reflectance values as features in a pattern classification setup nevertheless can result in an over-dimensionality and ill-conditioned scenario induced by the high dimensionality and small-sample-size. This necessitates the use of large training datasets if conventional algorithms are to be employed for classification tasks. This however is not guaranteed in a general remote sensing setup. In fact, in many hyperspectral applications (for example, the detection of isolated targets), the amount of ground truth pixels available to the analyst may be less than the dimensionality of the data. As mentioned in the previous chapter, another ramification of having a high dimensional feature space is over-fitting of decision boundaries by classifiers [2], and consequently, poor generalization capacity. With this in mind, we study the performance

of the proposed multi-classifier, decision fusion systems using various experimental datasets and simulated signal fidelity conditions over the next three chapters.

Due to its potential and challenges, hyperspectral image analysis has been a topic of active research over the past decade. In [17], Compact Airborne Spectrographic Imager (CASI) data was employed for identifying subsets of the available spectral bands for robust target material classification. In [18], [19], [20], wavelet based analysis of hyperspectral data was performed to study the distribution of signal energy of hyperspectral reflectance signatures at different scales and resolutions. In [21], Lin *et al* proposed a parametric projection pursuits algorithm for projecting high dimensional reflectance signatures onto lower dimensional spaces. In [22], Hsu *et al* performed dimensionality reduction by using Fourier and wavelet transform basis of the available spectral and spatial data, and then chose the "best" bases as those that provided the best approximation of the training data.

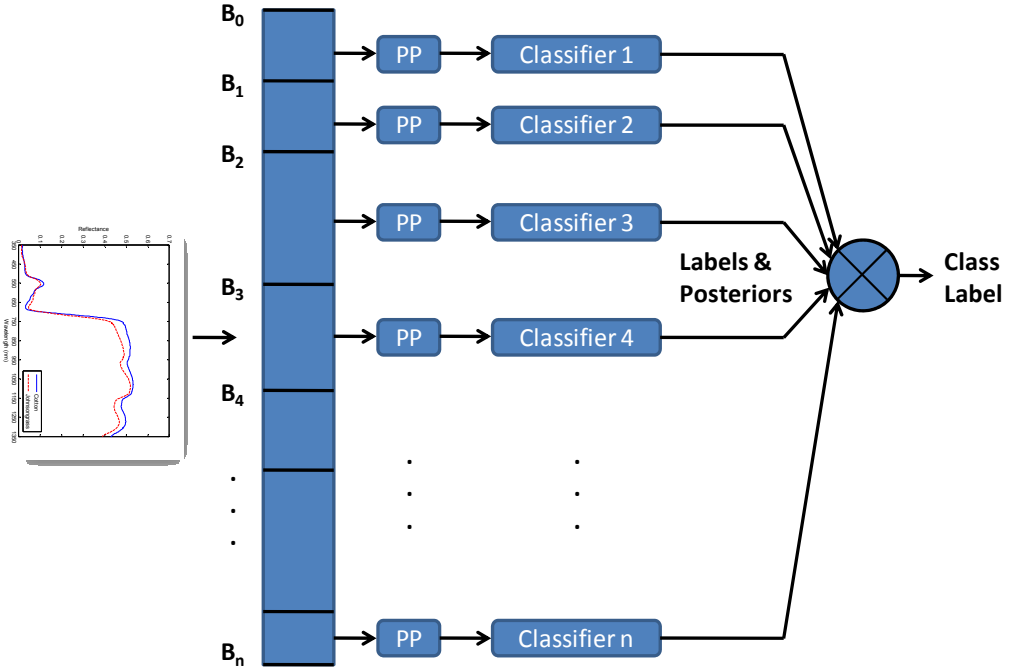


Figure 2.3 The overall block diagram representation of the proposed system.

PP refers to "pre-processing". B_0, B_1… B_n are spectral bands that define the partitioning of the spectrum.

2.7 Proposed System Architecture

The solution proposed to alleviate the small-sample-size problem and high dimensionality of hyperspectral data is based on a divide-and-conquer approach. Fig. 2.3 illustrates the block level functionality of the proposed system. Note that this system is designed to work for supervised classification tasks. Training data is employed to intelligently partition the hyperspectral space into a set of contiguous subspaces. An appropriate pre-processing (for example, LDA) is performed on each subspace. This is followed by a bank of classifiers (each assigned to a particular subspace) making local classification decision in each subspace. Each of these local pre-processing operations and classification formulations are well conditioned, and exploit information in a certain

distinct subset of the spectrum. The classification results (in the form of class labels and posterior probabilities) from each subspace are merged into a single classification decision per test pixel / signature using an appropriate decision fusion mechanism. This dissertation will study the effect of each part of the proposed system on the overall classification performance, under varying data fidelity conditions. In particular, the following studies will be conducted in the proposed framework:

1. *Techniques for band-grouping (subspace identification)*: Under this category, various "intelligent" band-grouping metrics will be examined for their efficacy in identifying partitions of the hyperspectral space that best serve the multi-classifier decision fusion system. As will be described in the next chapter, the subspace identification task as approached in this work is based on a bottom-up band-growing procedure that monitors a certain performance metric of the group that is being grown. A new metric that better identifies subspaces in the current context is also proposed. Effect of various types of performance metrics on the overall classification performance of the system will be studied.
2. *Suitable pre-processing of input data in each subspace*: After subspace identification, it is desirable to improve class separation, and reduce the dimensionality of subspaces for improved classification performance. It has already been established that PCA projections are not optimal projections in classification tasks. LDA projections on the other hand are well suited for classification tasks where class conditional density functions are uni-modal. Because they improve class separation by design, LDA transformations are employed at the pre-processing level. Note that although LDA transformations

over the entire hyperspectral space are likely to be ill-conditioned due to a limited sample size, this is not the case for LDA projections at the subspace level. In severe pixel mixing conditions however (such as when a significant percentage of target pixels are mixed with background pixels), the class conditional density functions are likely to be multi-modal, and LDA projections may no longer be optimal (even in the proposed divide-and-conquer framework). In such situations, benefits of projecting data into a kernel space will be explored. In particular, efficacy of KDA projections to improve class separation in kernel induced spaces will be studied in the MCDF framework.

3. *Impact of decision fusion strategies on classification performance:* The decision fusion strategy employed plays an important role in merging the classification decisions from individual subspaces. Benefits of hard and soft decision fusion techniques will be studied in the proposed framework. Further, an adaptive weight assignment technique is proposed that will weigh the classification decisions (hard, e.g., labels; and soft, such as posterior probabilities) of each classifier based on its relative strength in accurately classifying training data.

Finally, the studies described above will be repeated with different simulated data fidelity conditions. In particular, pixel mixing will be simulated by linearly mixing target and background pixels. This will simulate a scenario where the spatial resolution of the sensor is not high enough to just capture the target in the pixels. Further, noisy data conditions will be simulated by adding noise to the hyperspectral signatures. Performance of various aspects of the proposed system will be studied in these simulated conditions.

2.8 Alternate Pattern Classification Application – Computer Aided Diagnosis of Benign and Malignant Tumors in Digital Mammography

To demonstrate that the proposed divide-and-conquer framework can serve as a useful approach for different types of classification tasks, an alternate classification task – robust detection of benign and malignant tumors in digital mammography is chosen in this dissertation. CAD tools have been developed over the years to aid radiologists in the interpretation of mammograms. Breast cancer is believed to be among the leading causes of cancer related deaths among women, and mammography is the modality of choice for detecting breast cancer [23], [24]. As is the case with many medical imaging modalities, significant research is being conducted for the design of Computer Aided Diagnosis (CAD) systems. A typical CAD system performs the following tasks in succession: (1) Image enhancement and segmentation, (2) Feature extraction, and, (3) Classification. Robust image enhancement and segmentation algorithms are now available for identifying regions of interest in mammogram images [25]. The features extracted from these segmentations are however oftentimes very high dimensional; for example, some CAD systems result in hundreds or even thousands of features [25], [26]. This has ramifications on the performance of the backend classification system in that the size of available training data (number of available training mammograms) does not match the required size needed to accurately model the statistical characteristics of high dimensional feature spaces. To alleviate this problem, many researchers employ some feature selection or dimensionality reduction method to reduce the size of the feature space. One popular choice in the medical imaging community for doing so is S-LDA, which employs a forward selection and backward rejection technique to identify a subset

of available features that are potentially useful for the classification task. This however is not necessarily the optimal solution to the problem of over-dimensionality. S-LDA does not perform an exhaustive search on all feature combinations to select the "best" features. Further, any approach that leaves out certain features and uses only a smaller subset for classification is clearly sub-optimal, in that it is not employing all the information available to perform classification.

In this dissertation, as an alternate application, the proposed divide-and-conquer approach is used to robustly classify mammogram images from very high dimensional feature spaces generated using state-of-the-art image enhancement, segmentation and feature extraction algorithms. The proposed approach partitions the high dimensional feature space into many smaller dimensional subspaces. A bank of classifiers (multi-classifier system) performs "local" classification in each such subspace, and an appropriate decision fusion system "fuses" these local classification results into a final malignant/benign classification for every mammogram image. In doing so, all the available information is employed for classification while the problems associated with overly high dimensional feature spaces are avoided. It is hence expected that the proposed system will more accurately classify malignant and benign mammogram images. A discussion on the pre-processing employed, the features extracted from these mammograms, the proposed classification system and experimental results with a mammography database is provided in section 5.3. Since the key motivation behind exploring the efficacy of the proposed framework for a CAD application is to demonstrate the generalization ability of the framework to other classification tasks, details on how to read mammograms to identify malignant tumors, the explanation

behind the choice of features used in CAD systems for mammography etc. are omitted from this work. The reader can review details of these concepts from [26].

REFERENCES

[1] A.K. Jain, R.P.W. Duin, Jianchang Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.22, no.1, pp.4-37, Jan 2000.

[2] R.O. Duda, P.E. Stark, D.G. Stork, *Pattern Classification*, Wiley Inter-science, October 2000.

[3] Pedro J Moreno. *Speech Recognition in Noisy Environments*, PhD Thesis*,* ECE Dept., Carnegie Mellon University, 1996.

[4] J.P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE* , vol.85, no.9, pp.1437-1462, Sep 1997.

[5] Kobatake, H., Yoshinaga, Y., and Murakami, M., "Automatic detection of malignant tumors on mammogram," *Proceedings of the International Conference on Image Processing.*, vol. 1, pp. 407–410, 1994.

[6] M.D. Farrell, R.M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," in *IEEE Geoscience and Remote Sensing Letters,* Vol. 2, No. 2, pp. 192-195, April 2005.

[7] A. Agatheeswaran, "Analysis of the Effects of JPEG2000 Compression on Texture Features Extracted from Digital Mammograms, Master of Science", *MS Thesis*, Mississippi State University, 2004.

[8] D.L. Swets and J. Went, "Using Discriminating Eigenfeatures for Image Retrieval," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 18, No. 8, pp. 831-836, August 1996.

[9] S. Prasad, L.M. Bruce, "Limitations of Principal Component Analysis for Hyperspectral Target Recognition," Accepted for publication in the *IEEE Geoscience and Remote Sensing Letters, 2008*.

[10] R. Pu, P. Gong, "Band Selection from Hyperspectral Data for Conifer Species Identification," Proceedings of Geoinformatics'00 Conference, Monterey Bay, pp 139-146, June 21-23.

[11] P. Watanachaturaporn, P.K. Varshney, M.K. Arora, "Multisource fusion for land cover classification using support vector machines", *Proc. 8'th International Conference on Information Fusion*, pp. 614 – 621, July 2005.

[12] B. Jeon, D.A. Landgrebe, "Decision fusion approach for multitemporal classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 37, no. 3, pp 1227-1233, May 1999.

[13] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[14] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, "Regularization studies on LDA for face recognition," in *Proceedings of the 2004 International Conference on Image Processing,* Vol. 4, pp 63-66, October 2004.

[15] W. Zheng, L. Zhao, C. Zou, "An efficient algorithm to solve the small-sample-size problem for LDA," in *Pattern Recognition*, Vol. 37, No. 5, pp 1077-1079, May 2004.

[16] Keshava, N., "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," in the *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 42, pp 1552-1565, July 2004.

[17] R. Garano, G. Gaston O. Heyman, "Remote sensing systems for identifying and mapping aquatic vegetation in estuaries and other near-shore environments along the pacific coast," *NASA Commercial Remote Sensing Program Office*, John C. Stennis Space Center, 1999.

[18] L. M. Bruce, C. Morgan, S. Larsen, "Automated detection of subpixel targets with continuous and discrete wavelet transforms," IEEE Transactions on Geoscience and Remote Sensing, vol. 39, no. 10, pp. 2217-2226, 2001.

[19] J. Li, L. M. Bruce, J. Byrd, J. Barnett, "Automated detection of Pueraria montana (kudzu) through Haar analysis of hyperspectral reflectance data," IEEE International Geoscience and Remote Sensing Symposium, vol. 5, pp. 2247–2249, 2001.

[20] Y. Huang, L. M. Bruce, T. Koger, D. Shaw, "Analysis of the effects of cover crop residue on hyperspectral reflectance discrimination of soybean and weeds via Haar transform," IEEE International Geoscience and Remote Sensing Symposium, vol. 3, pp. 1276-1278, 2001.

[21] H. D. Lin, L. M. Bruce, "Projection pursuits for dimensionality reduction of hyperspectral signals in target recognition applications," IEEE International Geoscience and Remote Sensing Symposium, vol. 2, pp. 960 - 963, 2002.

[22] P. H. Hsu, Y. H. Tseng, “Feature extraction of hyperspectral data using the best wavelet packet basis,” IEEE International Geoscience and Remote Sensing Symposium, vol. 3, pp. 1667 – 1669, 2002.

[23] National Cancer Institute, "National Cancer Institute Fact Sheet: Improving Methods for Breast Cancer Detection and Diagnosis," 2006. Available: http://www.cancer.gov/cancertopics/screening/breast

[24] A. Jemal, T. Murray, E. Ward, A. Samuels, R.C. Tiwari, A. Ghafoor, E.J. Feuer, and M.J. Thun, "Cancer Statistics, 2005," CA: A Cancer Journal for Clinicians, vol. 55, no. 1, pp. 10-30, 2005.

[25] J. E. Ball, L.M. Bruce, “Level Set-Based Core Segmentation of Mammographic Masses Facilitating Three Stage (Core, Periphery, Spiculation) Analysis,” Proc. Of EMBS, France, 2007.

[26] J.E. Ball, "Three stage level set segmentation of mass core, periphery, and spiculations for automated image analysis of digital mammograms." Ph.D. in Electrical Engineering. Starkville, MS: Mississippi State Univ., May 2007.

CHAPTER III

MULTI-CLASSIFIERS, DECISION FUSION AND CONFIDENCE BASED WEIGHT ASSIGNMENT FOR HYPERSPECTRAL CLASSIFICATION

3.1 Introduction

There is a growing interest in using multiple data sources for robust ATR and land cover classification. Data fusion in this context typically exploits multiple, independent observations of a phenomenon and performs a feature level or a decision level fusion for various recognition and identification tasks. For example, in Watanachaturaporn *et al* [1], different types of data (IRS-1C LISS III images, NDVI and DEM), collected in the Himalayan region were fused for land cover classification. Jeon *et al* [2] used decision fusion techniques for multi-temporal classification. Similarly, Memarsadeghi *et al* [3] studied the fusion of data from Hyperion and ALI sensors in the PCA and wavelet domains, for improved invasive species forecasting. More recently, Fauvel *et al* [4] have studied the use of multi-classifiers and decision fusion for the classification of urban images. Chanussot *et al* [5] have studied the use of fuzzy fusion techniques for detecting linear features in SAR multitemporal images. In particular, multi-source data fusion facilitates accurate image analysis and classification in scenarios where data from a single sensor or source lacks resolution or fidelity in the spatial or spectral domain [6], [7].

In this chapter, a divide-and-conquer approach is presented that employs such data fusion techniques to exploit hyperspectral data, which otherwise typically suffers

from the small-sample-size problem. Conventional hyperspectral ATR systems employ dimensionality reduction schemes for projecting the data onto a lower dimensional subspace, which is then used by a single classifier for labeling tasks [8]. Although these dimensionality reduction schemes are successful in reducing the ground truth requirement for unbiased modeling by the classifier [8], [9], these projections are not necessarily optimal from a pattern classification perspective [10], [11]. For example, a PCA projection may discard useful discrimination information if it were oriented along directions of small global variance, a LDA projection will be inaccurate for multimodal class distributions, etc. Another factor that governs the efficacy of such dimensionality reduction techniques is the amount of training signatures required to learn the projections. For example, if the number of training signatures is insufficient for a given feature space dimensionality, the sample scatter and covariance matrices are likely to be ill-conditioned, and the transformations such as PCA and LDA may not yield optimal projections. Similarly, other techniques such as best-bands selection [12] are also likely to be sub-optimal for ATR and land cover classification tasks, considering the fact that they do not fully utilize the rich spectral information in hyperspectral (or multispectral) signatures for final classification.

In this work, the hyperspectral space is partitioned into contiguous subspaces such that the discrimination information within each subspace is maximized, and the statistical dependence between subspaces is minimized. Each subspace is then treated as a separate source in a multi-source multi-classifier setup. Various decision fusion schemes are employed to merge classification outputs (labels/posterior probabilities) from the multi-classifier system and their efficacy is studied. In doing so, we do not discard potentially

useful information in the hyperspectral signatures, and also overcome the small-sample-size problem, since the number of training signatures required per subspace is substantially lower than if we directly used all the bands with a single classifier system. In fact, the minimum number of training signatures required in this scheme is governed by the size of the largest subspace formed during partitioning, which is typically much smaller than the size of the original hyperspectral space.

Previous approaches to band grouping [13], [14] use a combination of correlation between variables (in this case, bands) and Bhattacharya distance to partition the hyperspectral space. In this work, the efficacy of higher order statistical information (using average mutual information) instead of simple correlation is studied, for a bottom-up band grouping [15], [16]. A confidence based adaptive weight assignment scheme is also proposed for decision fusion - where the weight associated with a classifier's decision depends on its confidence in recognizing training data. The advantage of this adaptive classifier weight assignment over a uniform weight assignment for hyperspectral classification is studied.

This chapter is organized as follows. In section 3.2, the proposed method of partitioning the hyperspectral space using mutual information is described. In section 3.3, details of the proposed technique of employing multi-classifier systems for hyperspectral classification problems are presented. A description of various decision fusion schemes employed in this work is also included. Relevant implementation issues are also discussed in this section, along with explanation of how these have been addressed. Section 3.4 contains a description of the handheld hyperspectral dataset used for experimental evaluation of the algorithm. In section 3.5, experimental results quantifying

the efficacy of the proposed algorithm are presented. Section 3.6 summarizes key results and observations from the various experiments in this chapter.

3.2 Subspace Identification

3.2.1 Subspace Identification using Band Grouping

Subspace identification is the first step in the proposed multi-classifier, decision fusion system. It involves intelligent partitioning of the hyperspectral feature space into contiguous subspaces such that each subspace possesses good class separation, and the statistical dependence between subspaces is minimized. A classifier is then dedicated to every subspace, and an appropriate fusion rule is employed to combine the local classification decisions into a final class label for every test signature.

In this work, a bottom-up band grouping algorithm is proposed for subspace identification. Figure 3.1 depicts the application of the band grouping procedure on hyperspectral signatures. Using labeled training signatures, each subspace is grown in a bottom-up fashion (i.e., continue to add successive bands to the subspace) until the addition of bands no longer improves some performance metric. At this point, growth of the current subspace is stopped and the procedure is repeated for the next subspace. The metric employed for band grouping should be such that it simultaneously ensures good class separation within a group as well as low inter-group dependence. While good class separation per group is important for accurate decision making at the subspace level, a low inter-group dependence ensures robust decision fusion of these local decisions. A band grouping threshold (t) controls the sensitivity of partitioning to changes in the metric. This threshold is the tolerance value for the percentage change in the metric used

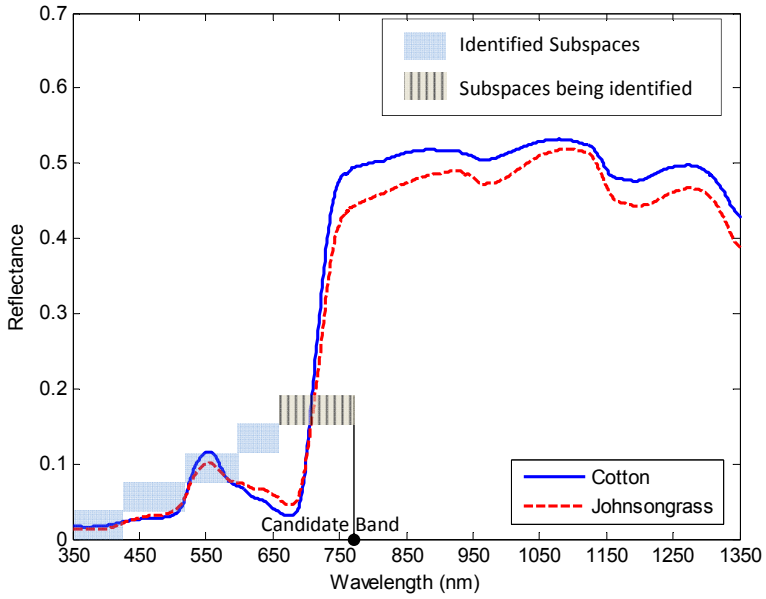


Figure 3.1 Illustrating the bottom-up band growing procedure for subspace identification.

