CAR-TR-821
CS-TR-3629

N00014-95-1-0521
May 1996

Multi-scale Discriminant Analysis and Recognition of Signals and Images

Kamran Etemad
Center for Automation Research
University of Maryland
College Park, MD 20742-3275

Abstract

This dissertation explores multiscale discriminant basis selection, as well as the improvement of classification reliability through context-dependent integration of soft decisions. These methods are applied to texture and radar signature classification, document image segmentation, and human face recognition.

The support of this research by the Defense Advanced Research Projects Agency (ARPA Order No. C635) and the Office of Naval Research under contract N00014-95-1-0521 is gratefully acknowledged, as is the help of Kathy Bumpass in preparing this report.

Preface

A successful pattern recognition scheme starts with efficient extraction of the most discriminant information elements from various, possibly imprecise, sources, followed by an intelligent combination of this information in a context-dependent framework of low complexity.

Conventional multiscale basis selection and feature extraction based on compression- and approximation-based criteria are not necessarily the best approaches for classification and segmentation purposes. Instead, a class separability based approach is preferable. In this dissertation, we explore methodologies for lower-dimensional adaptive multi-scale discriminant basis selection. Depending on the task, these methodologies are applied to local windows or to the whole pattern. Our tools in this analysis are derived from theories of wavelet packets and multi-scale local bases on the one hand, and from the statistical theory of discriminant cluster analysis on the other hand. The goal is to find efficient multi-scale representations that yield maximum between-class separations and minimum within-class scatters.

We also investigate the effectiveness of soft decisions in representing the vagueness, uncertainty and imprecision of the classification sources. Based on

the principle of least commitment in designing pattern recognition and consensus-theoretical concepts, we try to improve the reliability of our classification system through integration of soft decisions obtained from various observations and/or sources. The combination of decisions is based on the discrimination power of each source and its relevance to the current observation. We use ideas from consensus theory, fuzzy neural learning, and evidential reasoning.

Our methods of multi-scale local/global basis selection and context-dependent decision integration are applied to in several different domains, including texture and document image classification and segmentation, radar signature classification, and human face recognition. The results show that superior or highly competitive performance can be obtained using small feature sets and simple classifiers. The resulting systems are typically of low complexity and, since no iterative computations are involved, most of the calculations can be done in parallel. The proposed ideas can be extended in several directions and can be applied to many pattern recognition and segmentation tasks.

Acknowledgements

I am pleased to have the opportunity to acknowledge the individuals who made contributions to this thesis. First and highest thanks to God, our ultimate supervisor, without whom any success is both meaningless and worthless.

I would like to thank my advisor, Dr. Rama Chellappa, for his support, encouragement and useful suggestions. It has been a great privilege to work with him and to be a member of his talented and dedicated research group.

Many thanks to Dr. Azriel Rosenfeld whom I have always admired for his remarkable work, attitude, and dedication to the research group at the Center for Automation Research. I highly appreciate his careful comments and suggestions. Also my friend and supervisor Dr. David Doermann who guided me in my research on document processing deserves special thanks and acknowledgment.

I am also indebted to Dr. Shihab Shamma; working with him was an insightful and exciting learning phase of my studies. My special thanks to Dr. Nariman Farvardin, one of my best professors who not only gave me a lot of insights in digital communication and source coding but also provided me with much valuable, friendly advice and caring suggestions.

I am grateful to all my dissertation committee members, including Dr. Ray Liu, for their useful comments and suggestions.

Last but not least, I would like to give my special thanks to my parents and my beloved wife, Shiva, for their love, support, encouragement and trust throughout. I could not have completed this thesis otherwise, and I dedicate it to them.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Pattern recognition is the study of theories and algorithms for automating the process of recognition through efficient representation of relevant information and its analysis using intelligent schemes. The success of pattern recognition systems depends not only on the power of the data processing algorithm, but also on the proper representation of input data so that all the salient aspects of data for the specific task at hand are captured and utilized while all the irrelevant information is discarded. With a poor knowledge representation, even a powerful and sophisticated algorithm may give inferior results. Improvement in efficiency of data representation may achieve more benefit with less effort. Another fact which is sometimes overlooked is the significance of managing intermediate results and decisions, in terms of representing or saving them in the right format so that a minimum amount of information is lost as far as end-to-end performance is concerned. One of the most important principles in designing pattern classification schemes is the principle of least commitment, stated by Marr [59], which simply says "don't do something that may later have to be undone". This principle is consistent with utilizing soft decisions as intermediate results and carrying them along until a crisp decision is required.

A general schematic of a context-dependent classification/recognition process is shown in Figure 1.1. The process starts with making a set of observations that can be ordered in time or space, possibly as results of windowing. In some applications only a few observations may be available. The first step

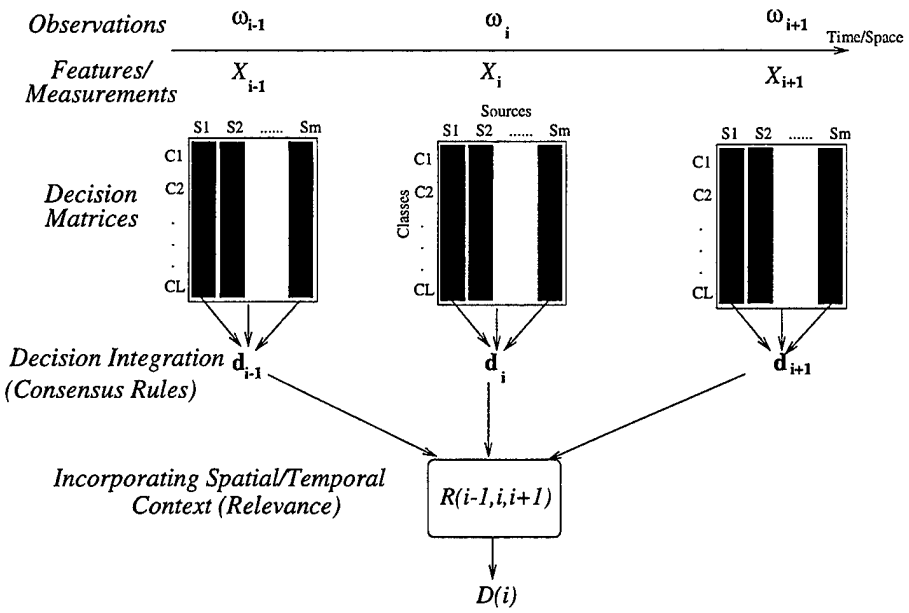


Figure 1.1: Context-dependent classification and recognition using decision integration.

is to find an effective and appropriate representation of the signal or image, which based on a given criterion, represents only the most relevant information in a compact form. The economy of clues in humans' recognition, and the fact that classification systems with small numbers of parameters have better generalization, are computationally more cost-effective and also can be trained and adapted faster, are motivations for efficient feature extraction techniques.

Feature extraction can also be thought of as, or be replaced by, measurements from various sources. These sources in general may be imprecise with certain levels of reliability or significance. Based on a consensus rule and an objective performance measure one needs to combine information or decisions provided by various sources/features to obtain more reliable performance. This requires an objective evaluation of decisions obtained from individual sources in terms of their impreciseness, uncertainty, or reliability.

Also attached to the concept of decision integration is the idea of incor-

porating context in the final decision. The context can be defined and used in a temporal or spatial sense or as any additional side information. Utilizing spatial/temporal context requires another level of decision integration using decisions obtained in a "neighborhood" around current observations based on their objectively defined relevance to the current data or decision.

In this thesis we investigate general methods of pattern classification through adaptive multi-scale local basis design. We also investigate the improvements obtained as a result of using soft local decisions as opposed to hard decisions and we link this idea to fuzzy neural networks, soft decision integration and context-dependent evidential reasoning. Our methodology is based on the following objective and observations.

Objective: The objective of this study is to develop a fairly general pattern and signal classification and segmentation scheme that is highly robust to signal and pattern distortions and can provide competitive results with low complexity. Our main applications of interest are document image processing, texture analysis, radar target classification, and face recognition. We will test our proposed schemes on these tasks and compare our results with other methods.

Observations: There are several primary observations that can lead us toward a reasonable approach to achieving the above objective:

1. **Best Representations for Approximation or Discrimination.** The best and most compact representation of a set of signals for compression or approximation purposes may not be appropriate for classifying them [2]. For discrimination purposes, instead of the description length, entropy, or rate distortion, the criterion should be class separability. In other words, we should seek a small-dimensional representation space with maximum discrimination power. In a feature space with high discrimination power, within each class the feature

points show small variation but feature points from different classes are highly separated. The most discriminating features may or may not correspond to high energy content or to major principal components of the signal(s).

2. **Multi-scale Representation and Classification:** Multi-scale signal/pattern representations and multi-scale classification have been found to be very effective in many signal processing applications ranging from signal compression and coding systems to pattern recognition schemes. Motivated by the success and plausibility of wavelets in classification systems, we will study appropriate choices of local basis functions for the detection, classification or segmentation of signals and images. The goal is to build, from a library of modulated waveforms, the best set of discriminant basis functions relative to which the given collection of signals shows the largest class separability which in turn results in simple and efficient algorithms for classification.

3. **Ambiguity/Impreciseness in Local Decisions:** In many signal/image classification and segmentation applications, because of a variety of constraints we have to base our decisions on local, incomplete or noisy views of the desired pattern. Therefore, there is usually some ambiguity, imprecision or fuzziness associated with our local decisions. In some applications the fuzziness is inherent to the problem and is not necessarily due to noisy or incomplete data. In such cases no hard decision can be accurate, and therefore it is more appropriate to use soft decisions to reflect mixed memberships. Expressing all initial decisions using real-valued soft decision vectors, one has to find a way to reduce their uncertainty and reach an acceptable confidence level using a set of consensus rules.

4. **Incorporating Context Information through Decision Integration:** The effective use of context information in human perception is one of the key

sources of its strength. But using context information in computer vision and pattern recognition efficiently is not a trivial problem. Many pattern recognition schemes attempt to identify relevant context information from different sources and incorporate it in their final decisions. The improvement due to the use of context information becomes more noticeable when primary decisions are imprecise due to a local/incomplete view of the patterns. The integration of soft local decisions over a "context area" within and across scales can reduce the level of uncertainty or increase the confidence of final decisions.

The organization of this dissertation is as follows. In the first two chapters we discuss some new ideas for pattern recognition in a general and analytical form. In the following three chapters we investigate the results of applying these ideas to various signal and image processing tasks.

Chapter 2 talks about best local/global discriminant basis selection and feature extraction. In this chapter we study multi-scale discriminant feature extraction for classification and segmentation purposes. These discriminant bases should be designed so that maximum separability of clusters in the feature space can be achieved using small-dimensional feature vectors. Depending on the task one may look for best features based on local windows on the signal or image, or for global features using all the data points in a signal/pattern. For segmentation tasks, for acceptable localization of region boundaries, one has to use small local windows and the challenge is to obtain consistent results with a limited and sometimes insufficient view of the signal/pattern. On the other hand, for recognition and classification of objects with major macroscopic structure, in order to capture all the geometrical relationships between object components, one may need to view the signal as a whole.

Chapter 3 focuses on the utilization of soft decisions and context-dependent decision integration rules. This chapter discusses methods of finding a consensus among a set of experts or imprecise information sources with different levels of reliability. Integration of spatial/temporal context information based on a relevance function is also discussed. The ideas presented in this chapter are based on evidential reasoning and similarity-based fuzzy decision systems discussed in the literature. After discussing our analytical proposals we test them on a variety of applications.