The signatures depicted above are the average of all hyperspectral signatures of Cotton and Johnsongrass in the experimental hyperspectral database.

for stopping growth of the subspace being identified. Let M_{i-1} be the performance metric of the subspace being identified without the addition of the i'th band, and, let M_i be the performance metric of the subspace with the i'th band included, then, the band grouping threshold, t is defined as

$$t = \frac{M_i - M_{i-1}}{M_{i-1}}. \tag{3.1}$$

In this work, the value of t is set to zero, that is, the growth of the subspace being identified is stopped when addition of the i'th band does not change the value of the performance metric being monitored. In addition to monitoring changes in the performance metric, upper and lower bounds are imposed on the size of each subspace during the band grouping procedure. The lower bound (chosen as 10 bands in this work)

ensures that the number of subspaces formed does not increase unreasonably. It also ensures that subspaces are not any smaller than would be supported by the approximately block diagonal statistical structure of the correlation or mutual information matrices of hyperspectral data. The upper bound (chosen as 25 in this work) ensures that the size of each subspace is not so large that supervised dimensionality reduction and classification algorithms start to fail because of ill-conditioned statistical estimates. This bound should be adjusted based on the amount of training data available for dimensionality reduction and classification.

3.2.2 Mutual Information and Subspace Identification

It can be inferred from the preceding discussion that the choice of performance metric plays an important role in the performance of the proposed system. Previously [13], [14], various combinations of Bhattacharya distance and feature cross-correlation have been studied as potential performance metrics. In recent work [17], Tsagaris *et al* have suggested the use of Mutual Information for defining blocks of bands of hyperspectral data in the context of color representation. In this work, a metric using Mutual Information is proposed for band grouping.

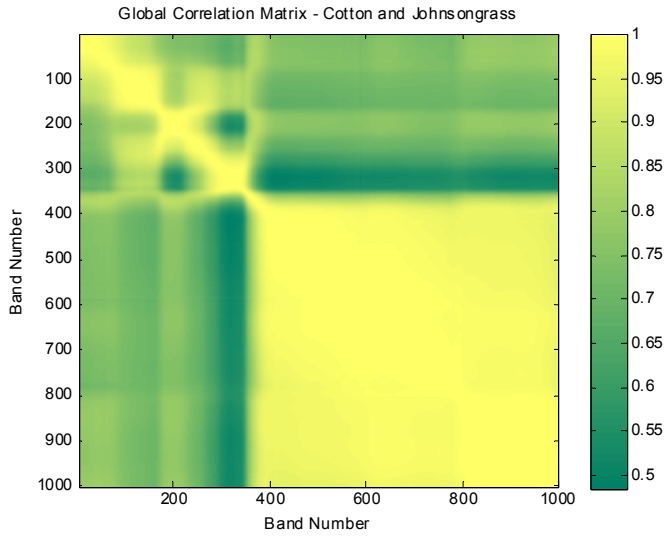
In the subspace identification process, a good class separation in every subspace reduces the local classification errors, while statistical independence between subspaces ensures diversity in the multi-classifier setup. A multi-classifier, decision fusion system will be beneficial if there is diversity in the subspaces or in the models (e.g., classifiers). Redundancy between subspaces is not desired in a decision fusion setup since it may lead to propagation of errors (e.g., in majority vote fusion, if two different subspaces produce

identical errors in classification, a single type of error contributes to two bad votes and so on). Instead of restricting the partitioning process to second order statistics (correlation), it is proposed that incorporating higher order statistics (as quantified by mutual information) into the metric shall generate a more meaningful partitioning of the hyperspectral space. Mutual information between two discrete valued random variables x and y is defined [18] as

$$I(x,y) = \sum_{i \in x} \sum_{j \in y} P(i,j) \log \frac{P(i,j)}{P(i)P(j)}. \tag{3.2}$$

Here, *P(i,j)* is the joint probability distribution of *x* and *y*, and *P(i)* and *P(j)* are the marginal probability distributions of *x* and *y* respectively. These can be estimated using histogram approximations. In the context of hyperspectral images, *x* and *y* represent reflectance values for a pair of bands. Figure 3.2 shows the global correlation matrix and the global average mutual information matrix for an experimental hyperspectral dataset. Details of this dataset are provided in section 3.4. Note that both statistical measures reveal an approximate block diagonal structure. It is this block diagonal nature of feature cross correlation (and mutual information) that allows us to partition this space into approximately independent and contiguous subspaces. Further note that the average mutual information matrix reveals a finer block diagonal structure as compared to the correlation matrix. Based on these observations, the metric employed for partitioning in this work is as follows

$$JMAMI_n = JM_n AMI_n, \tag{3.3}$$



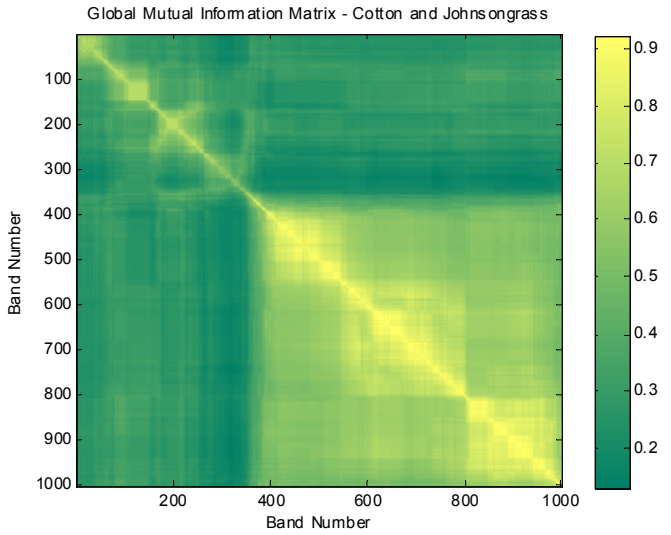


Figure 3.2 Global correlation matrix and mutual information matrix for experimental hyperspectral data.

AMI_n is the minimum average mutual information between a candidate band and the remaining bands in the current (n'th) subspace, and JM_n is the between class Jeffries Matsushita (JM) distance of the current subspace, and is given by

$$JM = 2(1 - e^{-BD}), \text{ where,}$$
$$BD = -\ln\left(\sum_{x \in X} p(x)q(x)\right) \quad . \tag{3.4}$$

BD is the Bhattacharya distance; $p(x)$ and $q(x)$ are the probability distributions of the two classes between which the distance is being estimated. As will be explained later, in this dissertation, both distributions are assumed to be Gaussian. *JM* distance is chosen to measure class separation, because unlike Bhattacharya distance it has an upper bound. This results in a normalized metric possessing lower and upper bounds. In a multi-class situation, JM_n is evaluated as the minimum pair-wise JM distance between classes in the current subspace. Previously, correlation has been employed for partitioning the space into approximately independent subspaces. The corresponding metric is similar to the one in (3.3) and is written as *JMCorr*, where mutual information is replaced by correlation.

3.2.3 The Jarque-Bera Test

Thus far, an intuitive argument towards employing mutual information for subspace identification has been presented. To reinforce the aptness of a mutual information based metric instead of a correlation based metric by means of a quantitative comparison, proof of deviation from normality of hyperspectral data is presented here. Towards this end, the Jarque-Bera (JB) test [19] was performed on experimental hyperspectral data on a per-band basis. The JB test is a two sided goodness of fit test which uses the following test statistic to accept or reject the null hypothesis

$$T = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right), \tag{3.5}$$

where n is the sample size, s is the sample skewness, and k is the sample kurtosis. The

null hypothesis is that the data under analysis comes from a normal distribution, whereas the alternative hypothesis is that it does not come from a normal distribution.

Figure 3.3 depicts the results of this test on experimental hyperspectral data (Cotton vs. Johnsongrass). H is either one or zero, depending on whether the null hypothesis was rejected or not. This hypothesis test was conducted at the 5% significance level. Each band of the hyperspectral data was treated as a random variable and was tested for normality by this test. It can be seen that at the per class level (e.g., Cotton only, or Johnsongrass only), there are several bands of the spectrum which deviate from normality (H = 1). A similar observation can be made at the global level (Cotton and Johnsongrass combined). Since various bands of the hyperspectral data deviate from marginal normality, we can infer that the data deviates from multi-variate normality. It

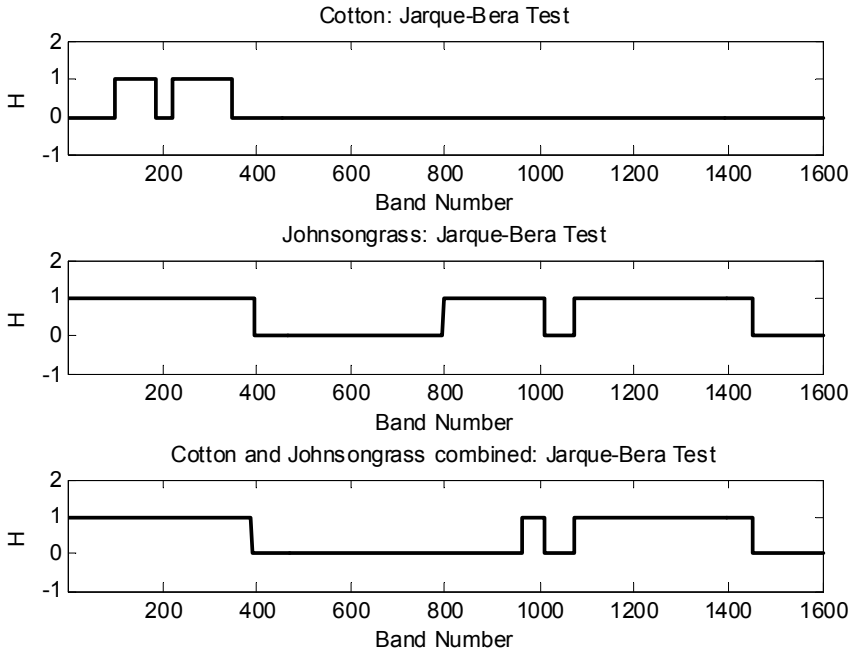


Figure 3.3 Jarque-Bera test for experimental hyperspectral data on a per-band basis for the first 1600 bands.

can hence be concluded that it is reasonable to expect a mutual information based metric to provide a more meaningful partitioning of the hyperspectral space, since it is not restricted to second order moments.

3.3 Multi-Classifiers and Decision Fusion (MCDF)

After partitioning the hyperspectral space into independent subspaces, each with good class separation, a multi-classifier system is employed followed by a decision fusion process to make classification decisions. The proposed Multi-Classifier and Decision Fusion (MCDF) system is essentially a bank of classifiers that make "local" decisions in the partitioned subspaces, followed by a decision fusion mechanism that fuses these individual decisions. These can be parametric classifiers such as maximum likelihood classifiers, or non parametric classifiers such as *k* nearest neighbors classifiers, neural network based classifiers etc. In this work, quadratic maximum likelihood classifiers are employed [20]. Figure 3.4 depicts the block level functionality of the proposed system. Training signatures are used to identify appropriate contiguous subspaces, which are represented by the bands on the edges. Training data is also used to ascertain subspace accuracies using the same bank of classifiers that will be used for classifying test signatures. This training accuracy assessment allows us to assign a confidence score to each subspace, which can then be used for pruning away subspaces with lower than acceptable scores, or for assigning weights to each classifier adaptively in the decision fusion process. Decision fusion can occur either at the class label level (hard fusion), or at the posterior probability level (soft fusion). The system is tested with decision fusion at both these levels.

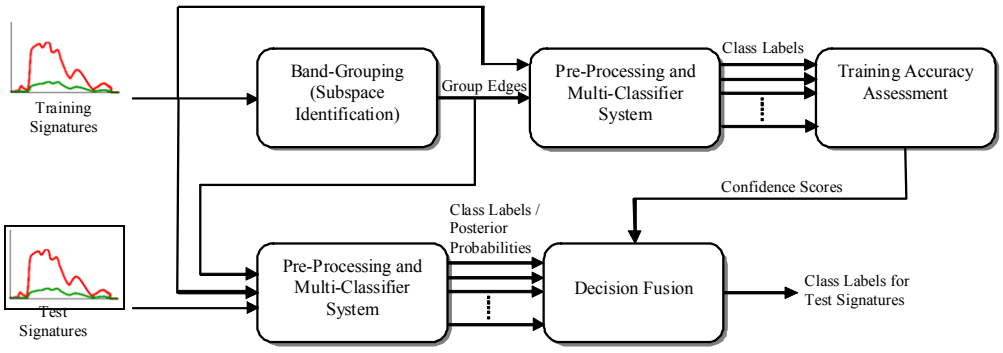


Figure 3.4 Block level functionality of the feature space partitioning, multi-classifier and decision fusion systems.

Since each subspace is of a much smaller dimensionality than the dimension of the original hyperspectral signature, a suitable preprocessing (such as LDA) may prove beneficial before making the local classification decisions. For uni-modal class conditional density functions, a LDA based dimensionality reduction is likely to improve class separation in the projected space. Recall that we impose an upper bound on the size of subspaces during the subspace identification process. One of the considerations during choosing an appropriate upper bound is for the within and between class scatter matrices to be well conditioned. Hence, LDA based dimensionality reduction at the local subspace level is going to be well conditioned for most subspaces, as opposed to a single LDA based projection on the entire hyperspectral space, which is likely to be ill-conditioned in the absence of a lot of training data.

3.3.1 Hard Decision Fusion – Majority Voting

In hard decision fusion, a final classification decision is arrived at based on a vote over individual class labels (hard decisions) from each subspace. Unlike soft fusion based techniques, the overall classification of majority voting based fusion is not very sensitive

to inaccurate estimates of posterior probabilities. However, in situations where posterior probabilities can be accurately estimated, soft fusion methods are likely to provide stable and accurate classification. A simple majority vote (MV) is given by

$$w = \underset{i \in \{1,2...C\}}{\arg\max} N(i)$$
$$\text{where}, N(i) = \sum_{j=1}^{n} I(w_j = i). \tag{3.6}$$

where I is the indicator function, w is the class label from one of the C possible classes for the test pixel, j is the classifier index, n is the number of subspaces (and hence the number of classifiers), and $N(i)$ is the number of times class i was detected in the bank of classifiers. The form of voting described in (3.6) is based on uniform weight assignment, i.e., every classifier in the multi-classifier, decision fusion system enjoys equal voting strength. An adaptive voting mechanism is proposed, where strong classifiers enjoy a greater influence in the final decision. One possible way of performing this adaptive voting mechanism is to weigh a classifier's vote based on its confidence score which can be learned from training data. Hence, (3.6) is modified to incorporate a non-uniform weight assignment:

$$w = \underset{i \in \{1,2...C\}}{\arg\max} N(i)$$
$$\text{where}, N(i) = \sum_{j=1}^{n} \alpha_j I(w_j = i), \tag{3.7}$$

where α_j is the confidence score (weight) for the j'th classifier. The voting scheme depicted in (3.6) is referred to as Majority Voting (MV), and the proposed voting scheme in (3.7) is referred to as Weighted Majority Voting (WMV) in this dissertation.

3.3.2 Soft Decision Fusion – Linear and Logarithmic Opinion Pools

Soft decision fusion entails the use of posterior probabilities, or more generally some class membership function from every classifier for making the final decision. Unlike hard fusion techniques, soft decision fusion schemes do not rely solely on class labels from each classifier to make the final decision. Two popular soft decision fusion schemes are linear and logarithmic opinion pools [21]. A linear opinion pool uses the individual posterior probabilities of each classifier ($j = 1, 2, \ldots n$), $p_j(w_i/x)$ to estimate a global class membership function

$$\begin{aligned} C(w_i \mid x) &= \sum_{j=1}^{n} \alpha_j p_j(w_i \mid x), \\ w &= \underset{i \in \{1,2 \ldots C\}}{\arg\max}\, C(w_i \mid x). \end{aligned} \tag{3.8}$$

Once again, the classifier weights (α_j, $j = 1, 2, \ldots n$) can either be uniformly distributed over all classifiers, or can be assigned based on the confidence score of each classifier. This is essentially a weighted average of posteriors across the classifier bank. In a Logarithmic Opinion Pool, the global class membership function is modified to be a weighted product of the posterior probabilities of all classifiers, instead of a weighted sum

$$\begin{aligned} C(w_i \mid x) &= \prod_{j=1}^{n} p_j(w_i \mid x)^{\alpha_j} \\ \Rightarrow \log C(w_i \mid x) &= \sum_{j=1}^{n} \alpha_j \log p_j(w_i \mid x). \end{aligned} \tag{3.9}$$

The logarithmic opinion pool has some advantages over a linear opinion pool [21]: (1) the resulting class membership function in a logarithmic opinion pool is uni-modal, and, (2) decisions from different classifiers are treated independently in the fusion

process. However, as pointed out in Benediktsson *et al* [21], [22], this fusion scheme has one serious drawback in that a single zero (or numerically small) posterior probability can potentially veto decisions from the remaining classifiers. Hence, logarithmic opinion pools must be used carefully, and should particularly be avoided where posterior probability estimation is not accurate.

In this work, the implementations of linear and logarithmic opinion pools with uniform weight assignment are referred to as LOP and LOGP respectively, while the implementations with the proposed adaptive weight assignment are referred to as WLOP and WLOGP respectively.

3.3.3 Confidence based Weight Assignment and Pruning

In the proposed system, training data is jackknifed into further training and test data. Recognition accuracies from each subspace obtained from this data are a measure of the confidence of the subspace in the recognition task. Based on the application, certain subspaces may be better suited for the recognition task. For example, if a certain portion of the hyperspectral space is more affected by noise and distortion effects than other regions, the corresponding subspaces are bound to be less reliable than the others for the recognition task. In such situations, it is hoped that training accuracy assessment captures the confidence of each subspace in the labeling decision. In this work, the effect of assigning non-uniform weights to the bank of classifiers, based on training accuracies in the subspaces is also studied. For this purpose, the training accuracy in a subspace is assigned as the weight of the corresponding classifier. Decision fusion performance based on such a weight assignment is compared to the performance with a uniform weight

assignment. For a two class target recognition problem, an accuracy of 50% or less is worse than a random decision. Hence, in this work, subspaces with training accuracy of less than or equal to 50% are pruned away and not considered in the decision fusion process.

3.3.4 Decision Fusion – Implementation Issues

The efficacy of soft fusion techniques is dependent on the accurate estimation of posterior probabilities. Although LOP and LOGP have been used previously for remote sensing classification [21], [22], these methods have not been tested for alleviating the small-sample-size problem commonly encountered when classifying hyperspectral data. A typical characteristic of hyperspectral data is that adjacent bands (and hence features) are highly correlated. For normally distributed data, a high cross-feature correlation sometimes results in rank deficient covariance matrices, which makes the estimates of class membership functions or posterior probabilities unreliable. Note that this problem is not commonly encountered with multispectral data since adjacent bands of a multispectral sensor are separated by a reasonable amount in the wavelength domain. With hyperspectral data, we need to address this issue for reliable estimation of posterior probabilities or class membership functions. In this work, quadratic maximum likelihood classifiers are employed. These classifiers assume Gaussian class distributions for the *i*'th class, $p(x/w_i) \sim N(\mu_i, \Sigma_i)$. Assuming equal priors, the class membership function for such a classifier is given by [20]

$$M(w_i \mid x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln\left|\Sigma_i\right|. \qquad (3.10)$$

It follows from the preceding discussion that for hyperspectral data, Σ_i can sometimes be rank deficient even in the presence of sufficient training data, resulting in an unstable inverse (and hence an ill-conditioned class membership function). To resolve this issue, the null space of Σ_i is discarded with the assumption that this space contains only redundant information (i.e., Σ_i is rank deficient only due to highly correlated data, not due to insufficient data). This assumption is reasonable in the proposed multi-classifier, decision fusion approach, since each classifier deals with a subspace of a much smaller dimension, and hence the small-sample-size problem is usually not encountered. Hence, to compute the inverse of Σ_i, the Singular Value Decomposition based pseudo-inverse method is used. Similarly, the determinant of Σ_i is estimated as the product of its non-zero significant singular values, in order to discard its null space. This results in stable estimates of class membership functions and posterior probabilities.

3.4 Experimental Hyperspectral Data

Hyperspectral data employed for testing the proposed system was collected using an Analytical Spectral Devices (ASD) Fieldspec Pro FR handheld spectroradiometer [23]. Signatures collected from this device have 2151 spectral bands sampled at 1nm over the range of 350 – 2500nm with a spectral resolution ranging from 3 – 10nm. A 25° instantaneous field of view (IFOV) foreoptic was used. The instrument was set to average ten signatures to produce each sample signature, and the sensor was held nadir at approximately four feet above the vegetation canopy. Hyperspectral signatures collected with an ASD spectroradiometer tend to have high levels of noise in the regions associated with longer wavelengths, particularly when the sensor has been in use for a longer period

of time or under high temperature conditions (due to overheating of the semiconductors). Thus the signatures were truncated at 1800nm. Also, the reflectance values in the regions 1350nm - 1430nm were removed from all signatures to avoid noise due to atmospheric water absorption.

Signatures in the dataset form two classes: (1) an agricultural row crop, Cotton variety ST–4961, and (2) a weed that is detrimental to the crop's yield, Johnsongrass (*Sorghum halepense*). In this study, 54 signatures of Johnsongrass and 35 signatures of Cotton are used. These signatures were measured in good weather conditions in Mississippi, U.S.A., in 2000-2004. A target recognition scenario is created using this data treating the weed (Johnsongrass) as the target class and the crop vegetation (Cotton) as the background class, as would be the case when remote sensing is used for precision agriculture applications. Challenging target recognition tasks are created by linearly mixing target test signatures with the background at various mixing ratios (MR). All experiments reported in this dissertation are performed using a leave-one-out testing procedure. Each test target signature sequestered during the leave-one-out testing is mixed linearly with a random background signature. To ensure an unbiased setup, the background signature used in this mixing is not used for training the system. This makes it a tough and realistic ATR problem because it creates a mismatched situation where the classifiers are trained on clean target and background signatures but tested on mixed (corrupt) target signatures. The mixing ratios (background percentage to target percentage) for test target signatures reported in this work are 30:70 (MR1), 40:60 (MR2) and 50:50 (MR3). With this setup, target recognition accuracies of these sub-pixel ATR tasks are estimated using the proposed MCDF system.

Figure 3.5 Experimental hyperspectral data.

Top left: Cotton (Whitney Cranshaw, Colorado State University, Bugwood.org); Top right: Johnsongrass (James H. Miller, USDA Forest Service, Bugwood.org); Bottom: Hyperspectral signatures of Cotton and Johnsongrass.

3.5 Experimental Setup and Results

In order to determine the efficacy of the proposed algorithms, various ATR experiments are setup with the dataset described in section 3.4. Five sets of experiments are presented with the following objectives: (1) To demonstrate the benefits of using a mutual information based metric for subspace identification, (2) To study the effect of adaptive weight assignment and uniform weight assignment on decision fusion performance when the signatures possess uniform fidelity throughout the spectrum, (3) To repeat the study in objective 2 when the signatures possess non-uniform fidelity across the spectrum, (4) To study the benefits of a LDA based pre-processing at the local subspace level in the proposed decision fusion setup, (5) To compare the performance of feature level fusion with that of decision level fusion, and, (6) To compare the performance of the efficacy of the proposed MCDF framework with current state-of-the-art feature extraction approaches.

All experiments were conducted in the mixed pixel classification framework as described in section 3.4. This simulates a challenging and realistic scenario – which is commonly encountered when the size of the target is smaller than the resolution of the sensor, resulting in mixing of target signatures with background signatures. In this work, efficacy of the proposed algorithms is gauged using overall recognition accuracies (which measure the system's capacity to recognize both target and background signatures.) Further, for accurate estimation of overall recognition accuracies, all experiments were conducted using the leave-one-out cross validation method [24].

3.5.1 Experiment 1: Consistent Fidelity of the Signatures across the Spectrum

This experiment demonstrates the benefits of subspace identification using the proposed mutual information based metric instead of a correlation based metric. Further, in this experiment, uniform weight assignment is compared with adaptive weight assignment when the signatures are of uniform fidelity across the spectrum. Subspaces are identified based on two metrics – *JMCorr* and *JMAMI*. Based on the subspaces identified using these metrics, dimensionality reduction (LDA) and local classifications per subspace followed by decision fusion are carried out to obtain the class labels of test signatures.

A comparison of the overall recognition accuracies at various mixing ratios and using different decision fusion strategies is provided in Figure 3.6 for both metrics. Error bars atop all bar plots (in this and in subsequent figures) indicate the 95% confidence intervals for the recognition accuracy estimates, taking into account the finite number of available training and testing samples. We can make the following observations from the results depicted in Figure 3.6. At low mixing ratios (e.g., MR1), all decision fusion schemes and both metrics result in a near 100% recognition accuracy. Thus, the overall approach (partitioning the spectrum into subspaces of contiguous bands followed by multi-classifiers and decision fusion) is very powerful. However, if the target abundance is substantially low relative to the background abundance, then the design parameters (band grouping metric, decision fusion scheme etc.) start playing a more critical role. *JMAMI* based partitioning almost always results in higher recognition accuracy as compared to *JMCorr* based partitioning. This improvement is higher at severe mixing ratios (when the target abundance is low). We can hence infer that *JMAMI* provides a

more meaningful partition of the hyperspectral space, resulting in more robust decision fusion in this multi-classifier setup. In the remainder of this chapter, we use the *JMAMI* metric for partitioning the hyperspectral space, unless otherwise mentioned. Further, for low target abundances, MV and LOP based fusion is more reliable than LOGP based fusion. This is expected because LOGP based fusion is sensitive to the accuracy of the posterior probability / class membership function estimates. These estimates are likely to be inaccurate under severe mixing conditions. Note also that in this setup, where the hyperspectral signatures were of consistent fidelity across the spectrum, uniform weight assignment (MV, LOP, LOGP) performs as well as adaptive weight assignment (WMV, WLOP, WLOGP).

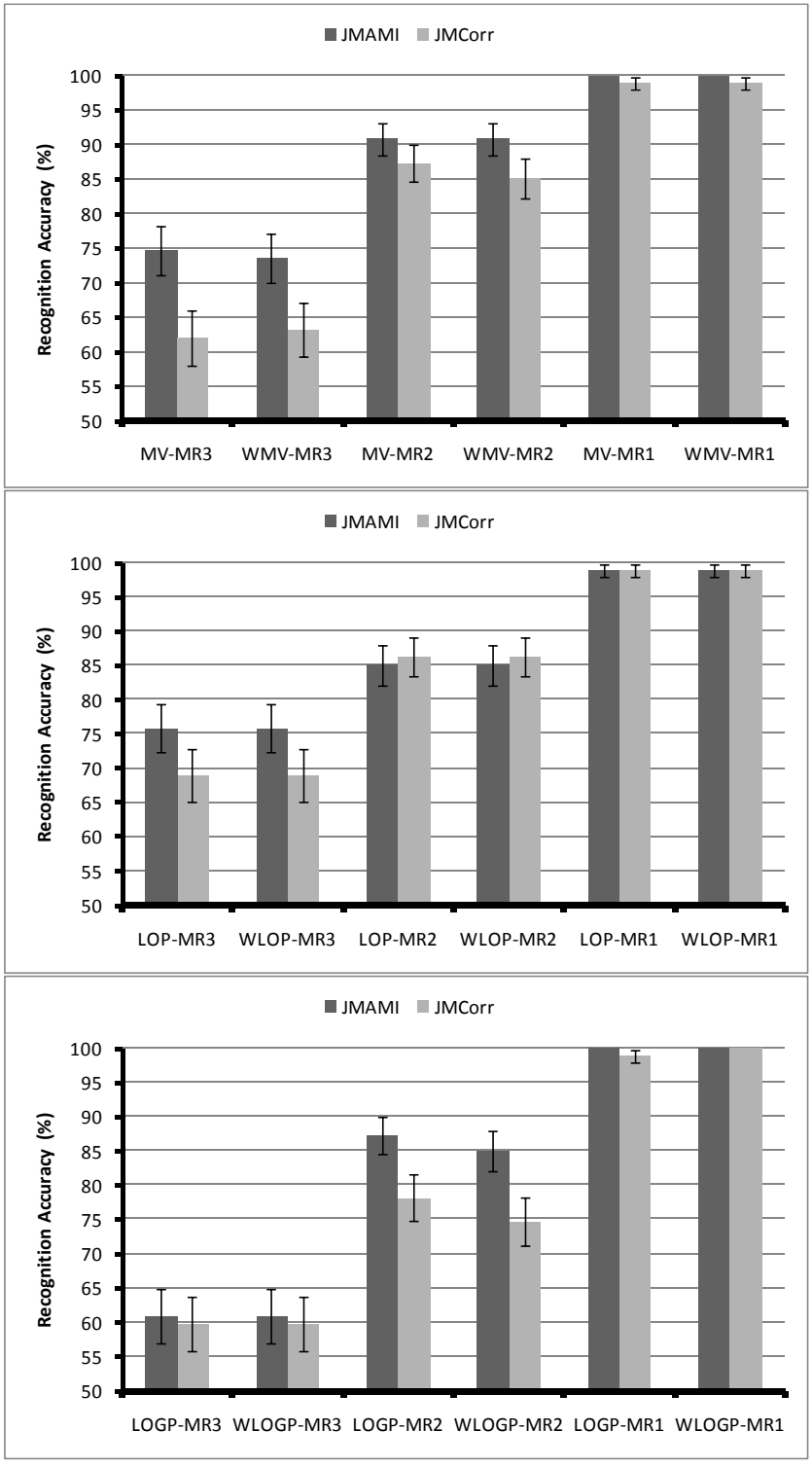


Figure 3.6 Mutual information vs. correlation based metric; uniform vs. adaptive weights.

3.5.2 Experiment 2: Non-Uniform Fidelity of the Signatures across the Spectrum

This experiment demonstrates the benefits of adaptive weight assignment over uniform weight assignment when the hyperspectral signatures have fidelity varying across the spectrum. As before, a mixed pixel classification experiment is performed, at mixing ratio MR1. Zero mean Gaussian noise is added to reflectance values in a part of the spectrum – bands 1 through 700 (approximately 350 to 1000 nm). Since many hyperspectral sensors are a composite of multiple sensors, each tuned to a specific region of the spectrum, it is possible to have a scenario where a subset of the spectrum is of poorer fidelity. Since this experiment studies noise performance, hundred iterations of noise addition and leave-one-out testing are performed and the results from these are averaged to estimate overall recognition accuracies.

Results from this experiment are summarized in Figure 3.7. Noise power was varied from 0.01 to 0.09 for hard decision fusion, and from 0.001 to 0.009 for soft decision fusion. Soft decision fusion techniques (LOP and LOGP) started to break down at noise powers at or above 0.01. This can be attributed to the inaccurate estimation of class membership functions at low signal to noise ratios. In particular, it is observed that LOGP is the least robust decision fusion scheme in the presence of additive noise. Hard decision fusion techniques are more robust to larger noise powers. Adaptive weight assignment consistently outperforms uniform weight assignment in this case. Further, the improvement in hard decision fusion using the proposed adaptive weight assignment is significantly higher than that in soft decision fusion. The improvement from using adaptive weights is relatively small at low and high noise levels, and large at moderate noise levels.

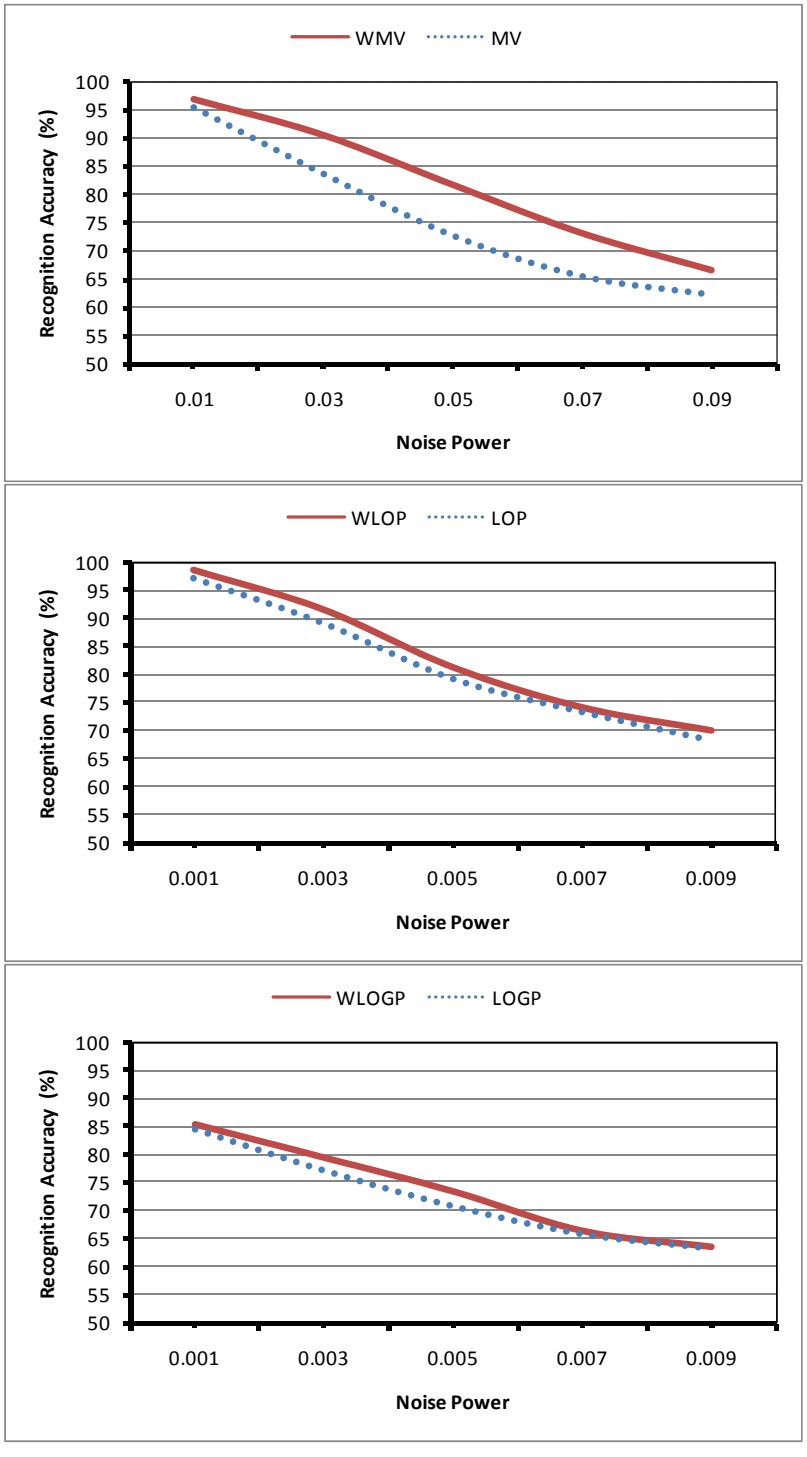


Figure 3.7 Performance of adaptive weights versus uniform weights.

3.5.3 *Experiment 3: LDA based Pre-Processing at the Subspace Level*

LDA based pre-processing per subspace has been performed before making local classification decisions in all experimental analyses presented in this chapter. This experiment shows that such a pre-processing at the subspace level in the proposed multi-classifier, decision fusion setup is indeed beneficial. Training data is used to learn the LDA transformation in each subspace and this transformation is applied to training and test data. Results from this experiment are summarized in Figure 3.8. It is clear that LDA based dimensionality reduction is beneficial for all decision fusion schemes at most mixing ratios. When the mixing becomes severe, the class conditional distributions are likely to become multi-modal, and hence the LDA transformations learned are likely to deviate significantly from optimality. This phenomenon can be observed at mixing ratio MR3, where for MV and LOGP, LDA based pre-processing reduces the overall

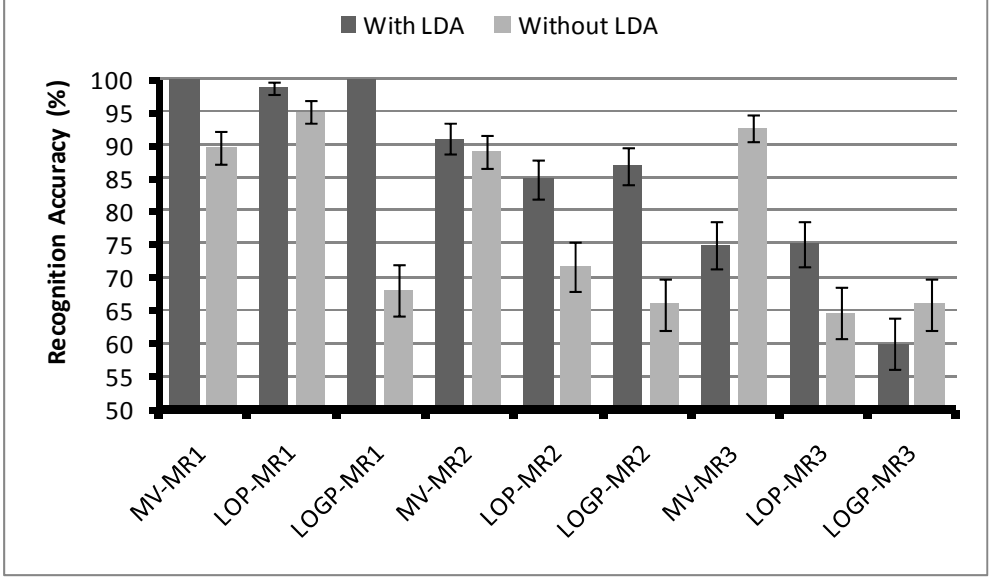


Figure 3.8 Effect of LDA based preprocessing at the subspace level on the decision fusion performance.

recognition accuracy. Recognition accuracies for WMV, WLOP and WLOGP (though omitted from this figure in order to maintain clarity), are very similar to those of MV, LOP and LOGP, since the signatures were of consistent fidelity throughout the spectrum.

3.5.4 Experiment 4: Feature versus Decision Level Fusion

In this experiment, the performance of feature level fusion is compared to that of decision level fusion. Feature level fusion refers to a simple concatenation of "optimized" features from every subspace after a suitable dimensionality reduction pre-processing (LDA in this case) per subspace. A single classifier is employed on this concatenated vector for classification. Decision level fusion is performed in a manner consistent with the MCDF implementation described previously – instead of concatenating the transformed features from every subspace, local classification decisions are made in each such LDA transformed subspace and these local decisions are then fused. Results from this experiment are reported in Figure 3.9. We can conclude from these results that

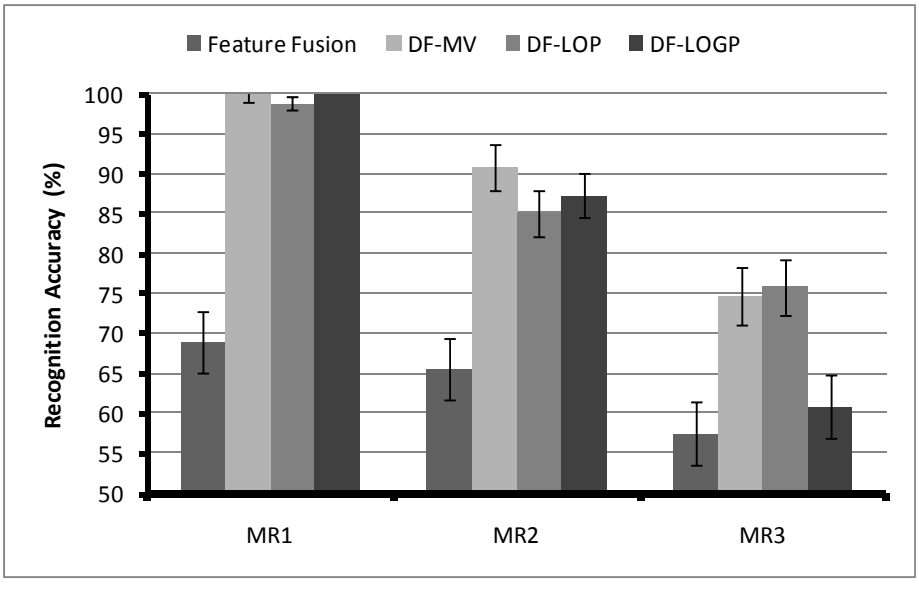


Figure 3.9 Feature versus decision level fusion.

decision level fusion consistently outperforms feature level fusion. Once again, recognition accuracies for WMV, WLOP and WLOGP (though omitted from this figure in order to maintain clarity), are very similar to those of MV, LOP and LOGP, since the signatures were of consistent fidelity throughout the spectrum.

3.5.5 Experiment 5: Comparison against Current State-of-the-art

Experiments 1 through 4 quantify the efficacy of the proposed MCDF framework in various classification scenarios, and demonstrate the usefulness of the framework for robust hyperspectral classification in small-sample-size conditions. In this experiment, the performance of the MCDF framework is compared with that of conventional algorithms employed by researchers for feature optimization and extraction in small-sample-size conditions. Towards this end, classification performance of the following feature extraction and classification systems is reported: (1) PCA, (2) R-LDA, (3) S-LDA, (4) BNDS, and (5) MCDF. For algorithms 1 through 4, a conventional single maximum-likelihood classifier is employed after each feature extraction method. These algorithms are described in the previous chapter, in sections 2.1 through 2.5. In the PCA approach, the final dimension was chosen to be equal to the number of significant eigenvalues in the spectral decomposition of the covariance matrix of the training data. In the R-LDA approach, a small constant (in this work, 1e-04) was added to the diagonal entries of the within-class scatter matrices to avoid unstable inverses in the LDA formulation. In the S-LDA algorithm, the upper limit of the intermediate feature space dimensionality in the forward selection, backward rejection procedure is set to 10. An entropy based band-selection technique was employed in the BNDS algorithm, where,

the “top” 10 features were selected. For algorithm 5, the MCDF framework with *JMAMI* based band-grouping and MV based decision fusion was employed for classification, as described in section 3.5.