Chapter 4 presents results of applying multiscale discriminant analysis to some real 1-D and 2-D signal classification problems. To test our method of classifying 1-D signals we use a set of low-resolution radar signatures for automatic target recognition. Then we investigate the effectiveness of applying similar methods to classification and segmentation of 2-D patterns/images, for which we use a set of texture images. The results of these analyses are compared to those using existing wavelet-based classification systems.

Chapter 5 treats layout-independent document page segmentation using adaptive multiscale discriminant features. We present an algorithm for layout-independent document page segmentation based on document texture which makes use of multiscale feature vectors and fuzzy local decision information to overcome the shortcomings of previous segmentation approaches when applied to complex documents. Multiscale feature vectors, computed using a wavelet packet tree which is designed based on document domain specific information, are classified locally using a neural network to allow soft/fuzzy multi-class membership assignments.

Chapter 6 focuses on analysis and recognition of human faces. In this chapter the discriminatory power of various human facial features is studied and a new scheme for Automatic Face Recognition (AFR) is proposed. The first part

of the chapter focuses on the Linear Discriminant Analysis (LDA) of different aspects of human faces in the spatial as well as wavelet domains. This analysis allows us to objectively evaluate the significance of visual information in different parts/features of the face for identifying the human subject.

The LDA of faces also provides us with a small set of features that carry the most relevant information for classification purposes. The features are obtained through eigenvector analysis of scatter matrices with the objective of maximizing between-class and minimizing within-class variations. The result is an efficient projection-based feature extraction and classification scheme for AFR. Although all of the face recognition experiments in this section are performed at a single scale, the underlying LDA-based feature extraction ideas can also be applied to wavelet decompositions.

1.2 Summary of Contributions

This dissertation reports the following new contributions ranging from analytical results to new applications:

- **Discriminant Local Basis Design:** The design of local bases for best discrimination performance using separability criteria; the application of separability measures for best basis selection or composition from an orthogonal or redundant dictionary of local waveforms.

- **Context-Dependent Multisource Soft Decision Integration:** Utilization of a consensus rule that integrates soft decisions based on their discrimination power and exploits spatial/temporal context using a corresponding relevance criterion.

- **Wavelet Packet Based Layout Independent Document Page Segmentation:** The application of texture-based adaptive multiscale features using

wavelet packets along with context-dependent soft decision integration for segmentation of document pages with complex layouts.

- **Discriminant Analysis and Recognition of Human Faces:** The application of linear discriminant analysis to objective analysis of human facial features and automatic face recognition using a simple projection-based method that utilizes separability measures for feature extraction and multisource soft decision integration.

  Beside these contributions, competitive results in automatic radar target recognition and texture segmentation using very small multiscale feature sets are also presented.

Chapter 2

Multi-scale Discriminant Features

2.1 Introduction

Classification of patterns as performed by humans is usually based on a small number of important attributes which often have multi-scale organizations. In practice we are usually confronted with pattern recognition tasks where physically or logically relevant information is not sufficiently well defined and understood. In such applications there is a need to devise algorithmic approaches to finding and evaluating a set of multi-scale classification attributes that show the maximum discriminatory potential in a small-dimensional feature space.

Recently the application of wavelets and multi-rate filter banks [70, 51] to multi-scale feature extraction has received significant attention. Wavelet-based features have been shown to be efficient representations for detection, classification and segmentation of 1-D signals, e.g. speech, music, and other acoustic or radar transients [51, 26, 25]. Successful texture and image analysis schemes based on wavelet or Gabor transforms have also been proposed [44, 13, 82]. Examples of texture and image segmentation using wavelet packets are given in [55, 29]. In addition to engineering tests, the evidence that some multi-resolution and spatial frequency analysis is performed by our visual and auditory systems, demonstrated by psychophysical studies [83, 93], shows the biological plausibility of wavelet-based methods.

Motivated by the success and plausibility of wavelet-based classification systems, in the first part of this chapter we review the basic methodologies for local basis selection found in the literature. We then present our proposed

discrimination-based signal decomposition scheme. Our objective is to build, from a library of modulated waveforms, a set of suitable basis functions relative to which the given collection of signals shows the largest class separability, which in turn results in simple and efficient algorithms for classification. We investigate appropriate algorithms for both orthogonal bases and redundant dictionaries of local functions.

Since classification systems with small numbers of parameters provide better generalization and adaptation performance at lower computational cost [34], we are interested in dimensionality reduction techniques. It is usually advantageous to sacrifice some information in order to keep the number of system parameters to a minimum. With this observation and our suggested basis selection idea, we also study the issue of optimal extraction of low-dimensional feature vectors from multi-scale decompositions of signals. Our approach focuses on the exploitation of class-specific differences obtained through inspection of a pre-defined class separation [34, 24] attainable from the multiscale decomposition, and on finding a linear map that provides the smallest set of features relative to which the given collection of signals shows the largest class separability. This in turn results in simple and efficient classification schemes. Although most of our discussions are about wavelet packet bases, the suggested basis selection method can be applied to other tree-structured local basis functions, e.g. libraries of local sine/cosine functions [19], and also to other tasks such as classification of acoustic transients and biomedical and satellite images. It is shown that simple search techniques can be devised if the basis functions are orthogonal and can be put in a tree structure. The multi-scale dimensionality reduction idea can be used for both orthogonal and non-orthogonal libraries of local basis functions, e.g. local sine/cosine functions, Gabor functions, and even composite

and redundant basis libraries [57].

The idea of designing local bases using class separability criteria has been studied concurrently by Saito and Coifman [72]. Their proposed algorithm is based on local energy features and additive discrimination costs and their tests are based on synthetic data. Part of this dissertation reports the results of similar but independent research which is applicable to non-additive costs, non-orthogonal bases, and arbitrary local features. Also, in the next few chapters the results of tests on several real signal and image classification and segmentation tasks are provided. Our approach to adaptive multi-scale local basis design is general in the sense of its applicability to different signal and image classification tasks.

The organization of this chapter is as follows:

In Section 2.2, a brief introduction to multi-scale signal representations with emphasis on wavelet packets is given. Several known [34, 24] measures of class separability are summarized and a new separability-based tree-structured local basis design is suggested in Section 2.3. Section 2.4 describes a related but independent idea of dimensionality reduction of multi-scale features, followed by its extension to redundant and non-orthogonal basis dictionaries, in Section 2.5. Some comments on multi-scale and context-dependent classification and segmentation are provided in Section 2.6.

2.2 Multi-scale Signal Representations

The optimal representation of signals in the time-frequency plane (or the so-called Phase Plane [19, 56]) is an active area of research, where the optimality is a task-dependent issue. In most time-frequency decompositions, signals are projected onto a set of waveforms or time-frequency atoms [57]. A general family

of time-frequency atoms can be generated by scaling, translating and modulating a single window function $g(t) \in L^2(\mathbf{R})$, where $g(t)$ is a real, continuously differentiable and $O(\frac{1}{t^2+1})$ function satisfying

$$|g| = 1; \quad \text{and} \quad \int g(t) \neq 0; \quad \text{and} \quad g(0) \neq 0; \tag{2.2.1}$$

Therefore any element of the dictionary is of the form

$$g_\gamma(t) = s^{-1/2} g(\frac{t-u}{s}) e^{i\xi t} \tag{2.2.2}$$

and can be identified by the triple $\gamma = (s, \xi, u) \in \mathbf{\Gamma} = (\mathbf{R^+} \times \mathbf{R^2})$, where s, ξ and u represent scaling, modulation, and translation factors, respectively [57]. These waveforms form a dictionary

$$\mathbf{D} = \{\mathbf{g_\gamma(t)} : \gamma \in \mathbf{\Gamma}\} \tag{2.2.3}$$

of basis functions which may or may not be orthogonal or even complete and may or may not have a tree structure. A function/signal is decomposed in a dictionary D by its projections onto the elements of D. The waveforms $\{g_{\gamma_n}\}$ must be selected adaptively based on the local properties of the desired signals, so that the expansion coefficients provide the desired information most "efficiently". The best decomposition strategy also depends on the characteristics of the dictionary.

The smallest possible dictionary is a basis of H, but general dictionaries are redundant families of waveforms/vectors. Examples of orthogonal bases are Wavelet Packet (WP) and Local Trigonometric Basis (LTB) functions; see Figure 2.1. On the other hand, the general family of Gabor functions forms a redundant dictionary of bases. In the following we review the theory of best signal decompositions using tree-structured local bases, where we focus on WP bases. Also, we review the concepts of best decompositions in the framework of

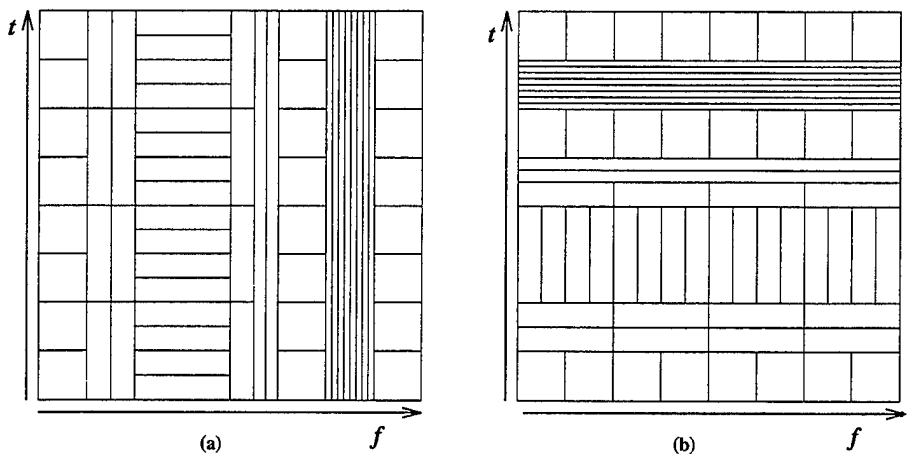


Figure 2.1: Partitioning of the phase plane for (a) wavelet packet basis (adaptive windowing along frequency axis); (b) local trigonometric basis (adaptive windowing along time/space axis)

redundant dictionaries. We then extend the idea of best approximation-based multiscale representation to finding the most discriminatory representations for classification purposes.

2.2.1 Multiscale Orthogonal Bases: Wavelets

Wavelet transforms [56, 22] and their generalized form, called wavelet packets, provide signal analysis through smooth partitioning of the frequency axis. The waveforms in WT and WP dictionaries have a tree structure and they form an orthonormal basis for $L^2(\mathbf{R})$.

We begin with an exact Quadrature Mirror Filter (QMF) [19] $h(n)$ satisfying

$$\sum_n h(n-2k)h(n-2\ell) = \delta_{k,\ell} \qquad \text{and} \qquad \sum_n h(n) = \sqrt{2} \tag{2.2.4}$$

Let $g(k) = (-1)^k h(k+1)$ and define the mappings F_i from $\ell^2(\mathbf{Z})$ onto "$\ell^2(2\mathbf{Z})$"

$$\begin{aligned} F_0\{s\}(i) &= 2\sum_k s(k)h(k-2i) \\ F_1\{s\}(i) &= 2\sum_k s(k)g(k-2i) \end{aligned} \tag{2.2.5}$$

which can be considered as convolutions followed by down-sampling operations. The map $F(s) = F_0(s) \oplus F_1(s) \in \ell^2(2\mathbf{Z}) \oplus \ell^2(2\mathbf{Z})$ is orthogonal, and satisfies alias cancellation and perfect reconstruction conditions

$$F_0 F_0^* = F_1 F_1^* = I \tag{2.2.6}$$

$$F_1 F_0^* = F_0 F_1^* = 0 \tag{2.2.7}$$

$$F_0^* F_0 + F_1^* F_1 = I \tag{2.2.8}$$

where F_0^* and F_1^* are the adjoint (i.e. upsampling and anticonvolution) operations corresponding to F_0 and F_1 respectively. This mapping is the basic block of all wavelet transform and wavelet packet trees [19]. Application of F to each node/subband s projects s onto two orthogonal subspaces $F_0(s)$ and $F_1(s)$ which correspond to the smoothed version of s and the remaining details respectively. Thus each node in the tree represents a subspace of its parent's space and each subspace is the orthogonal direct sum of its two children. The functions g and h represent the low-pass and high-pass filters, respectively. Also $H1$ and $G1$ are the frequency responses of the corresponding 1D filters, used in the filter bank implementation of the system, shown in Figure 2.2. In this Figure V and W are orthogonal subspaces generated at each level of decomposition.