Figure 3.10 depicts the overall recognition accuracy and false alarm rates using these algorithms, at the three mixing ratios, MR1, MR2 and MR3. PCA is expected to perform poorly, and that is observed in this figure. Not only does PCA based feature extraction result in poor overall classification accuracy, the associated false-alarm rate is also very high. Regularizing the scatter matrices in the R-LDA approach does not yield superior classification performance either. LDA applied on a reduced subset of features based on a forward selection and backward rejection approach (S-LDA) does yield better classification performance. Entropy based band selection (BNDS) performs slightly better than S-LDA, but at the expense of a larger false-alarm rate. Finally, the proposed MCDF framework outperforms the other algorithms at most mixing ratios. It also generates the least amount of false alarms.

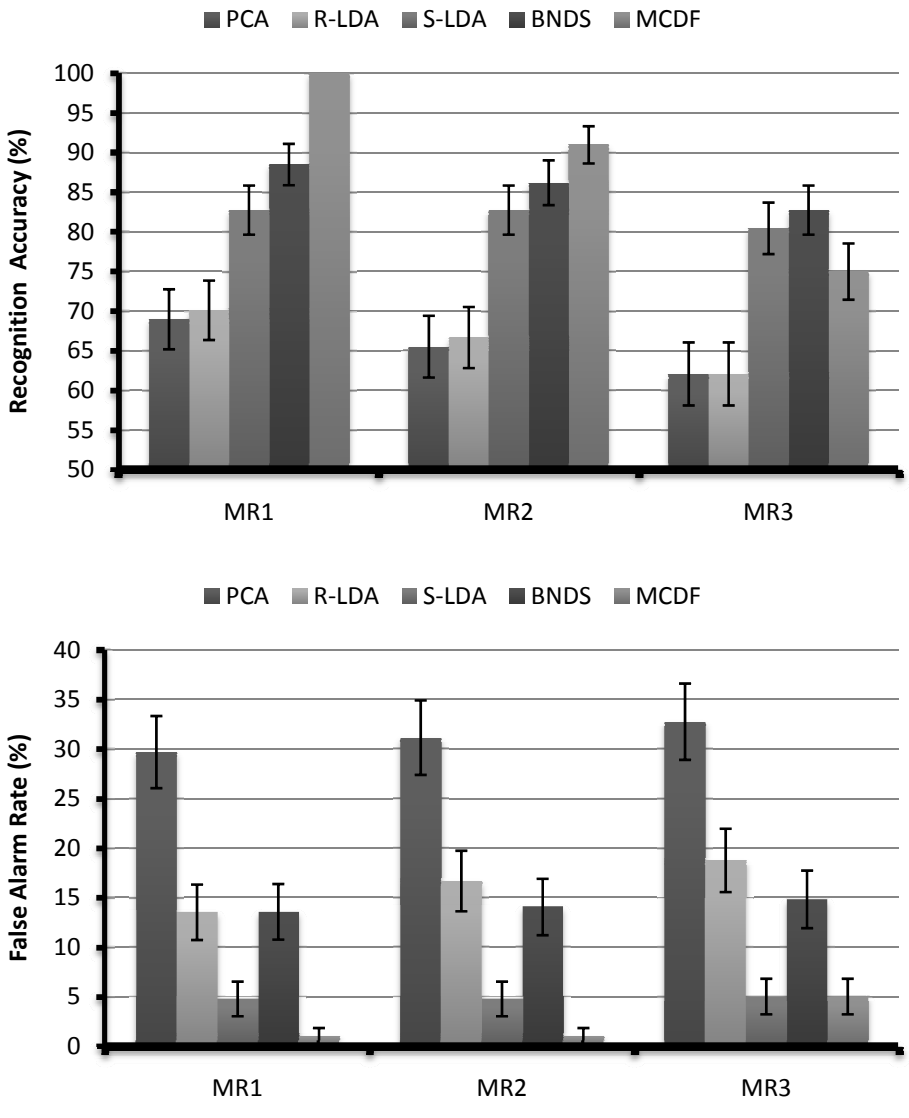


Figure 3.10 Comparison of the MCDF framework with current state-of-the-art.

3.6 Conclusions

In this chapter, a new classification framework for hyperspectral classification is proposed, based on spectral band grouping, multi-classifiers and decision fusion – providing a robust classification performance. The proposed classification system also

alleviates the small-sample-size problem commonly encountered in hyperspectral classification applications – since the training of supervised dimensionality reduction and classification algorithms is performed in many local subspaces, each of a much smaller dimension as opposed to the original high dimensional hyperspectral space.

A justification was provided for using higher order statistical information provided by a mutual information based metric for the subspace identification task. Experimental evidence was provided for the same. An adaptive weight assignment scheme was also proposed for the decision fusion process, which proved to be beneficial in scenarios where the hyperspectral signatures were of non-uniform fidelity across the spectrum. Experimental evidence was also provided to justify the choice of a LDA based dimensionality reduction scheme at the subspace level before invoking the multi-classifier and decision fusion systems. LDA based dimensionality reduction at the subspace level improved classification performance of the system under light and moderate mixing conditions. As pixel mixing increased, although the LDA based MCDF approach performed better than conventional approaches, there was room for improvement. It was also demonstrated experimentally that decision level fusion consistently outperforms feature level fusion. It was observed in these experiments that between hard and soft decision fusion approaches, hard decision fusion (MV and WMV) were most robust to poor signal fidelity (e.g., due to pixel mixing and additive noise). Also, among the soft decision fusion approaches, LOGP and WLOGP were least robust to poor signal fidelity. Hence, in the remainder of this dissertation, only MV and LOP based decision fusion results will be reported. Finally, classification performance of the proposed MCDF system was compared against the performance of conventional feature

extraction and optimization approaches that are currently employed for classification in small-sample-size conditions.

Although the proposed MCDF system provides a robust classification performance with experimental hyperspectral data, there is room for further improvement by using more sophisticated dimensionality reduction schemes. Kernel methods such as kernel principal component analysis, kernel discriminant analysis and support vector machines have recently gained popularity in many pattern classification tasks, and more recently, in remote sensing and hyperspectral image analysis tasks [25], [26], [27]. In the next chapter, kernel methods are incorporated into the proposed MCDF system to improve the system's classification performance when class conditional densities are multi-modal and decision boundaries are non-linear (for example, under severe pixel mixing conditions).

REFERENCES

[1] P. Watanachaturaporn, P.K. Varshney, M.K. Arora, "Multisource fusion for land cover classification using support vector machines", *Proc. 8'th International Conference on Information Fusion*, pp. 614 – 621, July 2005.

[2] B. Jeon, D.A. Landgrebe, "Decision fusion approach for multitemporal classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 37, no. 3, pp 1227-1233, May 1999.

[3] N. Memarsadeghi, J. Le Moigne, D.M. Mount, J. Morisette, "A new approach to image fusion based on cokriging," *Proc. 8'th International Conference. on Information Fusion*, pp. 25-28, July 2005.

[4] M. Fauvel, J. Chanussot, J.A. Benediktsson, "Decision Fusion for the Classification of Urban Remote Sensing Images," *IEEE Trans. Geoscience and Remote Sensing*, Vol. 44, no. 10, pp 2828-2838, 2006.

[5] J. Chanussot, G. Mauris & P. Lambert, "Fuzzy fusion techniques for linear features detection in multi-temporal SAR images," *IEEE Trans. Geoscience and Remote Sensing*, vol. 37 no. 3, pp 1292-1305, May 1999.

[6] G. Simone, A. Farina, F.C. Morabito, S.B. Serpico, L. Bruzzone, "Image fusion techniques for remote sensing applications," *Information Fusion*, vol. 3, pp. 3-15, 2002.

[7] J.A. Benediktsson, J.R. Sveinsson, "Multisource remote sensing data classification based on consensus and pruning," *IEEE Trans. Geoscience and Remote Sensing*, vol. 41, pp 932-936, 2003.

[8] M.D. Farrell, R.M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," *IEEE Geoscience and Remote Sensing Letters,* vol. 2, no. 2, pp. 192-195, April 2005.

[9] D.L. Swets and J. Went, "Using discriminating eigenfeatures for image retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-836, August 1996.

[10] A. Cheriyadat, L.M. Bruce, "Why principal component analysis is not an appropriate feature extraction method for hyperspectral data," *Proc. IEEE International Geoscience and Remote Sensing Symposium*, Vol. 6, pp 3420-3422, July 2003.

[11] S. Prasad, L.M. Bruce, "Limitations of Subspace LDA in Hyperspectral Target Recognition Applications," *Proc. IEEE International Geoscience and Remote Sensing Symposium*, July 2007.

[12] R. Pu, P. Gong, "Band selection from hyperspectral data for conifer species identification," *Proc. Geoinformatics'00 Conference*, Monterey Bay, pp 139-146, June 21-23.

[13] A. Cheriyadat, L.M. Bruce, A. Mathur, "Decision level fusion with best-bases for hyperspectral classification," *Proc. IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data,* pp 399-406,October 2003.

[14] S. Kumar, J. Ghosh, M.M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1368-1379, July 2001.

[15] S. Prasad, L.M. Bruce, "Information theoretic partitioning and confidence based weight assignment for multi-classifier decision level fusion in hyperspectral target recognition applications," *Proc. SPIE Defense and Security Symposium*, Orlando, Florida, USA, April 2007.

[16] S. Prasad, L.M. Bruce, "Hyperspectral feature space partitioning via mutual information for data fusion," *Proc. of IEEE Geoscience and Remote Sensing Symposium* , Barcelona, Spain, July 23-27, 2007.

[17] V. Tsagaris, V. Anastassopoulos, G. A. Lampropoulos, "Fusion of hyperspectral data using segmented PCT for color representation and classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 43, no. 10, pp. 2365-2375, October 2005.

[18] T. Cover, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, Inc., 2006.

[19] C. M. Jarque, A. K. Bera, "A test for normality of observations and regression residuals," *International Statistical Review / Revue Internationale de Statistique*, vol. 55, no. 2, pp. 163-172, August 1987.

[20] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[21] J. A. Benediktsson, J. R. Sveinsson, "Multisource remote sensing data classification based on consensus and pruning," *IEEE Trans. Geoscience and Remote Sensing*, vol. 41, pp 932-936, 2003.

[22] J. A. Benediktsson, P. H. Swain, "Consensus theoretic classification methods," *IEEE Trans. on. Systems, Man, and Cybernetics*, vol. 22, no. 4, pp. 688-704, August 1992.

[23] Analytical Spectral Devices FieldspecPro FR specifications. Available: http://asdi.com/productsspecifications-FSP.asp.

[24] P. A. Devijver, J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall, London, 1982.

[25] M. Fauvel, J. Chanussot & J.A. Benediktsson, "Kernel principal component analysis for feature reduction in hyperspectral image analysis," *Proc. 7th Nordic Signal Processing Symposium*, pp 238-241, June 2006, Reykjavik, Iceland.

[26] J. Munoz-Mari, L. Bruzzone, G. Camps-Valls, "A Support Vector Domain Description Approach to Supervised Classification of Remote Sensing Images," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, issue 8, pp. 2683 – 2692, Aug. 2007.

[27] Mingmin Chi, L. Bruzzone, "Semisupervised classification of hyperspectral images by SVMs optimized in the primal," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, issue 6, pp. 1870 – 1880, Jun. 2007.

CHAPTER IV

INFORMATION FUSION IN KERNEL INDUCED SPACES FOR ROBUST HYPERSPECTRAL CLASSIFICATION

4.1 Introduction

ATR systems based on remotely sensed hyperspectral images can suffer from the curse of dimensionality because the hyperspectral reflectance signatures often have a dimensionality much greater than the number of available training (ground-truth) pixels. Thus, conventional hyperspectral based ATR and land cover classification systems often project the high dimensional reflectance signatures onto lower dimensional subspaces before employing a classification algorithm [1]. As mentioned in previous chapters, in the context of hyperspectral classification and target recognition, some commonly used dimensionality reduction (or feature extraction/reduction) techniques are LDA, PCA, S-LDA and band selection [2], [3]. These techniques aim at avoiding the curse of dimensionality (commonly referred to as the Hughes phenomena in the remote sensing community), and reduce the amount of training data required for robust classification. Although these dimensionality reduction schemes successfully reduce the ground truth requirement for unbiased modeling by the classifier [1], these projections are not necessarily optimal from a pattern classification perspective [3]. For example, a PCA projection may discard useful discrimination information if it were oriented along

directions of small global variance, while an LDA projection will be inaccurate for multimodal class distributions. Another factor that governs the efficacy of such dimensionality reduction techniques is the amount of training signatures required to learn the projections. As an example, if the number of training signatures is insufficient for a given feature space dimensionality, the sample scatter and covariance matrices are likely to be ill-conditioned, and transformations such as LDA may not yield optimal projections.

In the previous chapter, a divide-and-conquer approach (MCDF) is proposed that addresses the small-sample-size problem by partitioning the hyperspectral space into smaller subspaces, and performing local pre-processing and classification in each subspace, followed by decision fusion to merge the local classification results for obtaining the final class labels of test signatures. Such an approach yielded a superior classification and recognition performance as compared to conventional dimensionality reduction techniques and alleviated the small-sample-size problem commonly encountered in hyperspectral data classification tasks. However, this approach had room for improvement, in that LDA was employed as the pre-processing procedure at the local subspace level to improve class separation before performing classification. This approach performs very well in pure pixel classification, or in mild mixing conditions (target pixels may be mixtures of target and non-target classes); however, performance significantly decreased when the target pixels were severely mixed with background pixels. In severe mixing conditions, the class conditional density functions are likely to be multi-modal, and the decision boundaries are hence likely to be highly nonlinear. Under such conditions, LDA is likely to fail, because it assumes uni-modal class

conditional distributions. Further, when a classification system is trained on "pure" data but tested with "mixed" data, it is important that the pre-processing transformation or the classification algorithm impart significant generalization in the overall classification procedure, to account for this statistical mismatch.

To address the inability of the previously proposed technique to reliably classify pixels in severe mixing conditions, in this chapter, a kernel based pre-processing at the subspace level is proposed, where the hyperspectral space is divided into smaller contiguous subspaces and a Kernel Discriminant Analysis (KDA) based projection is performed in each subspace before classification. Finally, local classification decisions from each subspace are merged using decision fusion. Kernel projections such as kernel PCA and KDA have recently become popular in many pattern classification tasks [6-9]. This chapter will explore the benefits of kernel projections in creating linearly separable class features which exhibit a stronger generalization capacity, and, Fisher's discriminant analysis in the kernel projected space for further improving class separation, all in the multi-classifier, decision fusion framework. In particular, the performance of the proposed system is studied in a difficult classification scenario - when one has pure training signatures or pixels, but mixed test signatures or pixels. This is a realistic scenario, since it is not uncommon to have "pure" training data of the target (e.g., acquired from hand-held sensors), but have mixed test pixels that need to be classified (e.g., acquired from satellite imagery with relatively poor spatial resolution.)

The outline of this chapter is as follows. In section 4.2, a brief summary of KDA and the implementation used in this work is presented. Section 4.3 provides a description of the multi-classifier decision fusion framework within which the KDA technique is

exploited, and the motivation for using KDA in this framework is presented. In section 4.4, a description of the experimental hyperspectral data employed for measuring the efficacy of the proposed system in target recognition tasks is provided. In section 4.5, experimental evidence is presented justifying the use of the proposed KDA based MCDF system. Section 4.6 concludes this chapter with a summary and discussion of experimental outcomes.

4.2 Discriminant Analysis in Kernel Induced Spaces

4.2.1 Conventional LDA

LDA seeks to find a linear transformation $\vec{y} = W^T \vec{x}$, where $\vec{x} \in \Re^m, \vec{y} \in \Re^n$ and $n \leq c - 1$, (c is the number of classes), such that the within-class scatter is minimized and the between-class scatter is maximized. The transformation W^T is determined by maximizing Fisher's ratio [10] which can be solved as a generalized eigenvalue problem. The solution is given by the eigenvectors of the following eigenvalue problem

$$S_w^{-1} S_b W = \Lambda W, \tag{4.1}$$

where S_b is the between-class scatter matrix and S_w is the within-class scatter matrix.

Note that $S_T = S_w + S_b$ is the total scatter matrix, which is related to the global covariance matrix by a scaling factor. Introductory discussion on LDA is provided here; the reader is referred to [10] for more discussion.

LDA assumes uni-modal class-conditional statistics. Hence, when class-conditional distributions are multi-modal (as is the case for mixed pixels), LDA transformations will not optimize class separation. Further, if class-conditional statistics vary in training and test conditions, linearly projected spaces will not guarantee a good

generalization capacity. For example, if the LDA transformation is learned from training data that has a linear decision boundary, but applied to test data that has a nonlinear decision boundary, the mismatch in training and test conditions will not be accounted for, and the classification of test data in the LDA projected space will be unreliable.

4.2.2 Kernel Discriminant Analysis

In kernel methods, the key motivation behind mapping data onto a higher dimensional space is to convert nonlinear decision boundaries in the input space into linear decision boundaries in the transformed space via an appropriate nonlinear kernel function [7]. The "kernel trick" allows for computation of algorithms in a kernel mapped space without explicitly evaluating the mapping, as long as the algorithm can be expressed in terms of dot products of vectors in the input space. In its most general formulations, the kernel trick states [7] that if an algorithm can be formulated in terms of a positive definite kernel, k_1, it is possible to construct an alternate algorithm by replacing k_1 by another positive definite kernel, k_2.

In machine learning applications, the most common use of the kernel trick involves a situation where the kernel k_1 is a dot product, although, the original formulation is not limited to this case. A positive definite kernel is also endowed with a reproducing property [7]. An example usage of the kernel trick in light of this property is as follows. Assume that an algorithm in the original (input) space can be represented entirely in terms of dot products of vectors in the input space, i.e., in terms of $< x, x' >$ where x and x' are vectors in the input space. Now consider a "kernel induced" space, created by mapping all points in the original space onto a higher (possibly infinite)

dimensional space - i.e., each vector x in the original space is mapped onto $k(.\ , x)$, a vector in the kernel induced space. The algorithm will still hold in this high dimensional kernel induced space. Further, the kernel trick and reproducing property can facilitate easy implementation of the algorithm in this space. To implement the algorithm in this kernel induced space, we need inner products of vectors in this space, $< k(.,x), k(.,x') >$. Instead of performing the mapping (from the input space onto the kernel induced space) explicitly and then evaluating inner products in the kernel induced space, the reproducing property allows us to replace these inner products by the values of the kernel function evaluated using vectors in the original space, $< k(.,x), k(.,x') >= k(x,x')$. For more explanation, and a more general formulation of the kernel trick and reproducing kernel Hilbert spaces, the reader is referred to [7].

Mika *et al* [8] extended the conventional Fisher's LDA technique to a high dimensional, kernel induced space by employing the kernel trick. Similarly, Baudat *et al* [9] proposed an alternative implementation to KDA, referred to as generalized discriminant analysis. In the kernel LDA setting, if Φ is a nonlinear mapping to a feature space F, the linear discriminant function that needs to be maximized is

$$J(w) = \frac{w^T S_B^{\Phi} w}{w^T S_W^{\Phi} w}, \tag{4.2}$$

where S_B^{Φ} and S_w^{Φ} are between-class and within-class scatter matrices [7] of the mapped training data in F, and w is a vector in F. If F is a very high dimensional space, obtaining a solution in the above formulation may become intractable. The solution proposed by Baudat *et al* [9] is as follows:

1) Evaluate the empirical kernel (Gram) matrix, K, as:

$$K_{ij} = \left\langle \phi(x_i), \phi(x_j) \right\rangle = k(x_i, x_j), \tag{4.3}$$

where $k(.,.)$ is the kernel function and $\{x_i\}$ is the set of all training data vectors.

2) Define a block diagonal matrix, W, as:

$$W = (W_l)_{l=1,2,\ldots,N}, \tag{4.4}$$

where W_l is an $(n_l x n_l)$ matrix with all entries equal to $1/n_l$. N here is the number of classes, and n_l is the number of samples in the l'th class.

3) Perform the eigenvalue decomposition of K as $K = P\Gamma P^T$

4) Compute the eigenvalues and eigenvectors (λ and β) of the system given by $\lambda\beta = P^T W P \beta$.

5) Compute $\alpha = P\Gamma^{-1}\beta$.