In the wavelet transform the decomposition process is iterated on the low-frequency component and at each iteration the high-frequency coefficients are retained intact. These iterations result in a pyramidal tree structure, which allows signal analysis by dyadically partitioning its spectrum more and more finely toward the low frequency regions. While for many classes of applications and signals this pyramidal multiresolution representation is appropriate, for others it becomes restrictive. For many classes of signals, e.g. textures, document images, and many acoustic signals, where a major part of the energy or "information"

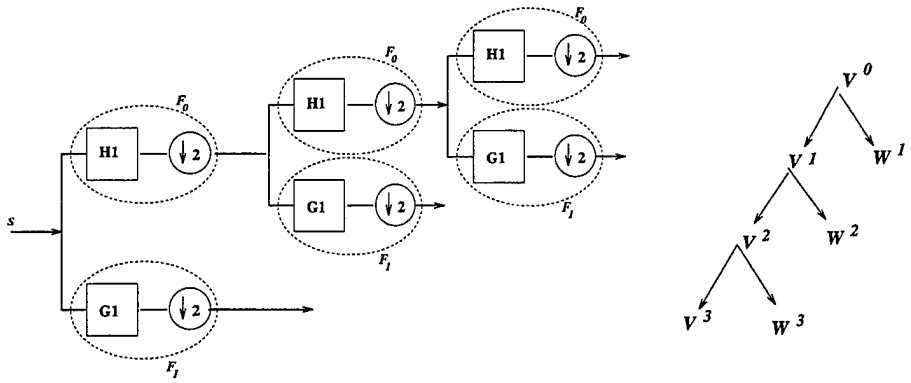


Figure 2.2: Computing the pyramidal wavelet transform by applying the F_0 and F_1 operations and multirate filtering. The filter-bank implementation (left); the tree structure (right).

lies in the mid to upper frequency ranges, the pyramidal wavelet transform is not suitable because, regardless of the spectral characteristics of the signal, it only allows finer and finer resolutions toward the lower frequency bands [29, 55]. On the other hand, fast wavelet packet analysis algorithms permit us to perform adaptive Fourier windowing of a signal by an optimal and smooth partitioning of the frequency axis.

Define the following sequence of functions:

$$\left\{ \begin{array}{l} W_{2n}(x) = \sqrt{2}\sum_k h(k) W_n(2x-k) \\ W_{2n+1}(x) = \sqrt{2}\sum_k g(k) W_n(2x-k) \end{array} \right\} \tag{2.2.9}$$

A Wavelet Packet Basis of $L^2(\mathbf{R})$ is any orthonormal basis selected from the functions $2^{k/2}W_n(2^k t - j)$ [19]. The three parameters $\{k, n, j\}$ have physical interpretations of scale, frequency (or sequency), and position, respectively. Thus each library of Wavelet Packet (WP) bases can be organized as a subset of a full binary tree. In WP analysis both low- and high-frequency components of the signal can be decomposed at each iteration, and thus the corresponding WP tree can grow in different directions.

Wavelet packet expansions correspond algorithmically to adaptive subband

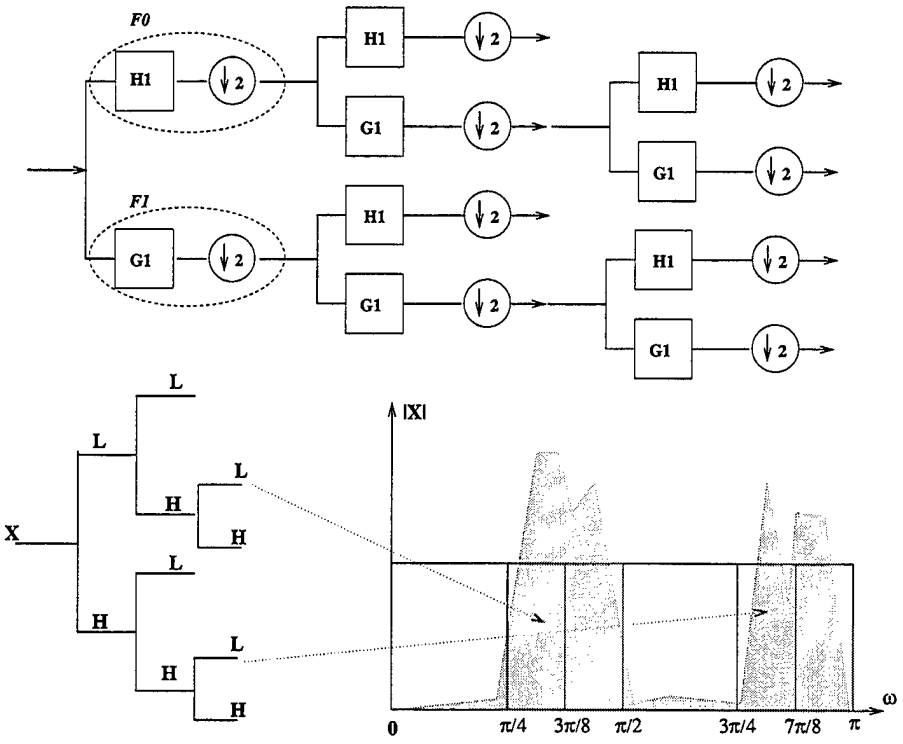


Figure 2.3: Filter bank structure for computing wavelet packets (top). An example of energy-based non-uniform subband decomposition (bottom).

decompositions of signals and images using multi-rate filter banks which are widely used in signal compression systems [89, 2, 68]; see Figure 2.3. In most applications of wavelet analysis to multi-dimensional signals and images, for simplicity, the signal space is assumed to be a separable Hilbert space and in filter bank implementations of 2-D wavelet packets separable filters along the row and column directions are used, i.e.

$$H_{ll}(\omega_x,\omega_y) = H(\omega_x)\cdot H(\omega_y) \qquad H_{lh}(\omega_x,\omega_y) = H(\omega_x)\cdot G(\omega_y) \quad (2.2.10)$$

$$H_{hl}(\omega_x,\omega_y) = G(\omega_x)\cdot H(\omega_y) \qquad H_{hh}(\omega_x,\omega_y) = G(\omega_x)\cdot G(\omega_y)$$

where H and G are the 1-D low-pass and high-pass filters respectively, defined above, and the first and second subscripts show the low-pass or high-pass characteristics of the filters in the row and column directions. Figures 2.4 and 2.5

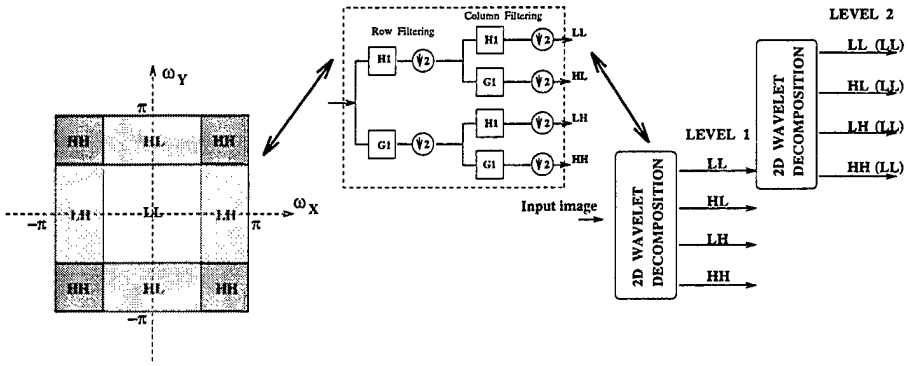


Figure 2.4: 2-D wavelet transform: the partitioning of the 2-D spectrum, the pyramid structure of the tree, and the filter-bank implementation.

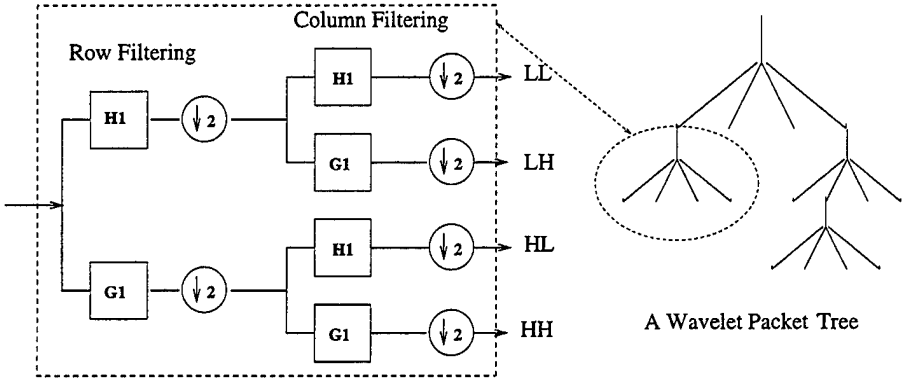


Figure 2.5: Multirate separable filter bank structure for computing two-dimensional wavelet packets (left); a 2-D wavelet packet tree, where each node can be decomposed into four child nodes (right).

show examples of the filter-bank implementations of the 2-D separable WT and WP, respectively.

Another way of optimally representing signals in the time-frequency plane is to perform adaptive smooth partitioning of the temporal/spatial axis, as illustrated in Figure 2.1. This leads to the "dual" or "conjugate" of wavelet packets, the so called "Local Trigonometric Basis"(LTB) functions [19]. It can be shown that it is possible to partition the real line into disjoint intervals smoothly and construct orthonormal bases on each interval. Dyadically partitioning the time axis forms a binary tree of local bases that can be adaptively designed to opti-

mize a predefined cost function. This idea has been studied in [19] in the context of best local basis selection using an entropy criterion for compression purposes.

Wavelet Packet and Local Trigonometric Basis functions are examples of tree-structured orthogonal basis functions. In the following we focus on wavelet packets but most of the results hold equivalently for LTB's or any other tree structured local basis. In particular, all results and algorithms are applicable to M-ary wavelet packet trees [19]. Since keeping all the coefficients in a WP tree leaves us with a redundant set, one asks about the optimal tree structure for a given task. In other words the flexibility of a WP tree enables us to form the WP tree based on a given task-dependent criterion.

In designing wavelet packet trees, one either takes a divide and conquer approach, starting from the most refined sub-space decomposition and moving upward in the tree by merging "adjacent" nodes "appropriately", or starts from the root and performs iterative decomposition of each node into its subspaces if this is "appropriate". In either case the "appropriate" choice is based on a pre-selected task-dependent criterion.

Let us consider the first approach. Using the fact that in wavelet packet trees, at each level, subspaces are orthogonal, and considering the redundancy between a parent node and its children nodes, one can evaluate the pre-selected cost function for the parent node and for the combination of its children, and by comparing the two values decide whether to retain the parent node or the children. Continuing this test for all nodes and levels provides the tree structure appropriate for the specific task based on the pre-selected criterion. The depth of the tree is limited by complexity and other considerations.

Depending on the specific application, criteria can be used to build the optimal wavelet packet tree. Coifman and Wickerhauser [19] have suggested the use

of "entropy" as a measure of energy spread among the transform coefficients. Let H be a Hilbert space. Let $s \in H$, $\|s\| = 1$ and let $H = \oplus \sum H_i$ be an orthogonal decomposition of H. They define

$$\varepsilon^2(s, \{H_i\}) = -\sum \|s_i\|^2 \ln \|s_i\|^2 \tag{2.2.11}$$

the entropy of s relative to the decomposition $\{H_i\}$ of H, as a measure of the distance between s and the orthogonal decomposition. For example, in the LST library case, one compares the entropy of the expansions in two adjacent windows to the entropy of the expansion in their union and picks the smaller one, continuing the comparison with the selection made for the next pair, etc. Thus the tree of basis functions is built so that maximum energy compaction among the fewest coefficients is obtained. Thus, the "best basis" paradigm permits a rapid (e.g. $O(N\log N)$) search among a large collection of tree-structured orthogonal bases to find most compact representation.

For signal compression applications, Vetterli et al. [68] suggest the minimization of the rate-distortion function [20] as a criterion for basis tree selection. This criterion is a compromise between description length and distortion in a compression scheme such as vector quantization. The WP tree is designed to minimize this function. This criterion seems to be appropriate for signal compression and coding applications.

Also for signal analysis and classification problems dominance of energy concentrations in subbands X_s

$$E_s = 1/n \sum_n |X_s[n] - \bar{X}_s|^2 \tag{2.2.12}$$

has been used as a criterion for further decomposition [51, 13], and the "Energy Map" is used as a feature set. The idea behind this approach is the assumption that the most interesting features come from high-energy components of

the signal. Figure 2.3 also illustrated the idea of energy-based wavelet packet decomposition.

2.2.2 Redundant Dictionaries

Although a signal can be completely characterized by its decomposition on an orthogonal basis, any such basis may not be rich enough to represent all potentially interesting microstructures. As in human language, a limited dictionary of words may suffice for expressing any idea, using composite words and sentences; but utilizing a more extensive dictionary enables one to find more compact and efficient ways of expressing ideas. There is an infinite number of ways to decompose a signal/image over a redundant dictionary of waveforms. In fact, it can be shown that in a finite-dimensional space, computing the optimal expansion of signals using a redundant dictionary of waveforms is an NP-complete problem. This justifies the use of suboptimal greedy algorithms. Thus an approximation-based greedy algorithm called matching pursuit is proposed. The problem is to find the optimal, i.e. most compact, decomposition of a signal f over a dictionary of normalized waveforms/vectors $D = \{g_\gamma\}_{\gamma \in \Gamma}$ whose linear combinations are dense in the signal space H. Matching Pursuit is a greedy algorithm that successively approximates a signal f with orthogonal projections on elements of $\mathbf{D}$.

Let $g_{\gamma_0} \in \mathbf{D}$. The vector/signal f can be decomposed into

$$f = < f, g_{\gamma_0} > g_{\gamma_0} + Rf \tag{2.2.13}$$

where Rf is the residual vector after approximating f in the direction of g_{γ_0}. The iterative approximation is performed by successive selection of the dictionary element closest to the decomposition residue at each step and computing the new residual term. Let $R^0 f = f$ and assume that at the k^{th} iteration, $R^k f$ has

already been computed. We choose $g_{\gamma_k} \in \mathbf{D}$ that best matches f

$$| < R^k f, g_{\gamma_k} > | = \sup_{\gamma \in \Gamma} | < R^k f, g_\gamma > | \tag{2.2.14}$$

and then project $R^k f$ onto g_{γ_k}:

$$R^{k+1} f = R^k f - < R^k f, g_{\gamma_k} > g_{\gamma_k} \tag{2.2.15}$$

which defines the residue of the $(k+1)^{\text{st}}$ order. The orthogonality of $R^{k+1} f$ and g_{γ_k} implies

$$||R^{k+1} f||^2 = ||R^k f||^2 - | < R^k f, g_{\gamma_k} > g_{\gamma_k}|^2 \tag{2.2.16}$$

$$f = \sum_{k=0}^{n-1} < R^k f, g_{\gamma_k} > g_{\gamma_k} + R^n f \tag{2.2.17}$$

Thus $R^n f$ is the approximation error of f after n iterations. In fact the original objective was to minimize this error for a fixed n. As part of our discriminant analysis we will revisit this idea and exploit it for best discrimination performance. Details about fast numerical computation of the matching pursuit algorithm and its orthogonal version can be found in [57].

2.3 Discriminant Local Basis

Most of the proposed basis selection algorithms are tailored to provide compact representations and effective signal compression. However, for classification purposes a criterion based on the difference between the patterns/signals of different classes, i.e. class separability, is preferable [28], because one may observe relatively high energy subbands on which the desired signals are quite similar and subbands of relatively low average energies that contain significant information about the differences between the signals. On the other hand the average energy and second central moments of the subbands may not be the only/best feature set for classification. For example, higher-order moments may be used as part of

a feature set and in such cases the decision criterion for further decomposition at each level should also take those features into consideration.

One of the main ideas of this study is to investigate the effectiveness of a separability or discrimination-based criterion for local basis selection. The process of analysis compares projections of a set of signals onto waveforms of a pre-selected library and picks up projections that contain the most discriminatory information. This selection permits discrimination of signals to a specified accuracy with the fewest waveforms. The tree structure selected based on class separability may not be optimal or even sub-optimal for representing or approximating individual signals and it does not even need to provide a "complete" basis, as is required for some other tasks, e.g. compression, identification, and modeling.

In the following we first review the basic ideas of class separability and its measures and then use those measures as our criteria for basis selection from an orthogonal or redundant dictionary of waveforms.

2.3.1 Class Separability Measures

In order to design an efficient classification system one has to select features that are most effective in showing the salient differences between the signals, so that signal clusters are well separated in the feature space.

Consider a collection of N signals $\{s_i\}_{i=1}^{N}$ from L different but known classes. Feature extraction is a mapping from a high and possibly infinite-dimensional signal space to a typically low-dimensional feature space:

$$T : s \in \mathbf{S} \rightarrow \mathbf{V} \in \mathbf{R^n} \tag{2.3.18}$$

The training set Γ is a set of prelabeled observations

$$\Gamma_h \;=\; \{(v_i, l_i) : i = 1, ..., N \text{ and } l_i \in \{1, 2, ..., L\}\} \text{ or} \tag{2.3.19}$$

$$\Gamma_s \;=\; \{(v_i, l_i) : i = 1, ..., N \text{ and } l_i \subset [0,1]^L\}$$

where Γ_H and Γ_S correspond to training based on hard decisions or soft decisions respectively. While a hard decision is a single label assignment, a soft decision is in the form of a real-valued vector where each component of this vector represents the closeness, degree of membership, or similarity between an observation and the prelabeled observations in the training set.

For best feature extraction or evaluation we need a means of quantifying the distance or separability of the clusters corresponding to different classes. Let $P(V|C_l)$ be the conditional density function which represents the spread of feature points $v \in V$ for each class C_l defined in the feature space. For L different classes we are seeking a measure of the distance or separation among the L clusters represented by the $P(V|C_l)$'s:

$$\text{Sep}(V, C) = d(P(V|C_1), P(V|C_2), ..., P(V|C_L)) \tag{2.3.20}$$

Examples of quantitative measures of class separability (CS) are Bayes error, variational distance, scatter matrix based measures, Bhattacharyya distance and divergence rate [34, 24]. Bayes error is the best measure of separability of distributions and for any selection of features it gives the minimum amount of attainable classification error. For a two-class $\{C_i, C_j\}$ problem with equal prior probabilities and uniform misclassification cost and feature vector V, the Bayes error can be simplified to

$$J_{i,j} = \int_V \min(P(V|C_i), P(V|C_j))dV \tag{2.3.21}$$

where the $P(\cdot|\cdot)$'s are conditional class density functions, shown in Figure 2.6. One attempts to minimize this error over different choices of feature vector V. For multiple-class problems Bayes error can be defined similarly by the areas of

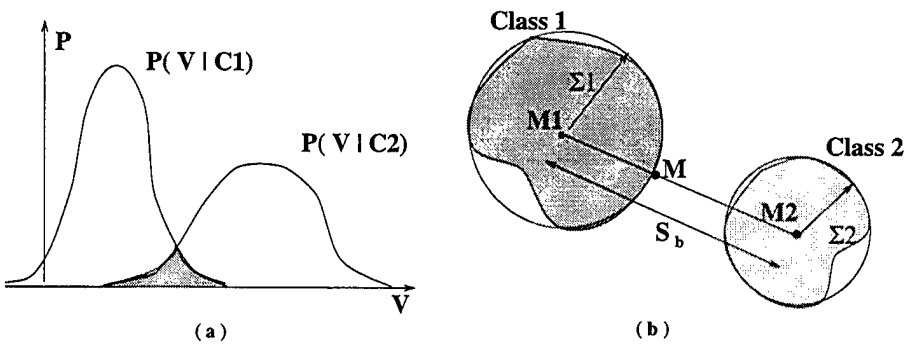


Figure 2.6: Class separability: (a) Bayes error; (b) within- and between-class scatter

the regions where the conditional distributions overlap. Theoretically speaking the Bayes error is the optimum measure of feature effectiveness, and despite its computational complexity, its estimated value is a popular criterion.

The problem with divergence rate and its symmetric variation (J-divergence)

$$\begin{aligned} \text{Divergence } J_{i,j} &= D(P(V|C_i), P(V|C_j)) \qquad (2.3.22) \\ &= \sum_{v \in V} P(v|C_j) \log \frac{P(v|C_j)}{P(v|C_i)} \\ J - \text{Divergence} J^s{}_{i,j} &= J_{i,j} + J_{j,i} \qquad (2.3.23) \end{aligned}$$

as measures of discrepancy between conditional distributions [20] is that they do not have metric properties. Also since divergence is defined for pairs of distributions, when the number of classes is more than two one needs to consider divergences for all pairs and use their summation:

$$d(P(V|C_1), P(V|C_2), ..., P(V|C_L)) = \sum_{i=1}^{L} \sum_{i<j \leq L} J^s{}_{i,j} \qquad (2.3.24)$$

Despite this fact and the computational complexity, divergence-based class separability measures are sometimes used as alternatives to Bayes error.

An elegant and yet simple way of formulating a criterion of class separability is based on within- and between-class scatter matrices, which are widely used in discriminant analysis [34]. The within-class scatter matrix (S_w) shows the

scatter of the sample vectors (V) of different classes around their respective mean/expected vectors M_l:

$$S_w = \sum_{l=1}^{L} Pr\{C = C_l\}\Sigma_l \tag{2.3.25}$$

where $\Sigma_l = E[(V - M_l)(V - M_l)^T | C_l]$ represents the spread of feature points in the i^{th} class. Also one can define the between-class scatter matrix (S_b) as the scatter of the conditional mean vectors M_i around the overall mean vector M:

$$S_b = \sum_{i=1}^{L} Pr\{C = C_i\}(M - M_i)(M - M_i)^T \tag{2.3.26}$$

In order to have good separability for classification one needs to have "large" between-class scatter and "small" within-class scatter simultaneously. There are several ways of defining a positive function as a measure of this combined separability criterion [34]:

$$J^1 = \mathrm{tr}({S_w}^{-1} S_b) \tag{2.3.27}$$

$$J^2 = \ln|{S_w}^{-1} S_b| \tag{2.3.28}$$

$$J^3 = \mathrm{tr}(S_b)/\mathrm{tr}(S_w) \tag{2.3.29}$$

In our experiments J^1 is used but the same results hold for J^2. We denote the objective function computed over subspace V by J_V. A similar but simplified version of this idea has been used in speaker identification and speech recognition problems, where it is called the "F-ratio" [33].