The projection of any point (z) in the input space that maximizes (4.2) in the kernel space can be obtained as

$$w^T \phi(z) = \sum_{i=1}^{M} \alpha_i k(x_i, z), \tag{4.5}$$

where $\{\alpha_i\}$ is the coefficient vector learned in the algorithm described above, M is the total number of training points $\{x_i\}$, and $k(.,.)$ is the kernel function. In this work, the algorithm described above is employed to perform the KDA projection on the feature space.

Such a KDA transformation can provide two key advantages in pattern classification tasks: (1) the kernel mapping onto the higher dimensional space F creates a linear class separation structure, which is easier to work with and provides a better generalization ability; (2) projection of data from the kernel space into a lower

dimensional space maximizes class separation which in turn ensures good classification performance in the KDA space. In scenarios where the original (input) space contains data that is already uni-modal and linearly separable, KDA may not prove significantly beneficial over conventional LDA. However, in scenarios where the class conditional distributions in the input space are multi-modal or are not linearly separable, discriminant analysis in the kernel space is likely to be beneficial. With this in mind, this technique was chosen as an alternate pre-processing transformation to ensure good classification performance of hyperspectral subspaces under severe mixed pixel (targets are sub-pixel) conditions.

Recently, Fauvel *et al* [6] employed kernel PCA for hyperspectral classification tasks. Although kernel PCA has previously produced promising results for some pattern classification tasks, like PCA, it is not designed to maximize class separation. LDA (in the original or kernel induced feature space) on the other hand, seeks to find a transformation that maximizes class separation as characterized by the Fisher's ratio. Hence, this work will employ KDA, and its benefits in robust hyperspectral target recognition will be studied.

4.2.3 Choice of Kernel and Kernel Parameters

The choice of kernel function and kernel parameters is expected to have a significant impact on the classification performance in a KDA projected space. The kernel function employed in this work is the Radial Basis Function (RBF) kernel, defined as [9]:

$$k(x_i, x_j) = \exp(-\left|x_i - x_j\right|^2 / \sigma^2), \tag{4.6}$$

where σ is a user defined parameter of the kernel. Although the key requirement for the kernel trick to hold is for the kernel function to be positive definite, the RBF kernel has been successfully applied in machine learning applications, such as in Support Vector Machine (SVM) implementations for pattern classification tasks. In various classification applications, this kernel function has resulted in induced spaces that result in a greater degree of generalization in learning decision boundaries. Further, this kernel function results in Kernel / Gram matrices that are full ranked [7]. This is a very important advantage over other kernels, because it ensures well-conditioned formulations of kernel based algorithms.

It has been pointed out in [7] that the value of σ (width of the kernel) governs the generalization of the decision boundaries learned in the kernel induced space. The larger this value, the better that classification algorithm would generalize to arbitrary test data, and vice-versa. In this chapter, classification performance of the proposed system will be studied over a wide range of this parameter space, in an attempt to identify appropriate parameter values for the classification task at hand.

4.3 KDA in a Multi-Classifier and Decision Fusion Framework

Although hyperspectral data provides a dense sampling of reflectance values across the spectrum, it does come at a price. The high dimensionality of hyperspectral data introduces two key challenges in classification tasks. (1) Small-sample-size problem: a high dimensional feature space necessitates a large amount of training (ground-truth) data for reliable statistical modeling of class-conditional distributions. (2) Hughes phenomenon: increased dimensionality typically reduces generalization ability of the

classification system because of "overtraining" of class-conditional statistics by the classifier. In other words, even if the classifier is able to successfully learn the statistics of the training data well, it may not be able to generalize well to test data bearing slightly different statistics. A significant "mismatch" between training and test conditions is likely to result in poor classification performance if the feature space dimensionality is high and conventional linear methods are employed for classification. Problem 1 was addressed in previous work [11] using a Multi-Classifier Decision Fusion (MCDF) framework. This chapter addresses problem 2 by introducing KDA in the MCDF framework.

4.3.1 A Multi-Classifier and Decision Fusion Framework for Classification

In the previous chapter, the MCDF framework was proposed for classification of hyperspectral data in small-sample-size conditions. In the MCDF framework, a high-dimensional hyperspectral space is partitioned into multiple contiguous subspaces, each of a much smaller dimension. An appropriate band-grouping algorithm is employed to identify these subspaces adaptively. Once these subspaces are identified, LDA based dimensionality reduction is carried out in each subspace. Finally, a bank of classifiers performs "local" classification in each of these subspaces independently, and the results from each classifier in this bank are merged into a final class label per pixel using an appropriate decision fusion rule. This framework results in robust classification of high dimensional hyperspectral data in small training sample size conditions. It was shown in the previous chapter and in recent publications related to this dissertation [11-13] that the MCDF framework outperformed other conventional single-classifier based paradigms, even when the training sample size was smaller than the dimensionality of the data.

However, under severe pixel mixing conditions, the MCDF approach started to break down, as did other conventional approaches. This was expected, because, the MCDF approach employed a LDA transformation per subspace, followed by a quadratic maximum-likelihood classifier for classification. Although LDA based pre-processing improves class separation, and hence classification performance of uni-modal data, it is not expected to perform well under pixel mixing, because class-conditional distributions are likely to become multi-modal with increasing pixel mixing. Further, since a simple linear projection (LDA) followed by a maximum-likelihood classification is employed per subspace in the MCDF approach, this framework is not expected to generalize well to severe mismatches in training and test conditions.

4.3.2 KDA in the MCDF framework

Kernel methods, including Kernel LDA and SVMs have recently shown to improve the classification generalization capacity [6], [8], [9]. In recent work [6], Kernel PCA has been employed for classification tasks, and has delivered promising results. In this work, PCA is not employed as a pre-processing of data because PCA is not designed for classification tasks. Instead, a Kernel Discriminant Analysis (KDA) approach is employed as a pre-processing in the MCDF framework to ensure robust classification, even under severe pixel mixing conditions.

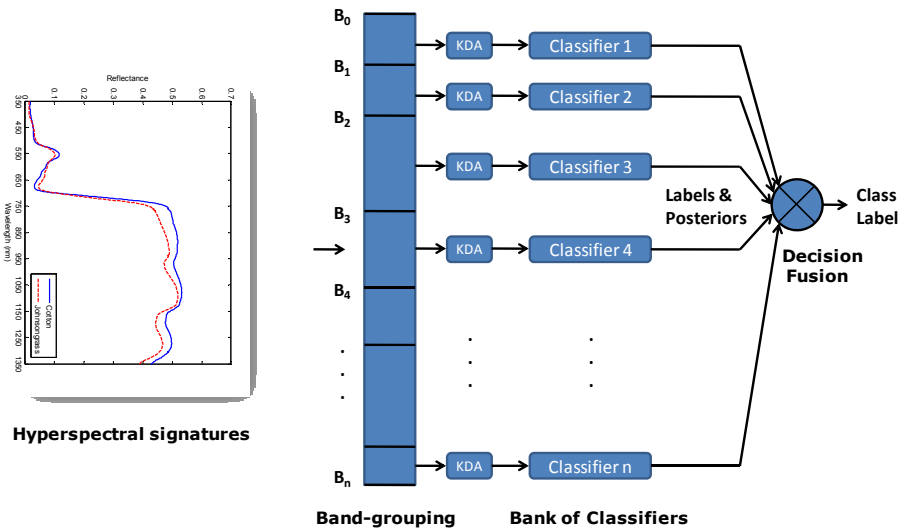


Figure 4.1 Illustrating the kernel based multi-classifier decision fusion framework.

Figure 4.1 illustrates the approach proposed in this chapter. The hyperspectral space is partitioned into multiple contiguous subspaces and training data is employed to "learn" the KDA transformation in each subspace (using (4.2) through (4.4)). This KDA transformation is then employed to project both training and test data to a reduced dimensional space. For a *c*-class classification task, the dimensionality of data in each subspace after the KDA projection would be *c*-1 [7], [10]. Hence, for a binary classification task, the dimensionality of data in each subspace after this projection is one. This reduced-dimensional training and test data is then employed to perform "local" classification of test data (signatures/pixels) in each subspace independently. Finally, a decision fusion mechanism is employed to merge these class labels into a single class label per test signature/pixel.

The motivation behind partitioning the hyperspectral space into multiple contiguous subspaces is similar to the explanation provided in [11] for employing the

MCDF framework. The key reason behind this partitioning is the observation that due to the dense spectral sampling in hyperspectral data, successive bands represent contiguous regions of the spectrum. . As a result, the resulting features (e.g. reflectance values) in these adjacent bands are highly correlated. To ensure a diverse collection of subspaces, adjacent bands (features) are grouped into subspaces such that the correlation (or mutual information) of bands within each subspace and the class separation of each subspace is simultaneously high. Different band-grouping metrics for this task were studied in [11] and chapter 3 of this dissertation. In this work, it was observed that unlike when LDA is employed in the MCDF framework [11], when KDA is used as a pre-processing at the subspace level, the choice of performance metric employed for band-grouping (/partitioning) of the hyperspectral space does not affect overall classification performance significantly. In other words, projections in kernel induced spaces at each subspace were so powerful in discriminating classes that the resulting MCDF system was not sensitive to intelligent partitioning of the hyperspectral space. Hence, in this chapter, a simple manual partitioning of the hyperspectral space into equal sized contiguous subspaces is performed. A window size (size of each group/subspace) equal to the dimensionality of the hyperspectral data will degenerate to the special case of a KDA based single classifier system. Although the KDA based MCDF system is not sensitive to the partitioning metric employed, it is sensitive to the size of each subspace, as will be explained later (in section 4.5.1).

4.3.3 Choice of classifier

In this work, a bank of quadratic maximum likelihood classifiers is employed for performing "local classification" on the KDA projection of each subspace. These classifiers assume Gaussian class distributions for the *i*'th class, $p(x|w_i) \sim N(\mu_i, \Sigma_i)$. Assuming equal priors, the class membership function for such a classifier is given by [14]

$$M(w_i \mid x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln\left|\Sigma_i\right|. \tag{4.7}$$

Other parametric or non-parametric classifiers such as *k* nearest-neighbors, neural networks, SVMs can also be employed in this bank of classifiers. However, it has been shown in [7] that data distributions in KDA projected spaces tend to be Gaussian in nature. In-fact, it has been shown in [7] that a KDA projection followed by conventional maximum-likelihood classification is as good (and in certain conditions, better than) as a SVM classifier. It is hence contended that quadratic maximum likelihood classifiers are a good choice in this setting.

4.3.4 Decision Fusion – Fusing "Local" Classification Decisions

After performing classification in local kernel induced subspaces, a decision fusion mechanism is needed to merge class label and posterior probability information from all individual subspaces into a final class label per test pixel/signature. Decision fusion can be either "hard" or "soft". Hard decision fusion involves fusion of individual class labels (hard-information). One popular example of such an approach is Majority Voting (MV) [11], [15], where for every test pixel, a voting mechanism is invoked over the results of all local classifiers in the bank of classifiers, and the signature is labeled as

belonging to the class that gets maximum number of votes. Soft decision fusion on the other hand entails the use of posterior probabilities, or more generally some class membership function from every classifier for making the final decision. Unlike hard fusion techniques, soft decision fusion schemes do not rely solely on class labels from each classifier to make the final decision. Two popular soft decision fusion schemes are linear and logarithmic opinion pools [11], [15]. A Linear Opinion Pool (LOP) uses the individual posterior probabilities of each classifier (j = 1, 2,… n), $p_j(w_i/x)$ to estimate a global class membership function

$$C(w_i \mid x) = \sum_{j=1}^{n} \alpha_j p_j(w_i \mid x), \qquad w = \underset{i \in \{1,2...C\}}{\arg\max}\, C(w_i \mid x). \tag{4.8}$$

The classifier weights (α_j, j = 1, 2, …n) can either be uniformly distributed over all classifiers, or can be assigned based on the confidence score of each classifier. This is essentially a weighted average of posteriors across the classifier bank. In a Logarithmic Opinion Pool, the global class membership function is modified to be a weighted product of the posterior probabilities of all classifiers, instead of a weighted sum. In this work, a uniform distribution of classifier weights is employed.

In previous work [11], it was observed that MV based decision fusion was least sensitive to signal fidelity conditions, while LOGP was most sensitive to signal fidelity and reliability of posterior probability estimates. When posterior probability estimates were reliable, LOP based decision fusion resulted in good classification in the MCDF framework. Hence, only results with LOP based decision fusion are reported in this chapter, although MV and LOGP were tested, and were found to be inferior to LOP.

4.4 Experimental Hyperspectral Data

Hyperspectral data was collected using an Analytical Spectral Devices (ASD) Fieldspec Pro FR handheld spectroradiometer [16]. Signatures collected from this device have 2151 spectral bands sampled at 1nm over the range of 350 – 2500nm with a spectral resolution ranging from 3 – 10nm. A 25° instantaneous field of view (IFOV) foreoptic was used. The instrument was set to average ten signatures to produce each sample signature, and the sensor was held nadir at approximately four feet above the vegetation canopy. Hyperspectral signatures collected with an ASD spectroradiometer tend to have high levels of noise in the regions associated with longer wavelengths, particularly when the sensor has been in use for a longer period of time or under high temperature conditions (due to overheating of the semiconductors). Thus the signatures were truncated at 1800nm. Also, the reflectance values in the regions 1350nm - 1430nm were removed from all signatures to avoid noise due to atmospheric water absorption. This resulted in hyperspectral signatures with 1600 bands.

Signatures in the dataset form two classes: (1) an agricultural row crop, Cotton variety ST–4961, and, (2) a weed that is detrimental to the crop's yield, Johnsongrass (*Sorghum halepense*). In this study, 54 signatures of Johnsongrass and 35 signatures of Cotton are used. These signatures were measured in good weather conditions in Mississippi, U.S.A., in 2000-2004. A target recognition scenario is simulated by employing the weed (Johnsongrass) as the target class, and the crop (Cotton) as the background class. Challenging target recognition tasks are created by linearly mixing target test signatures with the background at various mixing ratios (MR). All experiments reported in this chapter are performed using a leave-one-out testing procedure [17]. Each

test target signature sequestered during the leave-one-out testing is mixed linearly with a random background signature. To ensure an unbiased setup, the background signature used in this mixing is not used for training the system. This results in a realistic and challenging ATR problem because it creates a mismatched situation where the classifiers are trained on clean (unmixed) target and background signatures but tested on corrupt (mixed) target signatures. The mixing ratios (background percentage to target percentage) for test target signatures reported in this work range from 10:90 (very light mixing) to 90:10 (severe mixing). With this setup, target recognition accuracies are estimated from these sub-pixel ATR tasks using the proposed kernel based MCDF system.

4.5 Experimental Setup and Results

To determine the efficacy of the algorithm proposed in this chapter, ATR experiments are setup with the dataset described in section 4.4. Three sets of experiments are presented with the following objectives: (1) To study the effect of window size on the classification performance of the kernel based MCDF system, (2) To study the effect of the kernel parameter, σ on the generalization capacity of the proposed framework, (3) Benchmarking the recognition performance of the proposed system against other popular state-of-the-art classification methods under "light", "moderate" and "severe" pixel mixing conditions. Experiments 1 and 2 provide an understanding of the performance variation of the proposed system over the parameter space, and experiment 3 helps in quantifying the benefits of the kernel based MCDF system for classification over previous approaches, in particular, S-LDA, LDA based MCDF and single classifier KDA.

All experiments were conducted in the mixed pixel classification framework as described in section 4.4. This simulates a challenging and realistic scenario – which is commonly encountered when the size of the target is smaller than the resolution of the sensor, resulting in mixing of target signatures with background signatures. In this work, efficacy of the proposed algorithms is gauged using overall recognition accuracies (which measure the system's capacity to recognize both target and background signatures.) Further, for accurate estimation of overall recognition accuracies, all experiments were conducted using the leave-one-out cross validation method.

4.5.1 Experiment 1: Effect of Window Size on the Efficacy of the Kernel based MCDF System

As is expected, the window size, (the size of each group in the partitioning of the hyperspectral space) is likely to have an effect on classification performance of the KDA based MCDF system. In previous work [11], it was seen that correlation and mutual-information matrices of experimental hyperspectral data were approximately block-diagonal – with strong correlation among successive bands, and relatively weaker correlation between bands that were placed farther apart in the spectrum. When choosing an appropriate partition of the hyperspectral space, it was found necessary to ensure that the smallest size of the partition be no smaller than what would be supported by the approximate block-diagonal structure of the correlation matrix of the data. A smaller window/group size would mean that correlated bands would actually get partitioned into separate groups, and this would lower the diversity in the bank-of-classifiers, thereby weakening the decision fusion system [11]. Further, when LDA is employed as a pre-

processing at the subspace level [11], [12] instead of KDA, an upper bound on the size of each window/group is also needed to ensure that the within-class scatter matrices estimated in the LDA formulation are well-conditioned. If the size (dimensionality) of any group is larger than what would be supported by available training data, the within class scatter matrix may be ill-conditioned, resulting in sub-optimal performance of the MCDF system.

In the KDA based MCDF framework proposed in this chapter, we no longer have the restriction of an upper bound on each group to ensure a well-conditioned formulation. This follows from the fact that the KDA formulation employed in each group relies solely on the empirical Gram matrix, and the property that the Gram matrix is always full ranked (assuming it was learned from "distinct" training data points) when an RBF kernel is employed [7]. This implies that we can choose an arbitrarily large window size and the KDA based MCDF formulation would still remain well-conditioned.

However, increasing the window size indiscriminately in the partitioning process will affect the diversity of the bank of classifiers, and this may adversely affect the decision fusion performance. In experiment 1, the window size is varied over a wide range of values to study the variation in classification accuracy of the proposed system over this range. The value of the kernel parameter, σ was set to one in this experiment. Figure 4.2 shows the average absolute value of correlation between all the KDA projected subspaces in the partition. This value is a measure of the diversity of KDA induced subspaces, and hence of the classifiers in the bank-of-classifiers. Note that high redundancy (high correlation) implies a poor decision fusion performance [11], and high diversity (low correlation) implies a stronger and more robust decision fusion

performance. Figure 4.2 depicts this value (estimated from training data) for different window sizes, ranging from 10 to 800. The upper bound on the window size is 800 (i.e. the largest window size that results in uniform partitions). Figure 4.3 depicts the overall recognition accuracy of the proposed system for different window sizes, varying from 10 to 1600. Recall that when the window size is 1600 the KDA based MCDF system degenerates to a KDA based single classifier system. Because there is only one subspace in the partition when the window size is 1600, the window size is not varied all the way to 1600 for figure 4.2, where the purpose is to measure correlation between different subspaces in the KDA domain. However, comparing figure 4.2 and figure 4.3 for window size ranging from 10 to 800, a very definite trend can be seen. From window size 10 to window size 50, the average absolute correlation drops as the window size increases. In this range, the overall accuracy of the proposed system increases (for all mixing ratios). However, after a window size of 50, further increase in window size results in an increase in correlation and a drop in the overall classification accuracy. Classification accuracy and correlation between subspaces are inversely related, which can be explained by the fact that when the collection of classifiers is more diverse, the resulting decision fusion is more robust.

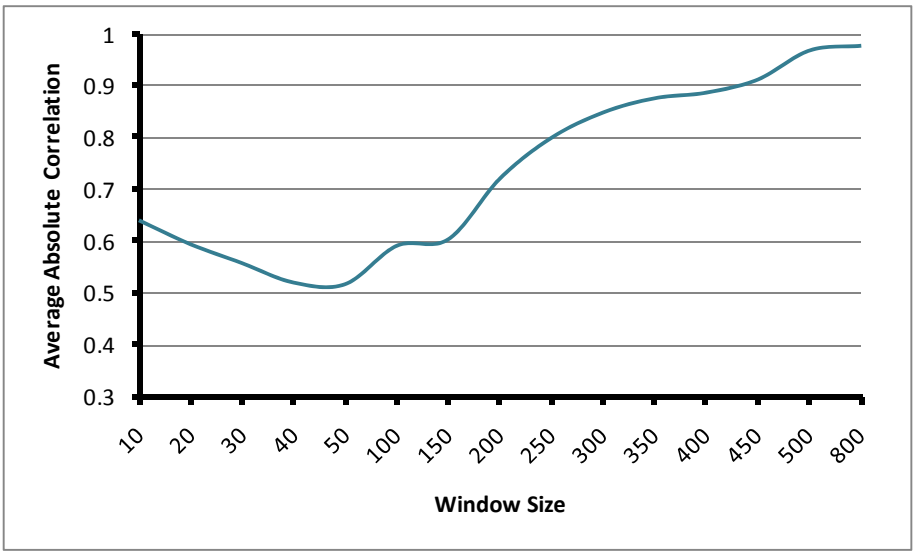


Figure 4.2 Average absolute value of correlation between KDA coefficients from each subspace for different window sizes.