2.3.2 Best Wavelet Packets for Discrimination

In this section we present our WP basis selection scheme which tries to find the best WP tree for classification purposes. First we need to mention that in the following we refer to each node as a subband or feature interchangeably although one may compute more than one feature from each subband. The algorithm is

based on a divide and conquer approach similar to the well-known Best WP basis selection proposed by Coifman et al. [19], but there are some important differences.

Algorithm

1. Select an appropriate wavelet/QMF filter or local sine/cosine transform. Call the operation F.

2. Let $T^{(0)}$ be the root node, let the iteration index $n = 0$, and go through the following iterations. Each iteration involves a decision about decomposing one node from the retained tree.

3. Perform one level of decomposition on each terminal node/subband p:

$$F(p) \rightarrow \underline{C}^{(p)} = \{c_1^{(p)}, c_2^{(p)}, ...c_M^{(p)}\} \tag{2.3.30}$$

4. For each parent node/subband p and its children nodes $\{c_i^{(p)},\ i = 1, ..., M\}$, compute the corresponding feature sets. These feature sets are typically computed through simple nonlinear operations and may or may not be based on local energies.

5. Compare the Combined Class Separability (CCS) obtained using all tree nodes $T^{(n)}$ selected so far, with the parent node $J(T^{(n)}, p)$, to the same CCS excluding node p but including all its children nodes $J(T^{(n)}, \underline{C})$:

$$\begin{aligned} T^{(n+1)} &= \{T^{(n)}, p\} \quad \text{if} \quad J(T^{(n)}, p) > J(T^{(n)}, \underline{C}) \\ T^{(n+1)} &= \{T^{(n)}, \underline{C}\} \quad \text{if} \quad J(T^{(n)}, p) \leq J(T^{(n)}, \underline{C}) \end{aligned} \tag{2.3.31}$$

   In other words, we decompose a node p if this decomposition gives us "additional" significant discrimination information.

6. Repeat steps $3-5$ for the updated tree; increase the iteration index ($n \rightarrow n+1$) until no further significant improvement of separation is observed by decomposing the terminal nodes. One can terminate the iteration earlier if the amount of achieved separation is larger than a preselected threshold.

7. Reduce the dimensionality of the feature vectors using a feature selection method (e.g. Backward Elimination, Forward Selection, or Branch and Bound) to sort the list of features in the order of their CS information importance.

Splitting each subband increases both the within- and between-class scatters, so it may or may not result in an increase of class separation as defined in (2.3.27). However, since windowing is performed in the frequency domain, it is more likely that such an increase will be observed at earlier levels of decomposition rather than later stages where the subbands are too small to reliably characterize the differences. This observation and the depth limitation described earlier explain how the algorithm terminates.

Note that for the special case of additive separability cost, i.e.

$$J(V^{(N)}) = \sum_{i=1}^{N} J(V_i) \tag{2.3.32}$$

(2.3.31) reduces to

$$\begin{aligned} T^{(n+1)} &= \{T^{(n)}, p\} \quad \text{if } J(p) > J(\underline{C}) \\ T^{(n+1)} &= \{T^{(n)}, \underline{C}\} \quad \text{if } J(p) \leq J(\underline{C}) \end{aligned} \tag{2.3.33}$$

which is consistent with [72]. The choice of additive cost may not be appropriate especially when there is a dependency or statistical correlation between features. For example, the combination of two features which carry significant but similar

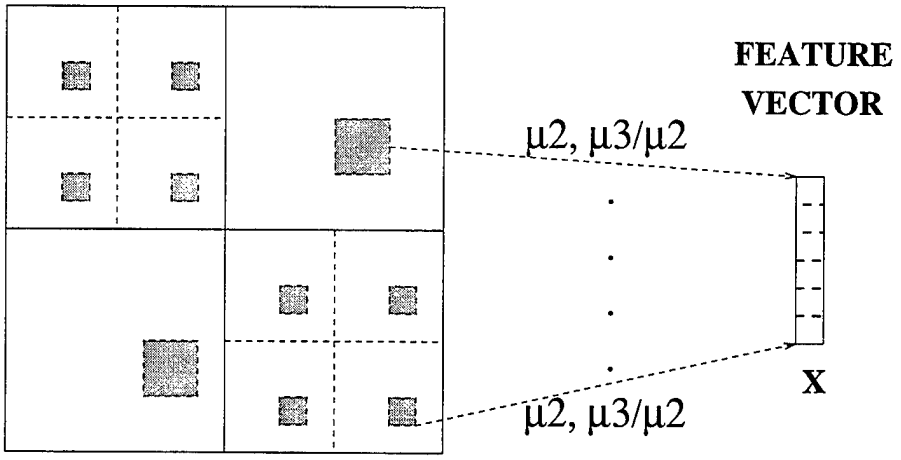


Figure 2.7: Computation of feature vectors for corresponding local windows in all subbands.

discrimination information does not provide us with twice the discrimination power.

Aside from the main idea of the algorithm, one can argue about the appropriate choice of the feature set for each node. Without claiming optimality, as a reasonable choice, we use features based on central moments of the corresponding subband signals, e.g.

$$\begin{aligned} \mu_n(W) &= \frac{1}{|W|}\Big(\sum_{x\in W}(f(x)-\bar{f_W})^n\Big)^{1/n} \qquad (2.3.34)\\ V &= \{v_i=\mu_2(W_i), v'_i=\mu_2(W_i)/\mu_3(W_i) \qquad i=0,1,..,N_{\text{subbands}}\} \end{aligned}$$

where W_i is the local window on the i^{th} subband. On each subband, $f(x)$ and $\bar{f_W}$ are defined as the intensity value at location x and average intensity on window W centered at x respectively, as shown in Figure 2.7. Depending on the nature of the signal or image classification task, W can be a 1-D or 2-D window. For segmentation tasks the window slides through the signal and at each location it covers a part of the signal, whereas in classification tasks there is

only one window covering the whole signal. For each subband or node, μ_2 shows the average energy whereas μ_3/μ_2 roughly represents the information about the shape of the spectrum in that subband.

2.3.3 Separability and Dimensionality Reduction

In order to design a simple and efficient classification and segmentation scheme one has to select features that are most effective in showing the salient differences between the signals, i.e. a selection that results in the best minimal set of features in terms of the separability of the signal clusters in the feature space. The reduction in dimensionality of the feature vectors can be achieved either by selecting them or combining them so that maximum classification information is retained. We start with the selection process and then we study Linear Discriminant Analysis as a tool for obtaining the best linear combination weights.

After the full tree of wavelet basis functions is selected, to simplify the feature vector, those nodes that do not actively contribute to the overall classification performance can be discarded. With this elimination process the pruned tree will no longer correspond to a "complete" basis, but completeness is not required for analysis and classification purposes.

The simplest but most unreliable method of selecting feature subsets (of size $m < n$) is to consider them individually and select from the top of the list of features, sorted based on the cost for each feature alone.

$$U_m = \{u_i, i = 1, ..., m : J(u_i) \geq J(v_j) \quad \forall v_j \in (V - U_{m-1}) \tag{2.3.35}$$

This selection is optimal only if the features are independent and the cost function is additive [24]. In many applications neither is the case. On the other hand, direct exhaustive search, even for moderate sizes of feature sets, is computationally prohibitive. So depending on the tolerated complexity, one can use

suboptimal Forward Selection or Backward Elimination methods, or so-called Branch and Bound search [34, 24].

Iterative comparisons can be initiated from the complete set of features (Ω) by eliminating the one that has the least effect on the overall cost and continuing the same elimination process for the remaining set until the minimum acceptable cost or maximum affordable number of features is obtained [34]. We call this stepwise process Backward Elimination:

$$V^0 = \Omega = \{v_1, v_2, ..., v_n\} \tag{2.3.36}$$

$$V^{k+1} = \{V^k - \arg_{v_i} \min\{J(V^k) - J(V^k - \{v_i\}), v_i \in V^k\} \tag{2.3.37}$$

Also, one can start with the selection of a single feature $V^1 = \{u_1\}$ that results in the largest cost $J(V^1)$. Then, fixing V^1, select from the remaining features a $V^2 = \{\{V^1, \{u2\}\}$ such that it provides the largest cost $J(V^2)$ and continue to include the most effective combination [34]. This is called Forward Selection:

$$V^0 = \text{Null} \tag{2.3.38}$$

$$V^{k+1} = \{V^k, \text{ argmax}_i\{J(V^k, \{v_i\}), v_i \in (\Omega - V^k)\}\} \tag{2.3.39}$$

One can also use variations of the so-called Branch and Bound method of selecting the best subset of nodes/subbands [34, 24]. This approach, although computationally more involved, can provide the optimal selection of nodes even when there is considerable dependence among features across nodes. This algorithm is a top-down search with backtracking which examines all possible combinations without exhaustive search. It is based on the monotonic property of the majority of feature selection criteria, namely for a nested set:

$$V^{(1)} \supset V^{(2)} \supset V^{(3)} \supset ... \tag{2.3.40}$$

$$J(V^{(1)}) \geq J(V^{(2)}) \geq J(V^{(3)} \geq ... \tag{2.3.41}$$

Just to illustrate the basic idea of pruning, the second approach is adopted in the following. Using J as our class separability criterion the algorithm for basis selection can be summarized as follows:

Unlike Mean Square Error(MSE), which is the most widely used criterion for signal representation, class separability measures are typically invariant under any non-singular, linear or non-linear, transformation. However, any singular mapping used for dimensionality reduction results in losing some discriminating information. Our objective is to find the mapping that for a given reduction in space dimensionality provides the maximum class separability. In other words, we are searching among all possible singular transformations for the best subspace which preserves class separability as much as possible in the lowest possible dimensional space, as illustrated in Figure 2.8. So we are seeking a linear transformation $\mathbf{A}$ from $\mathbf{R^n}$ to $\mathbf{R^m}$ with $m < n$ such that

$$A : X \subset \mathbf{R^n} \rightarrow Y \subset \mathbf{R^m} \tag{2.3.42}$$

$$A = \text{argmin}_{A_o}\{|J_X - J_{A_o{}^T X}|\} \tag{2.3.43}$$

where $J_X = \text{tr}(S^X)$ and $J_Y = \text{tr}(S^Y)$ are separabilities computed over the X and $Y = A^T X$ spaces respectively. Thus A optimizes J_Y, i.e. minimizes the drop in cost $|J_X - J_{A^T X}|$ incurred by the reduction in the feature space dimensionality. It can be shown that for such an optimum A

$$\{\lambda^Y{}_i\} \subset \{\lambda^X{}_j\} \; i = 1, ..., m \; , \; j = 1, ..., n \tag{2.3.44}$$

where the λ^X's and λ^Y's are the eigenvalues of the corresponding separation matrices S^X and S^Y. This observation and the fact that

$$J_Y = \text{tr}(S^Y) = \sum_{i=1}^{m} \lambda^Y{}_i \tag{2.3.45}$$

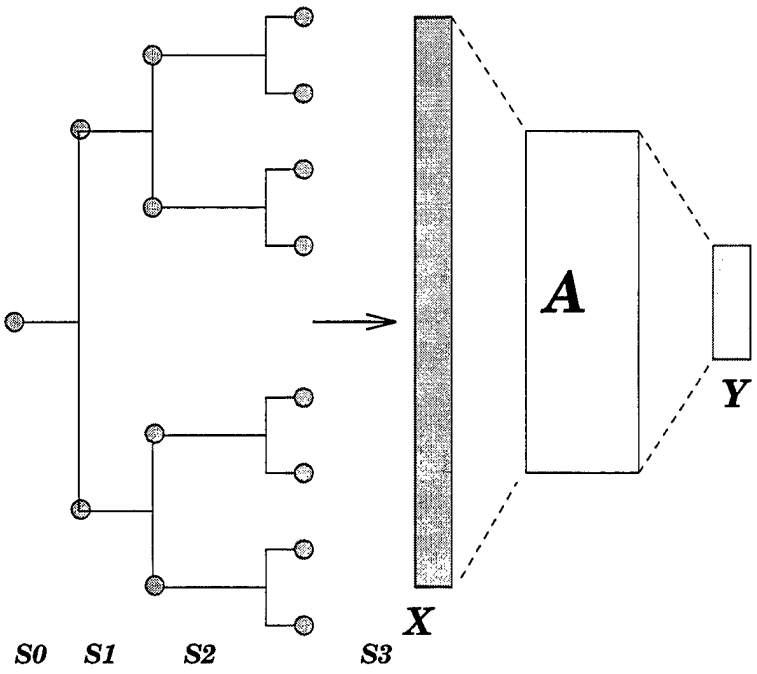


Figure 2.8: Dimensionality reduction of the feature vectors obtained from a balanced or pruned wavelet packet tree.

suggest that one can maximize (or minimize) J_Y by taking the largest (or smallest) m eigenvalues of S^X. Following our earlier observations, and having determined the separation matrix, we perform eigenvalue analysis of the separation matrix S^X on the augmented database:

$$\text{eig}\{S^X\} = \{(\lambda_i, u_i), i = 1, ..., N_S - 1,\ \lambda_i > \lambda_{i+1}\} \tag{2.3.46}$$

To reduce the computational cost for large dataset sizes one can use the following equality [78, 34]:

$$S_b u_i = \lambda_i S_w u_i \tag{2.3.47}$$

This shows that the u_i's and λ_i's are generalized eigenvectors of $\{S_b, S_w\}$. From this equation the λ_i's can be computed as the roots of the characteristic polynomial

$$|S_b - \lambda_i S_w| = 0 \tag{2.3.48}$$

and then the u_i's can be obtained by solving

$$(S_b - \lambda_i S_w)u_i = 0 \tag{2.3.49}$$

only for the selected largest eigenvectors [78]. Note that the dimensionality m of the resulting set of feature vectors is $m < \text{rank}(S) = \min(n, N_S - 1)$. Now define

$$\Lambda^{(m)} = \{\lambda_i, i = 1, ..., m < N_S - 1\} \tag{2.3.50}$$

$$U^{(m)} = \{u_i, i = 1, ..., m < N_S - 1\} \tag{2.3.51}$$

so that $\Lambda^{(m)}$ and $U^{(m)}$ represent the set of m largest eigenvalues of S^X and their corresponding eigenvectors. Considering $U^{(m)}$ as one of the possible linear transformations Ω from $\mathbf{R^n}$ to $\mathbf{R^m}$, with $m < n$, one can show that

$$\Omega = \{U : X \subset \mathbf{R^n} \rightarrow U^T X = Y \subset \mathbf{R^m}, m < n\} \tag{2.3.52}$$

$$U^{(m)} = \text{argmin}_{U \in \Omega}\{|J_X - J_{U^T X}|\} \tag{2.3.53}$$

where $J_X = \text{tr}(S^{(X)})$ and $J_Y = \text{tr}(S^{(Y)})$ are separabilities computed over the X and $Y = U^T X$ spaces respectively. This means that $U^{(m)}$ minimizes the drop $|\text{Sep}(X) - \text{Sep}(U^T X)|$ in classification information incurred by the reduction in the feature space dimensionality, and no other $\mathbf{R^n}$ to $\mathbf{R^m}$ linear mapping can provide more separation than $U^{(n)}$ does; thus $A = U^{(n)}$.

Therefore, the optimal linear transformation from the initial representation space in $\mathbf{R^n}$ to a low-dimensional feature space in $\mathbf{R^m}$ based on our selected separation measure results from projecting the input vectors x onto m eigenvectors corresponding to the m largest eigenvalues of the separation matrix S^X. These optimal vectors/direction can be obtained from a sufficiently rich training set and can be updated if needed. Note that the idea of multi-scale dimensionality reduction can be applied to multi-scale classification systems regardless of the

criterion used for basis selection, e.g. it can be used on the pyramidal wavelet transform, balanced or unbalanced wavelet packet tree, or local trigonometric functions.

2.3.4 Separability and Redundant Dictionaries

Although all of our discussion has been limited to complete and orthogonal dictionaries of bases, the idea of separability-based multi-scale basis design can also be applied to non-orthogonal and redundant dictionaries. In particular, if the initial multi-scale signal representation is obtained through linear operations or "projections" [57], one can absorb the matrix A into these operations. For example, if projections of the signals onto a set of multiscale "templates" $\{\phi_i, i = 1, ..., n\}$ are used, then application of A to these templates, $\{A^T\phi_i, i = 1, ..., m < n\}$, provides a small number of "composite waveforms" on which the projections of the input signals show the largest differences, i.e.

$$V = \{v_i\} = \{< s, \phi_i >\} \tag{2.3.54}$$

$$U = \{u_i\} = \mathrm{AV} = \mathrm{Av}\{_i\} = \{< s, A\phi_i >\} \tag{2.3.55}$$

The original library of multi-scale basis functions can be a redundant dictionary composed of wavelet packet bases, local sine/cosine functions, or families of Gabor functions. Also "composite" signals generated using this method are task-dependent and do not in general have any specific structure like the wavelet tree structure. They can be stored as a set of multi-scale signal templates/vectors to be used in signal projection and feature extraction processes.

For example if a set of Gabor functions Φ with index set Γ is used as the starting dictionary of basis functions

$$\phi_{(\sigma,f,d)} = \exp\left(-\frac{(t-d)^2}{2\sigma}\right) \times \cos\left(2\pi f(t-d)\right) \tag{2.3.56}$$

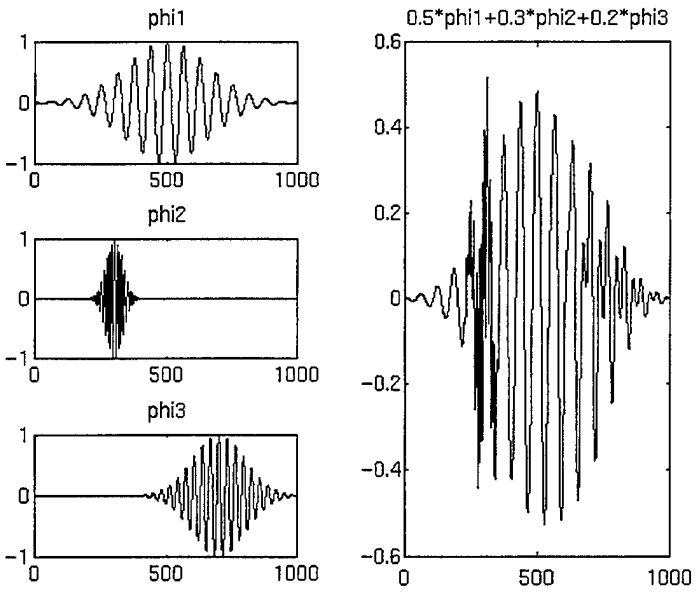


Figure 2.9: Obtaining multiscale composite templates from 1-D Gabor functions using linear combinations corresponding to rows of the dimensionality reduction matrix A.

$$\Phi \;=\; \{\phi_\gamma\}_\Gamma \quad \text{where} \quad \Gamma = \{\gamma\} = \{(\sigma, f, d)\} \tag{2.3.57}$$

and features are computed based on inner products or projections, then according to (2.3.54) a small set of multiscale templates for classification can be obtained based on linear combinations of Gabor wavelets according to the rows of the matrix A. As Figure 2.9 shows, the resulting composite templates may not be symmetric and may not resemble any known local basis. An alternative way of applying the separability idea to redundant dictionaries is a greedy algorithm similar to matching pursuit proposed by [57] or a sequential multi-scale hypothesis testing technique [26].

We can call this method the discrimination or Separation Pursuit (SP) method, which through a greedy sequential search algorithm similar to matching pursuit, tries to suboptimally find the best decomposition for classification purposes. The main difference between SP and MP is that SP uses a different

criterion that needs to be evaluated on a set of prelabeled training functions F rather than individual signals $f \in F$.

Let $\mathbf{F}$ be a matrix whose columns are training signals/vectors of all L classes. The problem is to find, from among all possible decompositions of $\mathbf{F}$ over a dictionary of normalized waveforms/vectors $\{g_\gamma\}_{\gamma\in\Gamma}$, a decomposition that results in projection coefficients with maximum discriminatory power. Like MP, define $R^0 f = f$ as the initial residue of decomposition. Let $g_{\gamma_k} \in \mathbf{D}$. Like matching pursuit at the k^{th} iteration,

$$\forall f \in F: \qquad R^k f =< R^k f, \gamma_k > g_{\gamma_k} + R^{K+1} f \tag{2.3.58}$$

The above equation can be rearranged and written in vector form as

$$R^{k+1}\mathbf{F} = R^k\mathbf{F} - g_{\gamma_k}^T . R^k\mathbf{F} . g_{\gamma_k} \tag{2.3.59}$$

where $R^k\mathbf{F}$ is the matrix of all residue vectors and

$$R^k\mathbf{F}_{\gamma_k} = g_{\gamma_k}^T . R^k\mathbf{F} \tag{2.3.60}$$

is the vector of projection coefficients. The iterative information extraction is performed by successive selection of the most discriminating dictionary element for the decomposition residue at each step and computing the new residual term according to (2.3.59). The most discriminatory element of the dictionary can be selected using any of the separability measures described in Section 2.2:

$$\gamma_k = \text{argmax}_{\gamma\in\Gamma} J(R^k\mathbf{F}_\gamma) \tag{2.3.61}$$

Fast numerical computation of the SP algorithm and its orthogonal version parallels those of MP and can be implemented according to [57].

Chapter 3

Multisource Soft Decision Integration

3.1 Introduction

After extracting multiscale discriminant features we need to find an effective framework for decision making. Effective is meant here in the sense that in the process of classification or recognition the system takes advantage of all the relevant information which is explicitly or implicitly embedded in the feature space. In fact it has been argued and shown that an important factor which typically degrades the classification and recognition performance of most systems lies is the loss of information as a result of under-utilization of information in the feature space [86]. This fact and the principle of least commitment suggest utilizing soft decisions as a more informative representation of intermediate decisions, and carrying soft decisions along until a crisp decision is required.

Consider a general pattern classification/segmentation problem with L different classes, based on m, possibly imprecise, sources with relative levels of expertise denoted by α's. Let $\{\omega_i\}$ be a set of arranged/ordered observations in time or space. These observations may be obtained from sliding windows that span the signal or image. For example, they may correspond to the successive 1-D windows used for speech recognition, or to the 2-D windows of an image segmentation system.