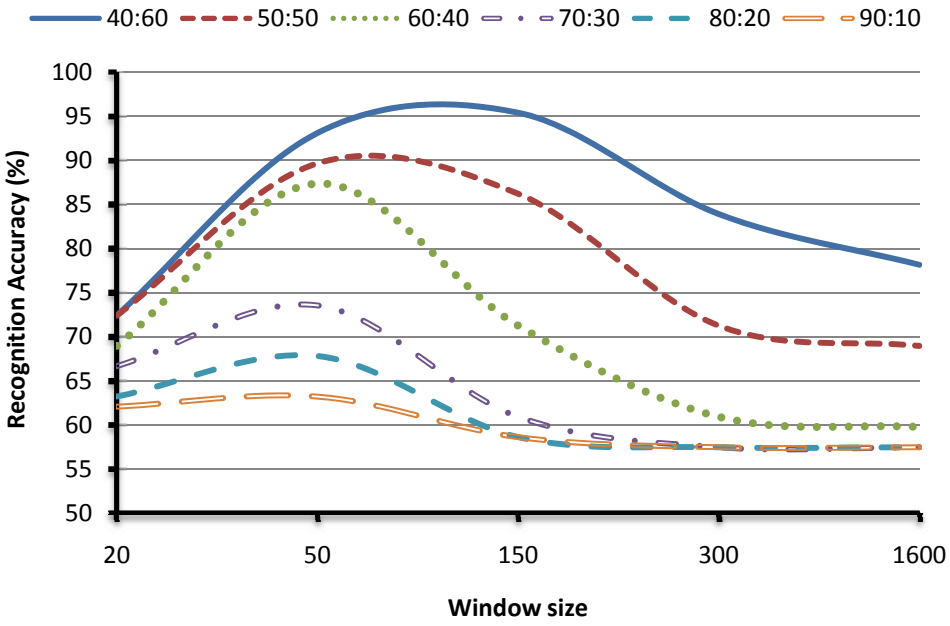


Figure 4.3 Accuracy vs. window size.

The large correlation between KDA spaces when the window size is small (e.g., between 10 and 50) is likely to be due to the fact that in that range, window size is smaller than what would be supported by the approximate block-diagonal structure of the correlation matrix of the data.

4.5.2 Experiment 2: Effect of Kernel Parameter on the Generalization Capacity of the Kernel based MCDF System

In this experiment, the generalization ability of the proposed system is studied as a function of the kernel parameter, σ. As was mentioned previously, the key motivation behind introducing a kernel based transformation in the MCDF framework is to improve the generalization ability of classification, that is, to ensure that the classification system is able to generalize well to arbitrary test data – even the kind that has a slightly different statistical structure as compared to the test data. This ensures a robust classification because in operational scenarios, it is rarely the case that we are able to train a classifier on data with spatial and spectral fidelity precisely similar to the actual test data.

In figure 4.4, overall classification accuracy is reported using the proposed KDA based MCDF system over a wide range of kernel parameter values, varying σ from 0.1 to 4.1. The window size in the partitioning process was set to 50. Results are reported for light pixel mixing (MR 10:90 and 20:80), moderate pixel mixing (MR 50:50) and severe pixel mixing (MR 80:20 and 90:10).

As explained in [7], the value of σ, the width of the RBF kernel has an impact on the generalization ability in the kernel induced space. As σ increases, the generalization capacity of a kernel based machine typically increases. Note that for light to moderate

pixel mixing conditions, the statistical structure of training and test data is very similar. This however is not the case for severe pixel mixing conditions, where not only the mismatch between training and test conditions is high, but with increased mixing, the class distributions are likely to be multi-modal in nature. This observation is reflected in the trends that can be seen in figure 4.4. For mild to moderate pixel mixing, overall accuracy increases with an increase in σ, obtaining the best classification accuracy at around σ = 0.6. However, a further increase in the parameter results in a drop in overall accuracy. For severe pixel mixing, it can again be seen that the overall accuracy increases

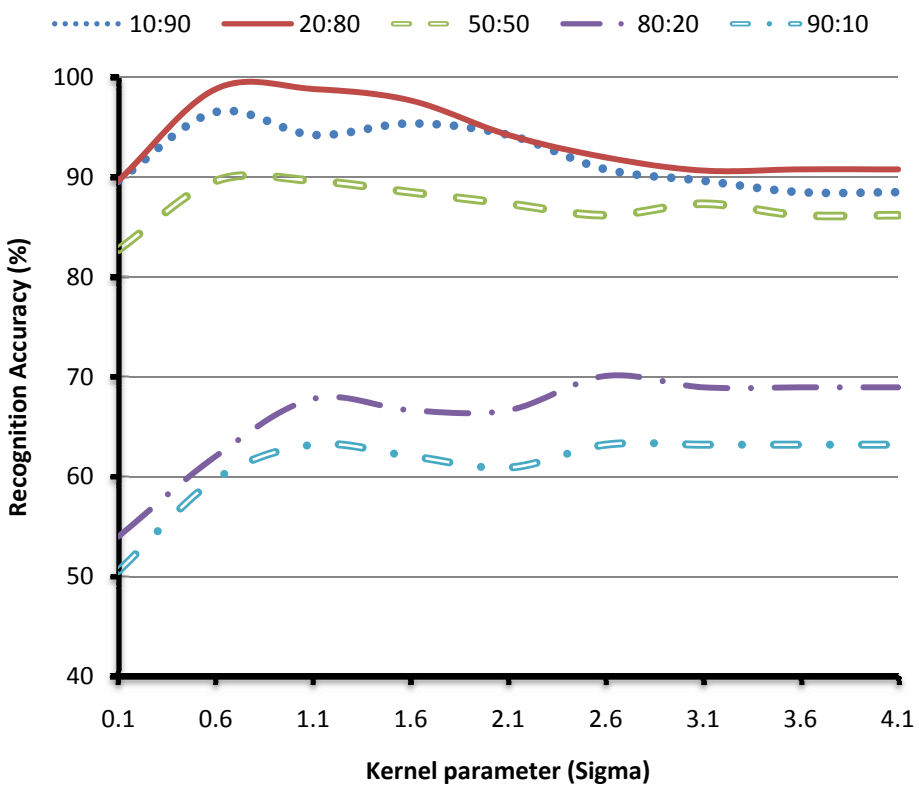


Figure 4.4 Accuracy vs. kernel parameter (sigma).

with increasing σ. Note that under severe pixel mixing, the maximum overall accuracy is attained with a relatively wide kernel (σ = 1) as compared to the mild and moderate pixel mixing case. This is due to the fact that under the severe pixel mixing case, more generalization (obtained by a wider kernel) is needed in the classification framework to account for multi-modality of class distributions and mismatch in training and test conditions.

From this figure, it follows that without any a-prior information about the extent of pixel mixing, a value of σ = 1 appears to be a good choice as the kernel parameter, as it provides high overall accuracy over a wide range of pixel mixing conditions.

4.5.3 Experiment 3: Benchmarking

The variation of performance of the proposed kernel based MCDF system with window size and kernel parameter provide us with valuable insight into the robustness and generalization capacity of this system as a function of the user-defined parameters. Based on the discussion in experiments 1 and 2, a window size of 50 and a kernel parameter σ = 1 is a reasonable choice for the type of experimental hyperspectral data employed in this chapter.

In this experiment, the recognition performance of the proposed system (using the parameters values: window size = 50 and σ = 1) will be compared against conventional state-of-the-art approaches for hyperspectral recognition. In particular, in this experiment, overall recognition accuracy will be compared in different pixel mixing conditions using (1) MCDF-KDA (the proposed system), (2) Single-KDA (employing a single KDA transformation on the entire hyperspectral space, followed by a single

maximum likelihood classifier), (3) MCDF-LDA (The multi-classifier and decision fusion framework using LDA as the pre-processing, instead of KDA), (4) S-LDA (Stepwise LDA), (5) Multi-KDA-FF (Feature fusion of multi-KDA projections, followed by a single classifier instead of a MCDF framework). S-LDA (also known as Discriminant Analysis Feature Extraction, or DAFE in the remote sensing community) is commonly employed by researchers in classification tasks when the training data size is small relative to the dimensionality of the data. It employs a forward selection and backward rejection algorithm to identify a smaller subset of available features (hyperspectral bands in this case) upon which a LDA transformation is applied. More details about this algorithm can be found in [18]. In this work, an area under Receiver Operating Characteristics (ROC) curve is employed to identify the smaller subset of hyperspectral bands upon which LDA is applied. This metric has previously shown to work well with hyperspectral data [18]. The size of the smaller subset upon which LDA is applied is chosen as 10, which is a reasonable value for the given amount of training data. Multi-KDA-FF still employs a partitioning of the hyperspectral space, followed by a KDA transformation in each subspace of the partition. However, the outcomes of KDA transformations from each subspace are not fed into a bank-of-classifiers, and instead are fused (concatenated) into one single feature vector per hyperspectral signature. Finally, a single maximum-likelihood classifier is employed for classification. This helps illustrates the benefits of decision fusion in the proposed MCDF-KDA system, instead of feature fusion.

Outcomes of these experiments for experimental hyperspectral datasets are depicted in figure 4.5. Note that in mild pixel mixing conditions (MR 10:90), the previously proposed MCDF-LDA system provides good classification accuracy. S-LDA and Single-KDA also perform well in these conditions. However, as pixel mixing becomes moderate (MR 40:60, 50:50) and severe (MR 60:40 and 90:10), the MCDF approach starts to break down. Performance of Single-KDA and S-LDA also starts to deteriorate. However, over this wide range of pixel mixing conditions, the proposed MCDF-KDA system to outperform other approaches (more so in moderate and severe pixel mixing conditions).

Finally, note that in all pixel mixing conditions, the feature fusion approach (Multi-KDA-FF) performs worse than most other approaches, and this clearly illustrates the benefit of having a multi-classifier and decision fusion system instead of performing feature fusion followed by a single classifier system for classification.

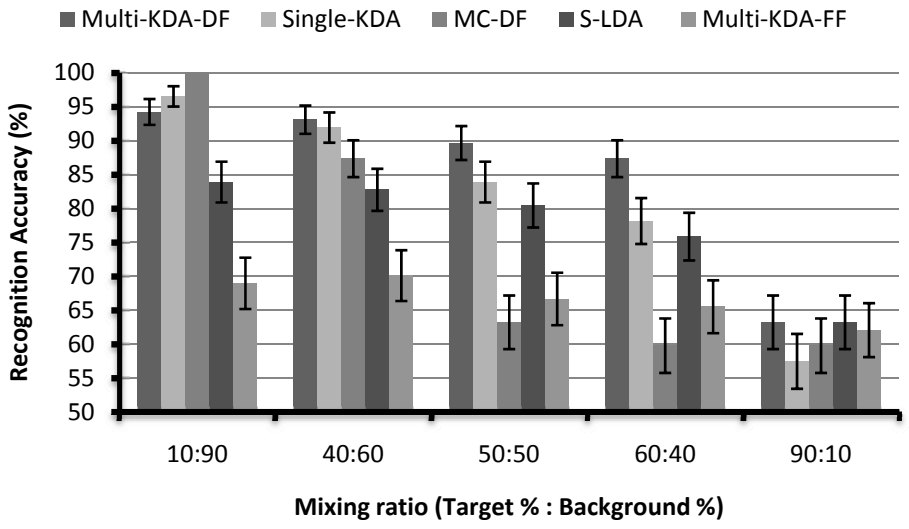


Figure 4.5 Accuracy at various mixing ratios for Cotton vs. Johnsongrass.

Bars atop each value indicate the 95% confidence interval.

4.6 Conclusions

In this chapter, a new kernel based multi-classifier and decision fusion framework is proposed for robust classification of high dimensional hyperspectral data. The proposed classification system has two significant advantages over conventional classification approaches: (1) It alleviates the small-sample-size problem commonly encountered in hyperspectral classification applications; (2) It ensures robust hyperspectral classification, even in severe pixel mixing and training-test 'mismatch' conditions. Although the previously proposed LDA based MCDF system alleviated the small-sample-size problem, the KDA based MCDF approach proposed in this chapter further ensures a robust classification in challenging classification scenarios.

Outcomes of experimental analysis in this chapter provided a justification for partitioning the hyperspectral space into smaller analysis windows, and for performing local KDA transformations and classifications in each window, instead of a single KDA transformation followed by a single classifier for classification. It was shown that the classifier diversity decreases with an increasing window size. A diverse ensemble of classifiers ensures a robust decision fusion based classification. This fact was exploited to determine the appropriate window size. Experimental results corroborated this observation, as it was noted that the overall accuracy of the MCDF-KDA system decreased as the window size was increased beyond the chosen window size.

The effect of varying the kernel parameter on the overall classification accuracy of the proposed system was studied. It was observed that increasing the value of the kernel parameter improved the generalization capability of the MCDF-KDA algorithm – i.e., the algorithm was robust even under severe training-test mismatch, but this came at

the cost of a slightly reduced performance in perfectly matched training-test conditions. After observing the overall accuracy vs. kernel parameter plots under various pixel mixing conditions, it was concluded that without any a-priori information about the extent of pixel mixing (or training-test mismatch), a kernel parameter value of 1 was a reasonable choice for the proposed system.

To conclude, the proposed MCDF-KDA algorithm provided a very robust hyperspectral classification performance for the given hyperspectral classification task, even with very little training data. This algorithm can be easily extended to any hyperspectral classification task, and the algorithm parameters (window size and kernel parameter) can be adapted to the dataset at hand by running experimental analysis similar to experiments 1 and 2 on the training dataset.

REFERENCES

[1] M.D. Farrell, R.M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," *IEEE Geoscience and Remote Sensing Letters*, vol. 2, no. 2, pp. 192-195, April 2005.

[2] R. Pu, P. Gong, "Band selection from hyperspectral data for conifer species identification," *Proc. Geoinformatics Conference*, Monterey Bay, pp 139-146, June 2000.

[3] S. Prasad, L.M. Bruce, "Limitations of Subspace LDA in Hyperspectral Target Recognition Applications," *Proc. IEEE International Geoscience and Remote Sensing Symposium*, July 2007.

[4] S. Prasad, L.M. Bruce, "Information theoretic partitioning and confidence based weight assignment for multi-classifier decision level fusion in hyperspectral target recognition applications," *Proc. SPIE Defense and Security Symposium*, Florida, USA, April 2007.

[5] S. Prasad, L.M. Bruce, "Hyperspectral feature space partitioning via mutual information for data fusion," *Proc. of IEEE Geoscience and Remote Sensing Symposium*, Barcelona, Spain, July 2007.

[6] M. Fauvel, J. Chanussot & J.A. Benediktsson, "Kernel principal component analysis for feature reduction in hyperspectral image analysis," *Proc. 7th Nordic Signal Processing Symposium*, pp 238-241, June 2006, Reykjavik, Iceland.

[7] B. Schlkopf, A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, December 2001.

[8] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K.-R. Muller. "Fisher Discriminant Analysis with Kernels," *Proceedings of IEEE Neural Networks for Signal Processing Workshop* 1999.

[9] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Proc. Neural Computation*, 2000.

[10] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Recognition*, Second Edition, Wiley-Interscience, 2000.

[11] S. Prasad, L.M. Bruce, "Decision Fusion with Confidence based Weight Assignment for Hyperspectral Target Recognition," *in IEEE Trans. Geoscience and Remote Sensing*, May 2007.

[12] S. Prasad, L.M. Bruce, "Overcoming the Small-Sample-Size Problem in Hyperspectral Classification and Detection Tasks," *in Proc. of IEEE Geoscience and Remote Sensing Symposium*, Boston, MA, July 2008..

[13] S. Prasad, L.M. Bruce, "A Robust Multi-Classifier Decision Fusion Framework for Hyperspectral Multi-Temporal Classification," *in Proc. of IEEE Geoscience and Remote Sensing Symposium*, Boston, MA, July 2008.

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[15] J.A. Benediktsson, J.R. Sveinsson, "Multisource remote sensing data classification based on consensus and pruning," *IEEE Trans. Geoscience and Remote Sensing*, vol. 41, pp 932-936, 2003.

[16] Analytical Spectral Devices FieldspecPro FR specifications. Available: http://asdi.com/productsspecifications-FSP.asp.

[17] P. A. Devijver, J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall, London, 1982.

[18] J.E. Ball, "Three stage level set segmentation of mass core, periphery, and spiculations for automated image analysis of digital mammograms." *Ph.D. in Electrical Engineering*. Starkville, MS: Mississippi State Univ., May 2007.

CHAPTER V

PRACTICAL APPLICATIONS OF THE MCDF FRAMEWORK

In this chapter, the MCDF framework is tested with three different practical classification tasks – (1) invasive species classification using satellite hyperspectral imagery, (2) multitemporal hyperspectral classification, and, (3) Computer Aided Detection (CAD) of malignant masses using digital mammogram images. In task 1, the previously employed MCDF framework is extended to satellite hyperspectral imagery, and its performance is compared to that from an S-LDA based feature extraction method. Results from this experiment will demonstrate that the MCDF framework can easily be extended to different hyperspectral sensors for robust statistical classification in small-sample-size conditions. In task 2, the MCDF framework is extended to a two-tier multi-classifier, decision fusion framework, wherein decision fusion is performed across both temporal and spectral dimensions for robust classification of multitemporal hyperspectral datasets. In task 3, the MCDF framework is tested on an entirely different classification task – as the classification backend of a digital mammography based CAD system. Current state-of-the-art CAD systems for mammography employ very high dimensional feature spaces for classification, and hence, it is expected that the MCDF framework will improve the robustness of such CAD systems. Results from these experiments

demonstrate that the MCDF framework can be extended to different high-dimensional, small-sample-size statistical pattern classification problems.

The outline of this chapter is as follows. In section 5.1, the MCDF framework is tested on an invasive species classification task. In section 5.2, an extension of the MCDF framework is proposed to robustly classify multitemporal hyperspectral data. In section 5.3, the MCDF framework is tested on an entirely different statistical pattern classification task – CAD of malignant and benign tumors using digital mammography.

5.1 Practical Application 1: Invasive Species Classification using Satellite Hyperspectral Data

Nonnative invasive species adversely impact ecosystems, causing loss of native plant diversity, species extinction, and impairment of wildlife habitats. Dispersal is a key issue in invasive species, because most invasive species disperse readily. During times of climate change, new invasive species may disperse into novel climate regions. The manner in which an invasive species will respond to climate change will vary according to the life history requirements of the species, its current range, its ability to disperse, and the conditions under which it can regenerate. Managers need to be on the alert for new threats by invasive species if climates change, and they must be ready to respond to situations as they arise.

Over the past decade federal and state agencies and nongovernmental organizations have begun to work more closely together to address the management of invasive species. In the 2005 fiscal year, approximately $500M was budgeted by U.S. Federal Agencies for the management of invasive species [1]. Despite extensive

expenditures, most of the methods used to detect and quantify the distribution of these invaders are ad-hoc, at best. Likewise, decisions on the type of management techniques to be used or evaluation of the success of these methods are typically non-systematic. More efficient methods to detect or predict the occurrence of these species, as well as the incorporation of this knowledge into decision support systems, are greatly needed.

In an attempt to demonstrate that the MCDF framework proposed in this dissertation will hold for data acquired from space-borne hyperspectral sensors as well, in this experiment, conventional classification approaches will be compared with the proposed MCDF approach for an invasive species detection and classification problem, using data acquired from the HYPERION imager [2].

HYPERION is a push-broom imager aboard the NASA Earth-Orbiter-1 mission satellite [2]. It possesses 220 spectral bands covering the spectrum from 400nm - 2500nm, each with a spectral resolution of 10nm. It has a swath of 7.5Km and a spatial resolution of 30m. Signatures in this dataset form two classes: (i) 115 samples of Tamarisk (*Tamarix ramosissima*), and (ii) 65 samples of Non-Tamarisk (a collection of native vegetation signatures in the vicinity, such as those of cottonwood, willow etc.). Figure 5.1 depicts some images and sample signatures from this dataset.

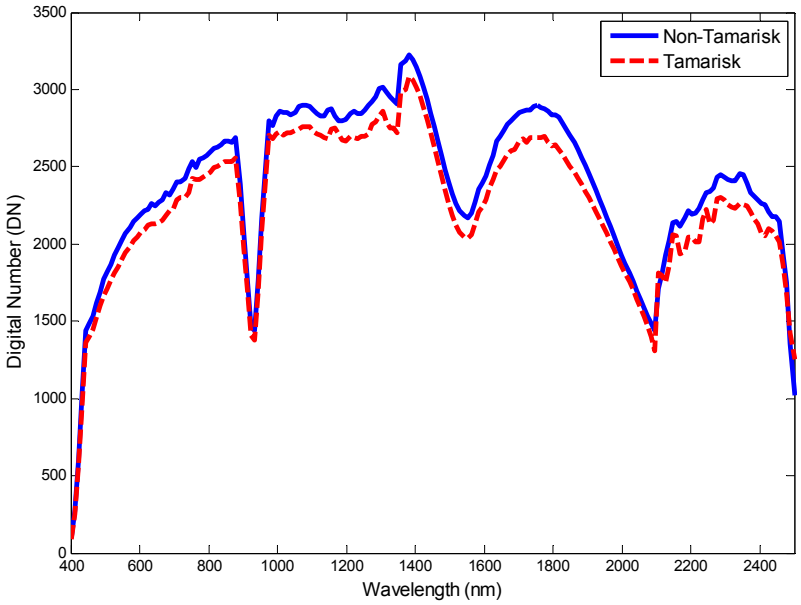


Figure 5.1 Experimental hyperspectral data.

Top: Tamarisk stand in Colorado. Native plants are unable to penetrate the thick stands of tamarisk [Photograph: Tim Carlson]; Bottom: Hyperspectral signatures from the dataset.

Tamarisk, an invasive species also known as salt cedar is a particular problem in the U.S.'s desert southwest, where it is displacing the native cottonwood, willow, and other native plants. Tamarisk shrubs, or trees, are extremely competitive against native vegetation because they aggressively consume the water supply. Since tamarisk can re-grow from root crown buds, even after burning, the current management practices for tamarisk involve combinations of chemical, mechanical, and biological techniques. Thus, detection of tamarisk through the use of remote sensing could greatly reduce the cost associated with this invasive species.

In this experiment, a Tamarisk vs. Non-Tamarisk classification is carried out, for accurate identification and mapping of Tamarisk invasion among other native vegetation. Due to limited ground-truth, a leave-one-out cross-validation is carried out, as described previously. In table 5.1, classification accuracy from two systems is reported – (i) S-LDA (current state-of-the-art), and, (ii) MCDF. The length to which the intermediate feature vector is allowed to grow in the S-LDA procedure is set to 10. In the MCDF framework, *JMCorr* and *JMAMI* are separately employed for band-grouping to study possible benefits of employing a mutual information based metric for satellite hyperspectral data. LDA based pre-processing is carried out in each subspace, followed by quadratic maximum-likelihood local classification. Finally, these local decisions are fused by a decision fusion mechanism. Results with both hard (MV) and soft (LOP) decision fusion are reported.

Table 5.1 Comparing classification performance of the conventional S-LDA technique, with that of the proposed MCDF technique for satellite hyperspectral data.

S-LDA (Baseline)	*MCDF – MV*		*MCDF – LOP*	
67.7 (3)	*JMCorr*	*JMAMI*	*JMCorr*	*JMAMI*
	76.4 (2)	**76.4 (2)**	**75.8 (2)**	**75.8 (2)**
	MCDF - WMV		*MCDF – WLOP*	
	JMCorr	*JMAMI*	*JMCorr*	*JMAMI*
	76.4 (2)	**76.4 (2)**	**74.7 (2)**	**74.8 (2)**

Values below are overall classification accuracies (expressed in percentage), and the 95% confidence interval for each value is provided in parenthesis.

Note that by employing the divide-and-conquer (MCDF) approach instead of the baseline S-LDA approach, the overall classification accuracy improved by 8 to 9%. Further, for this dataset, where the spectrum is not as densely sampled compared with handheld ASD data, employing a mutual information based metric (*JMAMI*) for band-grouping does not yield any advantage. Finally, the "weighted versions" of the decision fusion schemes, namely, WMV and WLOP did not provide any additional benefit either, as is expected for this dataset owing to the near uniform fidelity of the hyperspectral signatures.

The benefits of a kernel projection (KDA) at the subspace level in the MCDF framework are evident from results of a sub-pixel classification task using the HYPERION dataset (Figure 5.2). As before, a sub-pixel classification task is simulated

by mixing target test pixels with random background (non-target) pixels, and classification is conducted in the MCDF framework using (1) LDA as a pre-processing (MCDF-LDA), and, (2) KDA as a preprocessing (MCDF-KDA), at the subspace level. The KDA algorithm used in this experiment is explained previously in chapter 4. LOP is employed as the decision fusion rule. Results are reported for three different pixel mixing ratios (background : target percentage) – MR1 (10:90), MR2 (20:80), MR3 (30:70). It can be seen that incorporating a KDA transformation in the MCDF framework improves its generalization ability – the classification accuracy in severe pixel mixing conditions is substantially higher for MCDF-KDA as compared to the MCDF-LDA approach.

It is important to point out here that unlike the handheld ASD data, for space-borne and air-borne data, a mixing ratio of 30:70 is quite severe, because the pixels

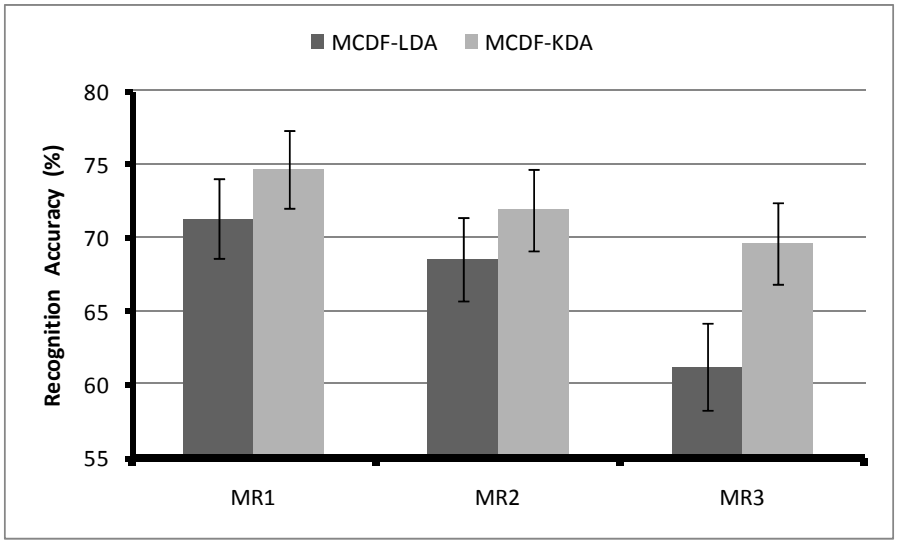


Figure 5.2 Overall recognition accuracy for the Tamarisk vs. Non-Tamarisk using Hyperion imagery, at different pixel mixing ratios.

Mixing ratios are: MR1 (10:90), MR2 (20:80), MR3 (30:70). Error bars indicate the 95% confidence interval of the classification accuracy.

already have much lower spatial resolution, and also suffer from atmospheric effects. It should also be noted that despite the fact that the MCDF based approach results in better classification accuracies as compared to conventional approaches, the best classification accuracy obtained with this dataset (76.4%, in table 5.1) is still lower than the very high classification accuracies obtained with handheld ASD data.

5.2 Practical Application 2: Multitemporal Hyperspectral Classification

Up to this point, experiments in this dissertation have exploited the MCDF framework towards the goal of robust classification of hyperspectral data. In this section, it will be demonstrated that this approach can be extended to a more generic data fusion scenario. In particular, an extension of the MCDF approach will be presented that will exploit both temporal and spectral information for robust classification of an invasive species classification dataset, i.e., decision fusion will be performed across the spectral and temporal dimensions.

Employing temporal information along with spectral information is expected to result in better classification performance, as compared to employing temporal information alone, or spectral information alone. In classification of vegetation species, spectral data can provide useful information about the cell structure, water stress, and other such biophysical characteristics, while the temporal evolution of this spectra can provide useful phenological information. However, just like hyperspectral classification tasks, multitemporal hyperspectral classification problems are also prone (in fact, even more so) to over-dimensionality of features and small training sample size problems.

Previously, Jeon and Landgrebe [3] have proposed a decision fusion system to classify multispectral, multitemporal data. The proposed majority vote based decision fusion technique improved the classification accuracy of three remotely sensed Thematic Mapper (TM) datasets, over a simple pixel-wise classifier that did not perform multitemporal decision fusion. Although this framework provided good recognition performance with multispectral data, extending this approach to hyperspectral data would necessitate addressing the inherent over-dimensionality of hyperspectral signatures in an appropriate manner. In this section, an extension of the MCDF approach [4] is presented to enable robust multitemporal, hyperspectral target classification. In the proposed approach, MCDF systems will be employed for each date in the dataset, and a second tier of decision fusion system will merge these results for final classification. The use of the divide-and-conquer approach per date ensures that for every date, the classification system is robust, even when working with a small-training-dataset. Finally, the fusion of class labels from every date further ensures that information from every time stamp in the temporal sequence is employed in the decision making process. The performance of the proposed system will be compared against that of conventional techniques, such as LDA and S- LDA. The efficacy of the system will be quantified by means of overall recognition accuracies.

Figure 5.3 depicts the overall block diagram of the proposed system. As in [4], the proposed framework incorporates a subspace identification procedure to partition the hyperspectral space into multiple contiguous subspaces and then employs a decision fusion mechanism to fuse local classification results from each subspace. However, in the proposed multitemporal hyperspectral system, multiple banks of classifiers and decision

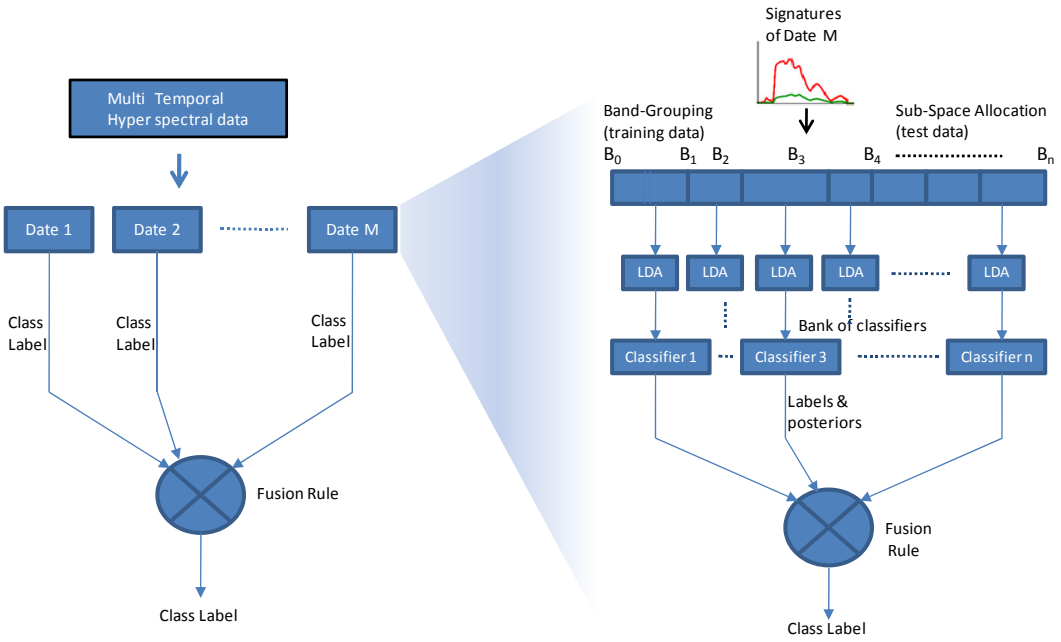


Figure 5.3 Illustrating the proposed system for robust classification of multitemporal hyperspectral data.

For every date, a band-grouping / subspace allocation procedure followed multi-classifiers and decision fusion is employed. Decisions from every date are merged into a final class label.

fusion systems are employed – one for each date. Finally, a global decision fusion system merges classification results from each bank (date) into a final class label per test pixel or signature. The resulting system is capable of providing reliable classification of test data even when relatively few training samples are available for each date.

It has been shown in chapter 3 that a divide-and-conquer approach that partitions the hyperspectral space and employs multi-classifiers and decision fusion for classification is robust in small-sample-size conditions. In this work, as in chapter 3, a metric based on the product of maximum Jeffries Matsushita distance and mutual information is employed for such partitioning of the hyperspectral space of every date.

This ensures that each subspace created in the process possesses good class separation, while the collection of subspaces is simultaneously diverse.

5.2.1 Experimental Hyperspectral Dataset

The multitemporal hyperspectral data used in this study was collected using an Analytical Spectral Devices (ASD) Fieldspec Pro FR handheld spectroradiometer [5]. Signatures collected from this device have 2151 spectral bands sampled at 1nm over the range of 350 – 2500nm with a spectral resolution ranging from 3 – 10nm. A 25° IFOV foreoptic was used, the instrument was set to average ten signatures to produce each sample signature, and the sensor was held nadir at approximately four feet above the vegetation canopy. Reflectance values in the regions 1350nm - 1430nm and 1800nm – 1980nm were removed from all signatures and then interpolated using piecewise cubic Hermite interpolation, to remove effects of atmospheric water absorption.

Figure 5.4 illustrates the experimental dataset. Signatures in this dataset form two classes: (i) an aquatic invasive species, Waterhyacinth (*Eichornia crassipes*), and (ii) another aquatic species, American lotus (*Nelumbo lutea*). A possible remote sensing application for such species may involve detecting and mapping Waterhyacinth in aquatic environments for appropriate chemical treatment and removal. The two aquatic species were grown under well-regulated environmental conditions at the R. R. Foil Plant Research Center at Mississippi State University. Data was collected in the range of ±2 hours of solar noon, every week from 24th June 2005 to 26th October 2005, for a total of twenty signatures per class per date [6].

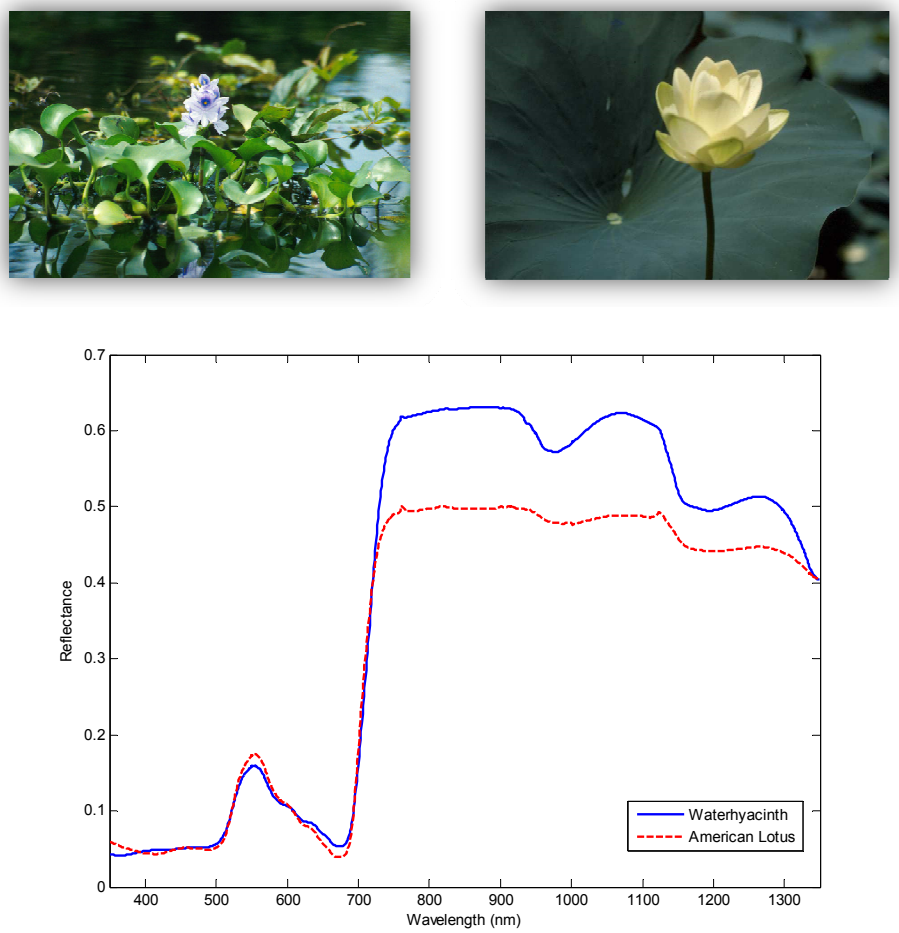


Figure 5.4 Experimental hyperspectral data.

Top left: Waterhyacinth (Ted Center, USDA); Top right: American Lotus (Robert H. Mohlenbrock, USDA); Bottom: Hyperspectral signatures of the two aquatic species.

5.2.2 Experimental Setup and Results

Target recognition experiments were carried out on hyperspectral data described above. All experiments reported here are performed using a leave-one-out testing procedure. Three different baseline experiments are reported, for comparing the performance of the proposed system with. The first baseline approach, called MT-SLDA employs a S-LDA [7], [8] approach per date, followed by a feature-space concatenation

(of the S-LDA output from every date) and a single classifier at the backend. In this approach, hyperspectral data from each date was reduced by means of a S-LDA (also known as Discriminant Analysis Feature Extraction, DAFE [8]) procedure. S-LDA employs a forward selection and backward rejection approach to choose a smaller subset (in this work, set as 10) of original features to apply LDA upon. The metric employed for forward selection and backward rejection in this experiment was the area under the Receiver Operating Characteristics (ROC) curve [7], [8]. This resulted in a reduced feature space dimensionality of one per date (since we only have two classes in this recognition task). Finally, the reduced dimensional space (in this case a scalar) is merged across all dates, and we come up with a single feature vector, which is of a much smaller dimensionality than the original multitemporal hyperspectral feature space. In this case (for a two-class problem), the dimensionality of this feature vector equals the number of dates in the multitemporal dataset. The S-LDA transformations per date are learned from training data, and applied to both training and test data samples. This is followed by a single classifier system.

In the second baseline approach, referred to as LDA-DF, we combine hyperspectral signatures from all dates into one single dataset, getting rid of the temporal information in the dataset by discarding date information, and randomly permuting the samples of each class. Finally, instead of using a conventional single classifier approach, we employed a multi-classifier, decision fusion approach, where the hyperspectral space was partitioned into multiple smaller subspaces, and LDA followed by classification was performed in each subspace independently. Finally, the local classification results (from each subspace/subset of the spectrum) were merged using decision fusion.

The third baseline approach, referred to as LDA also involves merging the hyperspectral signatures of all dates, that is, getting rid of the temporal information in the dataset by discarding date information, and randomly permuting the samples of each class. Since in this scenario, the number of training samples is sufficiently large for a conventional LDA transformation to be estimated, an LDA based projection of training and test data (learned from the training data) is performed, followed by a single classifier system. This approach quantifies the ability of conventional classification systems to classify the available hyperspectral data without using any temporal information. This approach is referred to as LDA in table 5.2.

The three baseline systems described above were compared against the proposed multitemporal decision fusion system illustrated in figure 5.3. All classifiers employed in this work were maximum-likelihood classifiers, assuming Gaussian class-conditional distributions [7], [9]. Majority voting [4] was employed as the decision fusion scheme in all of this work.

Experimental results with the proposed algorithm and the three baseline systems are provided in table 5.2. The accuracy reported is the overall recognition accuracy, along with the 95% confidence interval, both expressed in percentage. Figure 5.5 depicts the individual accuracies per date – i.e., using a multi-classifier, decision fusion system for each date separately. Note that although the individual classification accuracies (figure 5.5) are all less than 100%, and vary from approximately 65% – 95%, the overall accuracy of the proposed system (MT-DF), which is comprised of a two level decision fusion (spectral and temporal) is a 100% (table 5.2). This illustrates the fact that when the temporal information is added to the spectral information (by fusing results obtained over different dates), the corresponding classification accuracy for this two-class recognition task improved significantly. From table 5.2, it is clear that the proposed MT-DF system significantly outperforms the other "baseline" approaches to classification. A LDA-DF system, which discards the temporal information, but employs a multi-classifier and decision fusion framework for classification using spectral information is the next best

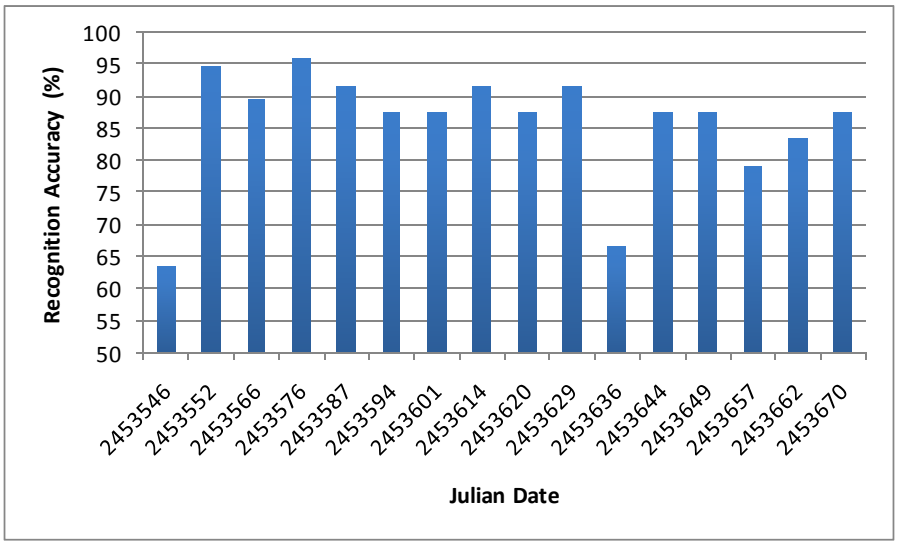


Figure 5.5 Individual classification accuracies per Julian date, using the multi-classifier decision fusion system on the hyperspectral signatures.