Consider a collection of N examples $\{\omega_i\}_{i=1}^{N}$ from L different, but known, classes. Feature extraction is a mapping from signal space to feature space:

$$T : \omega \in \mathbf{\Omega} \rightarrow \mathbf{X} \in \mathbf{R^n} \tag{3.1.1}$$

so the training set Γ is a set of pairs

$$\begin{aligned} \Gamma_s &= \{(x_i, l_i) : i = 1, ..., N \text{ and } l_i \subset [0,1]^L\} \text{ or} \\ \Gamma_h &= \{(x_i, l_i) : i = 1, ..., N \text{ and } l_i \in \{1, 2, ..., L\}\} \end{aligned} \tag{3.1.2}$$

where Γ_s and Γ_h correspond to training based on soft decisions or hard decisions respectively. Equivalently, we let $x_i^s = x^s(\omega_i)$ denote measurements, e.g. discriminant feature values or vectors, obtained from a source s. The soft classifier is a map $F(.)$, typically non-linear, from the feature space X to the points in the "fuzzy" cube $[0,1]^L$. Thus

$$d : \Re^n \rightarrow [0,1]^L \tag{3.1.3}$$

$$d^s(x_i) = d_i^s \tag{3.1.4}$$

$d_i^s = d^s(x_i)$ is a decision based on measurement x_i from source s. In general this decision is a vector of size L, whose j^{th} element shows the decision (or in fuzzy terms, the fit value) associated with class j:

$$d^s(x_i) = [d^s(x_i, c_j)]_{j=1}^L \tag{3.1.5}$$

Thus, the classifier has L non-binary outputs, one for each class, where each output takes values in $[0,1]$, Figure 3.1. Some authors put a constraint on the summation of the soft decisions made for all classes:

$$\sum_{j=1}^{L} d^s(x_i, c_j) = 1 \tag{3.1.6}$$

These conditions restrict the decision points to a hyperplane in the L-dimensional decision space.

Now let $d_i = g(\{d_i^s, \alpha_s : s \in S\})$ be the decision based on the consensus of all sources, each of which may be imprecise with reliability α_s. The decision

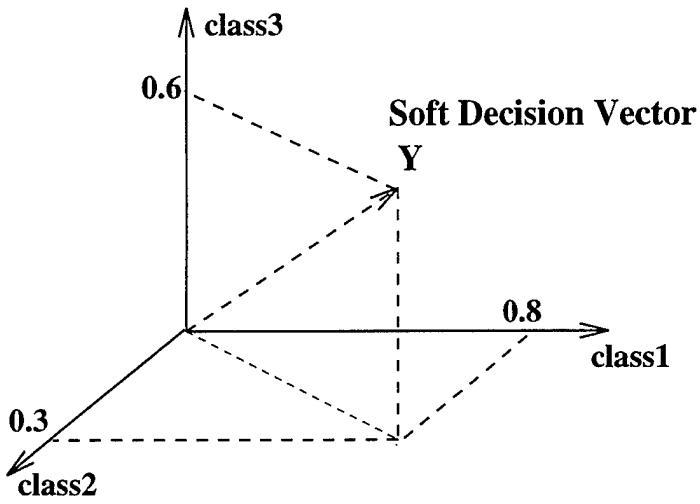


Figure 3.1: Soft decisions for L classes are vectors in an L-dimensional space.

integration is finding the best function g that, through effective combination of the decisions obtained from the individual sources, based on a consensus rule, achieves a more reliable result.

Based on the temporal or spatial arrangement and interrelationships of the observations $\{\omega_i\}$, one can define a notion of neighborhood or context area around each window of observation. Let $D_i = D(\omega_i)$ be the final decision about event ω_i which is a function of the decisions obtained from other windows in the context area of ω_i, and possibly other information Z, i.e. $D_i = h(\{d_j : j \in N_i\}, Z)$. Context-dependent classification and recognition involves the definition of function h based on a reasonable assumption about the interrelation of observations within an area/interval.

In this chapter we first discuss the issue of similarity-based soft/fuzzy classification based on a single source. Then we talk about consensus of experts through decision integration using objectively defined measures of the reliability or importance of information sources. We incorporate the spatial/temporal context information through defining a relevance function that describes the

interrelationships among observations within a neighborhood.

3.2 Fuzzy Partitioning of Feature Space

Let $X = \{x\}$ be a universe of discourse with generic elements denoted by x. Membership in a classical set A of X is often viewed as a characteristic function

$$\chi_A : \qquad X \rightarrow \{0,1\} : \text{ such that } \chi_A(x) = 1 \Leftrightarrow x \in A \tag{3.2.7}$$

which assumes that the set has precisely defined boundaries and each element (i.e., each observed example) has either full or no membership in set A. This assumption results in hard partitioning of the feature space, and as we will discuss later, there is a loss associated with such partitioning.

A fuzzy set B is, on the other hand, characterized by a function f_B which associates with each x a real number in $[0,1]$ that represents the "grade of membership" of x in B. The closer the value of f_B is to 1, the more x belongs to set B. So, while in hard decision each observation is labeled as one of the possible classes, soft classification attaches to each observed pattern a group of membership grades. The fuzzy set membership functions simply but efficiently encode a complete ordering among the set elements. Such orderings carry a lot of information about the relative location of an observation/measurement in the feature space with respect to clusters of prelabeled data. They are also useful for discriminating between values in relation with a variety of semantics (e.g. preference, uncertainty, or similarity) that a fuzzy set based representation may bear in different tasks.

In this chapter our study of fuzzy memberships and soft decisions is mostly related to grades of similarity and dissimilarity suggested by a group of experts or classification resources. In this context the elements with membership 1 are viewed as prototype elements of the fuzzy set, while other membership grades

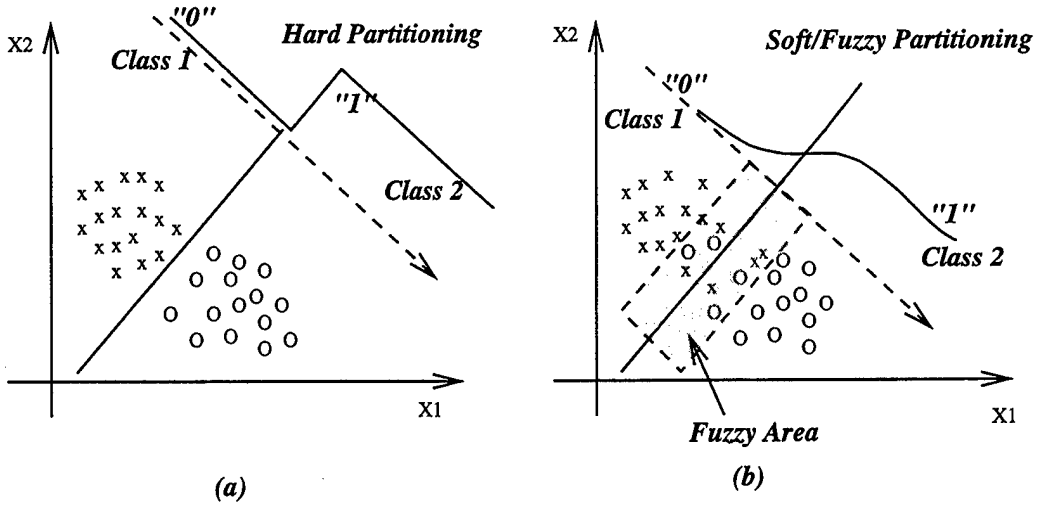


Figure 3.2: Hard partitioning (a) and soft partitioning (b) of the feature space.

estimate the closeness of the elements to the prototypes. An observation may belong to a class to some extent and meanwhile belong to another class to another extent, and membership grades are attached to quantitatively indicate these extents. Such a partition is referred to as a fuzzy or soft partition of the feature space. Formally, a fuzzy partition of a feature space is a family of fuzzy sets $\{C_i, i = 1, ..., L\}$ on universe X such that

$$\forall x \in X \quad 0 \leq f_{c_i} \leq 1 \tag{3.2.8}$$

$$\sum_{x \in X} f_{c_i}(x) > 0 \tag{3.2.9}$$

$$\sum_{i=1}^{m} f_{c_i}(x) = 1 \tag{3.2.10}$$

In a multidimensional feature space the concept of fuzzy membership is equivalent to soft/fuzzy partitioning of the space, where decision regions are not separated by sharp hyperplanes, but there are transition or fuzzy areas between any two decision regions. Figure 3.2 schematically shows how a soft/fuzzy partitioning of feature space may represent the memberships and similarities more

realistically than a hard decision. This allows for classification information to be utilized in subsequent analysis.

3.2.1 Learning Membership Functions

After considering the effectiveness of soft decisions, one has to devise systematic approaches to training the classifier to form the required soft decision boundaries. Here we mention two major approaches to creating such membership functions. The first method relies on probability measures of fuzzy events and in particular on the so called fuzzy mean and fuzzy variance of a fuzzy set [86]:

$$\mu_c^* = \frac{\sum_{i=1}^n f_c(x_i)x_i}{\sum_{i=1}^n f_c(x_i)} \tag{3.2.11}$$

$$\Sigma_c^* = \frac{\sum_{i=1}^n f_c(x_i)(x_i - \mu_c^*)(x_i - \mu_c^*)^T}{\sum_{i=1}^n f_c(x_i)} \tag{3.2.12}$$

Note that these definitions are different from their classical counterparts in that each example contributes to the mean and variance of a class based on its partial membership in that class.

After estimating the mean and variance based on a prelabeled training set, and assuming that the cluster of points for each fuzzy set follows a normal distribution, one defines a Gaussian-shaped membership function as [86]

$$f_c(x) = \frac{P_c^*(x)}{\sum_{l=1}^L P_l^*(x)} \quad \text{where} \tag{3.2.13}$$

$$P_c^*(x) = \frac{1}{(2\pi)^{N/2}|\Sigma_l^{*-1}|^{1/2}} \exp[-1/2(x - \mu_c^*)^T \Sigma_i^{*-1}(x - \mu_c^*)] \tag{3.2.14}$$

Based on the same idea and using μ_c^* and Σ_c^* one can define other types of membership functions, e.g. triangular, exponential, or trapezoidal.

It has been argued that including some of the mixed classes in the training set with their corresponding best mixed labels helps in terms of better estimating the mean and variance and therefore in the final performance. Including such

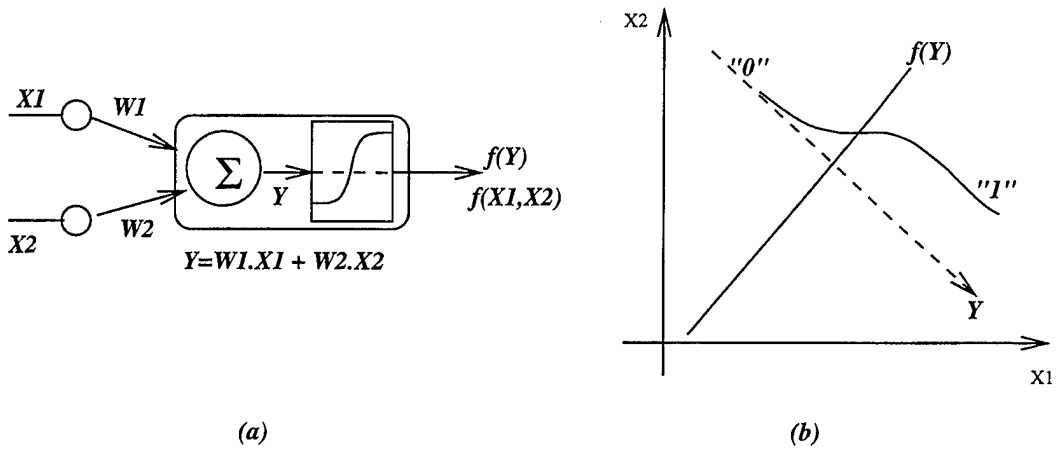


Figure 3.3: Each node in the hidden layer of a MLP network forms a "soft hyperplane" in the feature space: (a) a basic connection in MLP, (b) the corresponding soft "hyperplane"

fuzzy cases in the process of training is sometimes referred to as a fuzzy training method. Obviously, reasonable fuzzy training requires a methodology of defining membership values for the training set.