Table 5.2 Overall recognition accuracy for the multitemporal, hyperspectral task using the proposed approach (MT-DF) and three baseline approaches.

	Recognition Accuracy (%)	**Confidence Interval (+/- %)**
MT-DF	100	0
LDA-DF	96.8	0.8
LDA	91.5	1.3
MT-SLDA	86.6	3.2

system. The next best system is the LDA system, which also discards the temporal information, but employs a conventional single-classifier approach for the classification task. Finally, for this classification task, the MT-SLDA system provides the lowest classification accuracy.

In this experiment, the recognition performance of the proposed multi-temporal, hyperspectral decision fusion system was compared with various baseline approaches. It was established that despite the classification performance of each individual date in the dataset being relatively low, when the "temporal" information was exploited in the proposed framework, the classification accuracy for this two-class problem rose to a 100%. This demonstrates one of the many possibilities of multi-source data fusion within which MCDF framework can be incorporated. Hyperspectral information can similarly be combined using a decision fusion framework with other modalities, such as RADAR, LIDAR, spatial information etc.

5.3 Practical Application 3: Robust Classification of Mammogram Images

In previous chapters, the proposed MCDF framework is designed and tested with hyperspectral data, which is typically comprised of a dense "near-continuous" sampling of the spectra. This helps ensure that the adjacent bands of hyperspectral data are highly correlated, thereby allowing for a bottom-up band-grouping approach to identify subspaces in the proposed MCDF framework. While there are many applications which would use such datasets, there exist other datasets, where the feature vectors may not possess a well-defined correlation structure to warrant an intelligent and automated band-grouping. It would become necessary in such circumstances to adapt the feature-grouping (subspace identification) technique to the dataset at hand. One example of such a high-dimensional classification task is the discrimination of malignant and benign tumors in digital mammography CAD systems. In this experiment, the MCDF system will be applied to features extracted from the front-end of a digital mammography CAD system, for robust classification of mammogram images, i.e., identifying benign versus malignant masses in the images.

Breast cancer is believed to be among the leading causes of cancer related deaths among women, and mammography is the modality of choice for detecting breast cancer [10], [11]. As is the case with many medical imaging modalities, a great amount of research is being conducted for the design of CAD systems. A typical CAD system performs the following tasks in succession: (1) image enhancement and segmentation, (2) feature extraction, and, (3) classification. Robust image enhancement and segmentation algorithms are now available for identifying regions of interest in mammogram images [12]. The features extracted from these segmentations are however oftentimes very high

dimensional; for example, some CAD systems result in hundreds or even thousands of features [8], [16], [17], [18]. This has ramifications on the performance of the backend classification system in that the size of available training data (number of available training mammograms) does not match the required size needed to accurately model the statistical characteristics of high dimensional feature spaces.

In this experiment, we propose employing the divide-and-conquer (MCDF) approach to robustly classify mammogram images from very high dimensional feature spaces generated using state-of-the-art image enhancement and segmentation and feature extraction algorithms. The proposed approach partitions the high dimensional feature space into many smaller dimensional subspaces. A bank of classifiers (multi-classifier system) performs "local" classification in each such subspace, and an appropriate decision fusion system "fuses" these local classification results into a final malignant/benign classification for every mammogram image. In doing so, the proposed system is employing all the available information for classification while avoiding the problems of overly high dimensional feature spaces, and hence it is expected that the system will more accurately classify malignant and benign mammogram images.

5.3.1 Mammography Background

Most end-to-end CAD systems follow a three step approach to classifying mammogram images. Mammograms are first pre-processed and enhanced (e.g. contrast enhancement) to remove noise and improve contrast. These images are then segmented to extract the "region of interest". Features pertinent to the malignant/benign classification task are extracted from these segmentations, and are then optimized before being input to

a classification system. The classification system builds models of feature spaces for malignant and benign classes using available training data and uses these to perform classification, i.e. assign a label of benign or malignant to future input cases.

Some common mammographic image enhancement methods include adaptive neighborhood contrast enhancement, gamma correction and wavelet and multi-scale enhancement. These methods are described in detail in [8]. In this study, for the periphery segmentation, image enhancement consists of contrast limited adaptive histogram equalization (CLAHE), as well as custom non-linear methods described in [8]. Figure 5.6 illustrates sample mammographic images from the experimental dataset.

Image segmentation in the context of mammogram images typically seeks to identify a "region of interest" in the image that is most likely to provide useful

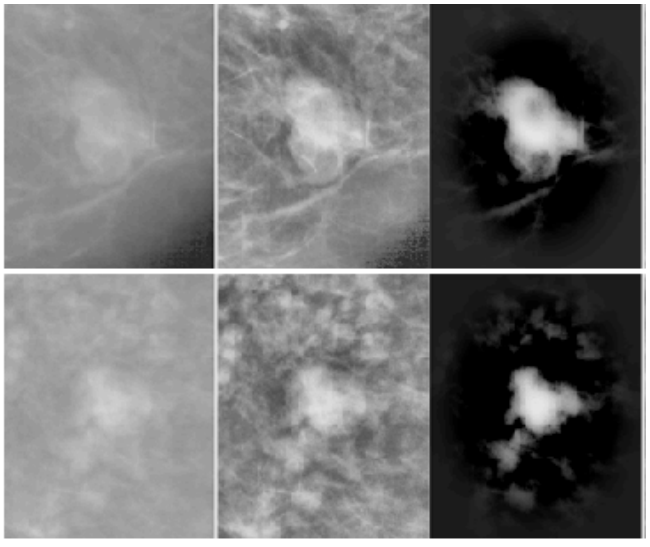


Figure 5.6 Sample images from the DDSM database.

Figure (left to right) shows the original mammograms, images after image enhancement to remove noise and further enhancement using Gaussian isotropic filtering [8]. Segmentation and feature extraction was performed on these enhanced images [8], [13].

information pertinent to malignant / benign classification. There are many approaches to mammographic mass segmentation. These include morphological approaches, texture analysis, gray level statistical thresholding and statistical models, level sets, bilateral subtraction of breast image pairs, functional minimization and region growing, multi-resolution analysis and fuzzy region growing, modified median filtering and template matching, adaptive filtering, the radial gradient index and probabilistic methods, vicinal support vector based segmentation, and steerable filters. These methods are discussed in detail in [8].

The segmentation methodology used in this work is based on the approach proposed by Catarious in [13]. Since the focus of this experiment is on the backend (multi-classifier and decision fusion framework) design, the reader is referred to [8], [13] for a detailed description of the segmentation method. This segmentation is chosen as our system's front-end because it performed very well for a benign / malignant classification on the mammogram database used by Catarious [13]. It provided a "successful" segmentation method and has been documented in previous publications.

Table 5.3 enlists the features extracted from the dataset. The final feature vector for every mammogram (generated after segmentation and feature extraction) is a concatenation of all the features listed in table 5.3. The resulting feature space is hence inherently very high dimensional (1032-dimensional in this case.) A high dimensional feature space adversely affects classification performance (unless there is sufficient training data to support the high dimensional space.) Hence, conventional single-classifier based systems typically employ a suitable feature selection and optimization algorithm for dimensionality reduction before classifying the mammograms.

Table 5.3 Features extracted from mammograms in the DDSM database for classification of mammogram images.

Feature type	Number of features	References
Patient age	1	[8, 10, 15]
Morphological features (Area, axis ratio, box ratio, eccentricity etc.)	14	[8, 16]
Statistical features (grey-level mean, standard deviation etc.)	3	[8, 16]
Features extracted from the Normalized Radial Length (entropy, mean, roughness etc.)	6	[8, 16, 17, 18]
Features extracted from the gray-level co-occurrence matrix (Energy, variance, covariance etc.)	1008	[8, 16, 17, 19]

LDA is a popular preprocessing / dimensionality reduction tool commonly employed by many researchers in the pattern classification community. As mentioned in previous chapters, one limitation of LDA based preprocessing is that to learn the LDA based linear transformations, the various scatter matrices (similar to covariance matrices) must be estimated in the original feature space from training data. For high dimensional data, a large amount of training data must be available to estimate these matrices. In the absence of a large amount of training data, estimates of these matrices are likely to be ill-conditioned, and LDA based preprocessing is likely to fail (or be suboptimal).

S-LDA with forward selection and backward rejection is commonly employed to mitigate affects of small-sample-size on LDA transformations. This algorithm is described in section 2.4. This forward selection and backward rejection approach results in a determination of the "best" feature subset, upon which if LDA is applied, the class separation in the resulting space will be high. Although this approach allows us to draw on the benefits of the LDA transformation for high dimensional feature spaces, it is still sub-optimal, in that the selection and rejection procedures outlined above do not perform

an exhaustive search on the feature space to find the optimal 'combinations' of features. Since this algorithm is commonly employed by many researchers as a feature optimization strategy for pattern classification (including medical CAD) systems, it is used as the baseline system in this work - classification performance of the proposed approach is compared against that of this S-LDA approach.

5.3.2 Experimental Dataset

This study uses an image database from the Digital Database for Screening Mammography (DDSM) database [14]. The 60 cases in our dataset consist of 30 randomly selected benign cases and 30 randomly selected malignant cases, where only 17 of the malignant cases are spiculated. Note that spiculated benign cases are relatively rare, and none were included in this study. Each test case consists of a mammographic image, the diagnosis (malignant or benign, which are validated with biopsies and follow–up visits), the patient's age, a physician supplied region of interest (ROI), and a radiologist assessment of whether the mass is spiculated. Only an indication of the presence of spicules is provided (a spiculation template is not provided). The original mammographic images are scanned with a Howtek scanner with a pixel size of 43.52 μm2 and 12 bits per pixel radiometric resolution [14]. Each mammogram is cropped to [2048 × 2048] pixels in an area around the physician supplied ROI. Further discussion of this dataset in terms of subtlety, margins and density are provided in detail in [8].

5.3.3 The Proposed MCDF Approach

It has been established in previous sections that when dealing with high dimensional feature spaces and a small training sample size, a multi-classifier and

decision fusion approach to classification can provide high classification accuracies, even when the conventional approaches such as PCA, LDA and S-LDA begin to fail. When employed on high dimensional, remotely sensed hyperspectral data, this framework provided a reliable classification mechanism, even under relatively poor signal fidelity conditions. This experiment proposes to employ this algorithm in the context of features extracted from mammogram images. The dimensionality of features extracted from these images can be as high as a few thousand.

The proposed framework is as follows: (1) Find a suitable partition of the feature space, i.e., identify appropriate subspaces (each of a much smaller dimension); (2) Perform "local" classification in each subspace; (3) Finally, employ a suitable decision fusion scheme to merge the local decisions into a final malignant/benign decision per mammogram image. In the work with hyperspectral imagery presented in the previous chapters, it was found that the correlation structure of the feature space was approximately block-diagonal. This permitted the use of a correlation or mutual information based metric in the partitioning of the corresponding feature space into multiple contiguous subspaces [4]. However, unlike hyperspectral data, where the feature space comprises of reflectance values over a continuum of wavelengths, features extracted from mammogram images typically do not possess a standard correlation structure. This is primarily because these features are created by concatenating various different kinds of quantities, such as morphological characteristics, texture information, patient history etc. Hence, in an attempt to define a suitable partition of the feature space derived from mammogram images, the feature space is broken down into small groups, each comprised of m adjacent features, where m is a small integer valued number,

determined experimentally. In previous work, Ball and Bruce [8], [12] found that when doing a forward selection and backward rejection of mammography features, patient age was always selected as an important feature in the final feature selection. Hence, in this work, patient age was injected into each partition/subspace generated above to strengthen each local classifier.

Other than the feature grouping (subspace identification) procedure described above, the remaining implementation of the MCDF system remains the same as described in section 3.3 and figure 3.4. LDA was employed as the pre-processing technique for each subspace/group, and quadratic maximum-likelihood classifiers were employed in the multi-classifier system. Decision fusion performance of both MV and LOP are studied in this experiment. Results from this experiment are provided in table 5.4. The partition size, *m* was varied from 2 (a very small number), to 15 (a reasonably large partition size, considering the limited size of the training data). As is common in medical image processing and CAD literature, classification performance, as quantified by the

Table 5.4 Classification performance of the proposed system with the DDSM dataset.

Stepwise LDA (Baseline)					***MV based fusion (Proposed)***				***LOP based Fusion (Proposed)***			
OA	***CI***	***SE***	***SP***	***(m)***	***OA***	***CI***	***SE***	***SP***	***OA***	***CI***	***SE***	***SP***
82	4	80	83	**2**	85	3.8	87	83	85	3.8	87	83
				3	**90**	**3.2**	**90**	**90**	**88**	**3.4**	**90**	**87**
				4	85	3.8	83	87	85	3.8	83	87
				5	80	4.2	77	83	82	4.1	80	83
				6	85	3.8	83	87	83	3.9	83	83
				7	82	4.1	80	83	82	4.1	80	83
				8	82	4.1	80	83	82	4.1	80	83
				15	78	4.4	73	83	78	4.4	73	83

OA: Overall Accuracy; CI: 95% Confidence Interval; SE: Sensitivity; SP: Specificity (all expressed in percentage); *m*: partition size

overall recognition accuracy, the specificity (proportion of true negatives correctly identified), and sensitivity (proportion of true positives correctly identified) are reported in this table. To conclude, the proposed multi-classifier, decision fusion system has the potential to significantly outperform the baseline single-classifier system, for small partition sizes (m). By employing the proposed system, the overall accuracy, sensitivity and specificity of the binary classification task improve by as much as 10%. Hence, the MCDF framework promises robust classification of mammographic masses [20] even though the dimensionality of feature vectors extracted from these mammograms is very high. This demonstrates that the benefits of the MCDF approach are not restricted to hyperspectral data alone.

REFERENCES

[1] Simberloff, D., I. M. Parker, P. N. Windle, "Introduced species policy management and future research needs," *Frontiers in Ecology and the Environment,* vol. 3, no. 1, pp. 12 - 20, 2005.

[2] HYPERION instrument specifications, available: http://eo1.gsfc.nasa.gov/Technology/Hyperion.html

[3] B.Jeon, D.A. Landgrebe, "Decision fusion approach for multitemporal classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 37, no. 3, pp 1227-1233, May 1999.

[4] S. Prasad, L.M. Bruce, "Decision Fusion with Confidence based Weight Assignment for Hyperspectral Target Recognition," *IEEE Trans. Geoscience and Remote Sensing*, vol. 46, no. 5, May 2008.

[5] Analytical Spectral Devices FieldspecPro FR specifications. Available: http://asdi.com/productsspecifications-FSP.asp.

[6] A. Mathur, L.M. Bruce, "Exploiting Hyperspectral Hypertemporal Imagery with Feature Clustering for Invasive Species Detection," *Proc. 2006 IEEE Intl. Geoscience and Remote Sensing Symp.*, pp 828-831, July 2006.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[8] J.E. Ball, "Three stage level set segmentation of mass core, periphery, and spiculations for automated image analysis of digital mammograms," *Ph.D. in Electrical Engineering*. Starkville, MS: Mississippi State Univ., May 2007.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, New York: John Wiley & Sons, 2001.

[10] National Cancer Institute, "National Cancer Institute Fact Sheet: Improving Methods for Breast Cancer Detection and Diagnosis," 2006. Available: http://www.cancer.gov/cancertopics/screening/breast

[11] A. Jemal, T. Murray, E. Ward, A. Samuels, R.C. Tiwari, A. Ghafoor, E.J. Feuer, and M.J. Thun, "Cancer Statistics, 2005," CA: A Cancer Journal for Clinicians, vol. 55, no. 1, pp. 10-30, 2005.

[12] J. E. Ball, L.M. Bruce, "Level Set-Based Core Segmentation of Mammographic Masses Facilitating Three Stage (Core, Periphery, Spiculation) Analysis," Proc. Of EMBS, France, 2007.

[13] D.M. Catarious, "A Computer-Aided Detection System for Mammographic Masses," PhD Dissertation in Biomedical Engineering, Durham, NC: Duke University, Aug. 2004.

[14] M. Heath, K.W. Bowyer, D. Kopans, W. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the Digital Database for Screening Mammography," in Digital Mammography, N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, Eds. Boston, MA: Kluwer Academic Publishers, pp. 457-460, 1998.

[15] D.M. Catarious, A.H. Baydush, and C.E. Floyd Jr., "Incorporation of an iterative linear segmentation routine into a mammographic mass CAD system," in *Medical Physics*, vol. 31, no. 6, pp 1512-1520, June 2004.

[16] H.D. Cheng, X.J. Shi, R. Min, L.M. Hu, X.P. Cai, and H.N. Du, "Approaches for automated detection and classification of masses in mammograms," in *Pattern Recognition*, vol. 39, no. 4, pp 646-668, April 2006.

[17] R. M. Haralick, I. Dinstein, and K. Shanmugam, "Textural features for image classification," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, pp 610-621, November 1973.

[18] A. Agatheeswaran, "Analysis of the effects of JPEG2000 compression on texture features extracted from digital mammograms," Masters thesis in Electrical Engineering. Starkville, MS: Mississippi State University, pp 20-37, 42-43, December 2004.

[19] B. Sahiner, H. P. Chan, N. Petric, M.A. Helvie, and M.M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," in *Medical Physics*, vol. 25, no. 4, pp 516-524, April 1998.

[20] S. Prasad, L.M. Bruce, J.E. Ball, "A Multi-classifier and Decision Fusion Framework for Robust Classification of Mammographic Masses," in *IEEE Engineering in Medicine and Biology Conference*, Vancouver, Canada, August 2008.

CHAPTER VI

CONCLUSIONS

6.1 Conclusions

A new paradigm for robust statistical classification of high dimensional data was proposed in this dissertation. The proposed divide-and-conquer framework partitions the high dimensional classification problem into multiple smaller dimensional classification problems, followed by decision fusion to combine these local classification results. This framework was primarily tested on hyperspectral classification datasets. With hyperspectral data, it was shown that an appropriate bottom-up band-grouping followed by a multi-classifier decision fusion setup resulted in very high classification accuracies. It was also shown that with the handheld ASD hyperspectral data, a mutual information based metric for the bottom-up band-grouping procedure provided a better partition of the hyperspectral feature space, resulting in greater classification accuracies in mild and moderate pixel mixing conditions. An adaptive decision fusion approach based on non-uniform classifier weight assignment was also proposed in this work that accounted for non-uniform fidelity across the feature space.

The basic MCDF approach based on linear transformations (LDA) at the subspace level resulted in very high classification accuracies when pixel mixing was not severe. An alternate nonlinear (KDA) preprocessing at the subspace level was also proposed in this dissertation to provide greater robustness within the MCDF framework, even under severe pixel mixing conditions. It was found that the nonlinear version of the MCDF framework (MCDF-KDA)

provided superior classification performance over a wide range of pixel mixing conditions, enabling the MCDF framework to be useful even when the pixel mixing is severe.

To demonstrate the aptness of the MCDF framework to general high dimensional classification tasks, this dissertation also tested the framework with three different practical classification tasks. In the first such task, it was demonstrated that the MCDF approach performed better than conventional methods at classifying invasive species from satellite hyperspectral data. It was also shown that the nonlinear extension of the framework (MCDF-KDA) provided an even more robust classification performance in mixed pixel conditions. The second practical classification task consisted of classifying aquatic invasive species from available multitemporal, hyperspectral data. Towards this goal, the MCDF framework was extended to a two-tier decision fusion system, combining information over spectral and temporal dimensions. The resulting classification systems' performance was far superior compared to conventional methods. The third practical classification task consisted of employing the MCDF framework as a classification backend to a CAD system for identifying malignant and benign tumors from mammographic images. Once again, an appropriate partitioning of the very high dimensional feature space, followed by local classification and decision fusion resulted in a far superior classification performance as compared to conventional methods.

Experimental results presented in this dissertation demonstrate that the MCDF framework is indeed a promising classification approach for high dimensional datasets. One can conclude from its ability to exhibit robust classification performance for a variety of datasets using very little training data that it should indeed be considered for classification tasks involving high dimensional small-sample-size datasets. Hyperspectral data possesses a statistical structure that facilitates a natural partitioning of the spectrum using a bottom-up band-grouping approach. However, for an arbitrary classification task, where the feature vectors do not possess such a statistical structure, the key issue that would need to be addressed to successfully employ the

MCDF framework is the identification of a suitable partition of the high dimensional feature space (as was done with the mammography dataset in the previous chapter).

6.2 Suggested Future Work

The MCDF framework is not only a powerful classification approach – it also provides a natural framework for extending the classification setup from a single-source to a multi-source setup. If additional sources of information are available for classification, such as, Radio Detection and Ranging (RADAR), or texture information derived from a high spatial resolution sensor in addition to hyperspectral data, these can be combined for better classification. It would be interesting to study the extension of the MCDF approach proposed in this dissertation to a classification task that simultaneously employs data acquired from multiple modalities.

In this dissertation, the MCDF framework is tested for two-class (binary) recognition problems. It can however be easily extended to be used in multi-class recognition problems, such as land-cover classification. It would be interesting to study the performance of this framework as applied to such tasks.