An alternative approach to adaptively defining membership functions is to use the nonlinear mapping characteristics of Multilayer Neural Networks (MLNN) and supervised learning algorithms to learn multidimensional membership functions based on a training set.

Consider a simple neural network with input layer X, connection weight matrix $W1$, and step function non-linearity for hidden and output nodes Z. As shown in Figure 3.3, each hidden node i in the first layer represents a hyperplane $W_i^T X$ in the space spanned by the input feature vectors. With sigmoidal non-linearities at each node the hyperplane becomes a fuzzy hyperplane:

$$Z = \text{Sigm}(W^T.X) \qquad \text{where} \qquad (3.2.15)$$

$$\text{Sigm}(y) = \frac{1}{1+\exp(-y)} \qquad (3.2.16)$$

In a three-layer network these hyperplanes can be combined to form any

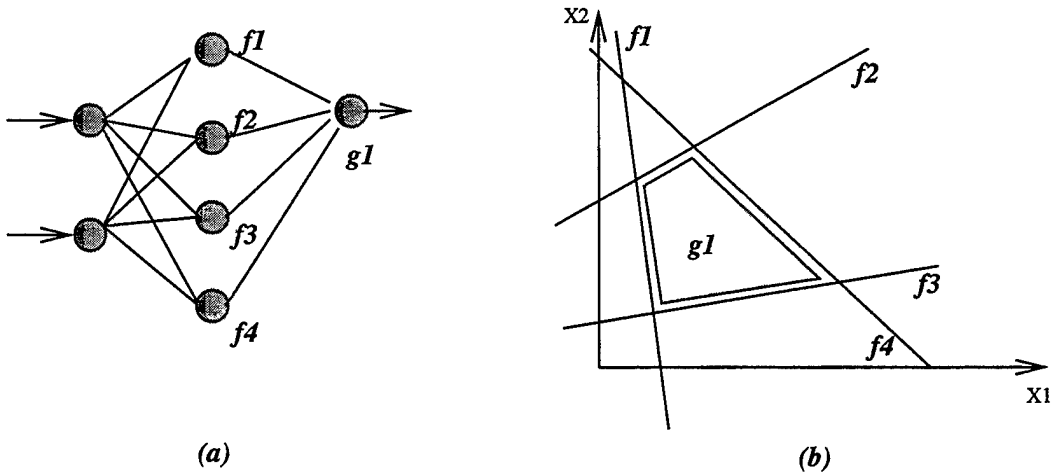


Figure 3.4: Creating soft decision boundaries with neural networks: (a) a two-layer neural network, (b) the corresponding decision region.

set of convex soft decision boundaries (see Figure 3.4). Also with three layers, even non-convex fuzzy decision regions can be formed. This process of creating decision regions is very similar to linear programming ideas, except that they involve a mild form of non-linearity.

This neural network based approach provides a flexible framework for implementing fuzzy training ideas. This type of training also requires the input-output pairs for all training, including mixed/fuzzy examples for which one needs to define a criterion for membership assignments. Note that this membership assignment has to be consistently and mathematically defined and applied to the training sets. Including such examples in the training set provides the network training algorithm with valuable information about the slope of the membership function in the transition regions.

3.3 Multisource Soft Decision Integration

A number of different approaches have been proposed for analyzing information obtained from several sources [52, 8, 34]. The simplest method is to form an extended data/feature vector, containing information from all the sources, and

treat this vector as the vector output of a single source. Usually, in such systems all similarities and distances are measured in the Euclidean sense. This approach can be computationally expensive; it is successful only when all the sources have similar statistical characteristics and comparable reliabilities. In many application this assumption is not valid and therefore a more intelligent alternative approach has to be taken. In fact there is a research field called Consensus Theory that deals with finding consensuses, among members of a group of experts/sources and studies desired and undesired characteristics of consensus rules [7, 46].

Consider a consensus rule C_S for n data sources with probability measures $\{p_1, p_2, ..., p_m\}$:

$$C_S : [P(\Omega, S)]^n \rightarrow P(\Omega, S) \tag{3.3.17}$$

where $P(\Omega, S)$ is the space of all probability measures with σ-algebra S. There are several properties that are reasonable or desirable for a consensus rule, for example:

- Marginalization Property (MP)

$$C_S((p_1, p_2, ..., p_m)|T) = C_S(p_1|T, p_2|T, ..., p_m|T) \tag{3.3.18}$$

- Null Set Property (NSP)

$$p_1(X) = p_2(X) = ... = p_m(x) = 0 \rightarrow C_S(p_1, p_2, ..., p_m)(X) = 0 \tag{3.3.19}$$

- Weak Setwise Function Property (WSFP)

$$C_S(p_1, p_2, ..., p_m)(X) = F(p_1(X), p_2(X), ..., p_m(X), X) \tag{3.3.20}$$

- Strong Setwise Function Property (SSFP), also called strong label neutrality or the context-free assumption

$$C_S(p_1, p_2, ..., p_m)(X) = G(p_1(X), p_2(X), ..., p_m(X)) \tag{3.3.21}$$

In [7] the relationships among these properties are studied and the following theorem is proved:

Theorem: Suppose there is a family of consensus rules $\{C_S\}$ in Ω; then

1. MP is equivalent to WSFP,

2. (MP and NSP) is equivalent to SSFP

3. SSFP is achieved if and only if there exist non-negative numbers (weights) $\{\alpha_1, ..., \alpha_m\}$ with $\sum_i \alpha_i = 1$ such that for all σ-algebras S with $X \in S$ and all $p_i \in P(\Omega, S)$,

$$C_S(p_1, p_2, ..., p_m)(X) = \sum_{i=1}^{m} \alpha_i p_i(X) \tag{3.3.22}$$

This summation represents the so called LInear Opinion Pool (LIOP), which is one of the most commonly used consensus rules. This rule has a number of advantages and disadvantages. It is simple, it yields a probability distribution, and it has the MP and NSP properties. The weights $\{\alpha_i\}$ in this rule have an intuitive interpretation of relative importance or reliability of sources. There are also has some disadvantages; for example, the LIOP is not externally Bayesian, i.e. an LIOP based decision maker does not necessarily satisfy Bayesian rules. In order to avoid some of the shortcomings of LIOP, some authors have discussed the application of the LoGarithmic Opinion Pool (LGOP)

$$C_S(p_1, p_2, ..., p_m)(X) = \frac{\prod_{i=1}^{m} (p_i(X))^{\alpha_i}}{\int \prod_{i=1}^{m} (p_i(X))^{\alpha_i} d\mu} \tag{3.3.23}$$

where $\sum_i \alpha_i = 1$. It has been argued that the result of LGOP is unimodal and less dispersed than that of LIOP. It is externally Bayesian, but it assumes

the independence of sources. Also it has some disadvantages, e.g. it considers "zero" opinions as vetoes, and it is computationally more complex than LIOP. Because of the product form of this rule, the weighting factors in LGOP have less intuitive interpretations.

One of the main problems with these consensus rules is the selection of weights. The weights should represent an objective measure of relative importance and expertise of sources.

In part of our study we will use LIOP and LGOP for decision integration. Our sources are multiscale features and their reliabilities are their normalized discrimination powers.

In our analysis our decisions are based on similarity and dissimilarity measures rather than model-based probabilistic measures. In this context the consensus rules are less restricted; for example, they do not have to provide probability distributions and they may not necessarily satisfy Bayesian rules. For clarity, in the remainder of this section we explain our methodologies based on a specific set of feature sets as sources with defined similarity and reliability measures.

Following our projection-based feature extraction, each projection of the input pattern onto a discriminant vector u_i creates a resource for classification information and therefore a decision axis with a certain level of reliability and discriminatory power. The level of significance or reliability α_i of the decisions based on u_i is directly related to the class separation along that axis which is equal to the corresponding (normalized) eigenvalue in the LDA:

$$\forall(\lambda_i, u_i) \in (\Lambda^{(m)} \times U^{(m)}) : \quad \alpha_i = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{3.3.24}$$

For any test vectorized input pattern/image ϕ, we project it onto each of

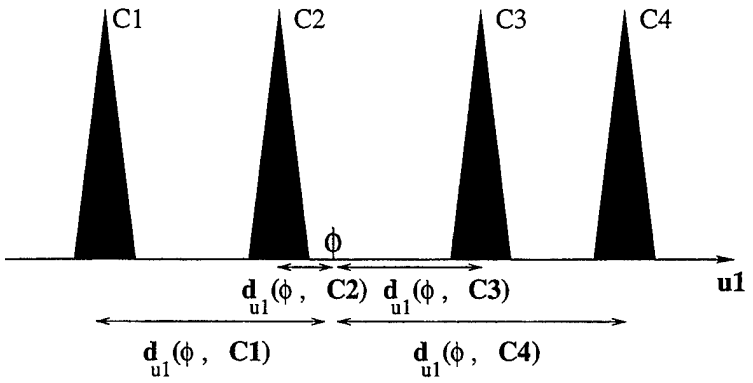


Figure 3.5: The raw distances between each test example and the known clusters along each discriminant axis result in the soft decision along that axis.

the top discriminant vectors u. Based on the distances between the resulting coefficients $\phi(u)$ and those of the existing templates ψ_u^c stored in the database, we estimate the level of similarity of the input image to each known class (see Figure 3.5):

$$\forall u \in U^{(m)} \quad : \quad \phi(u) =< \phi, u > \tag{3.3.25}$$

$$\forall c \in \widetilde{C} \quad : \quad \Delta_u(\phi, c) = |\phi(u) - \psi_u^c| \tag{3.3.26}$$

$$\pi_u(\phi, c) \quad = \quad 1 - \frac{\Delta_u(\phi, c)}{\sum_{c \in \widetilde{C}} \Delta_u(\phi, c)} \tag{3.3.27}$$

where $\pi_u(\phi, c)$ reflects the relative level of similarity between input ϕ and class c according to source $s \equiv u$ which has reliability α_u. Using our initial notation for soft decisions, we can put the $\pi_u(\phi, c)$'s into a decision vector

$$d_\phi^u = [\pi_u(\phi, c)]_{c=1}^L \tag{3.3.28}$$

Having determined our decision axis and the reliabilities, we can apply a probabilistic or an evidential scheme of multi-source data analysis to combine the soft decisions made based on the individual imprecise sources to obtain a more precise and reliable final result. The normalized similarity measures (π's) indicate the proportions of evidence suggested by different sources. They can

be interpreted as the so-called basic masses of evidence or they can be used as rough estimates of posterior probabilities given each measurement. From this stage on, a probabilistic or an evidential reasoning approach can be taken to combine the basic soft decisions. A comparative study of various probabilistic and evidential reasoning schemes is given in [52].

Similarly, working with distances as dissimilarity measures, one can combine basic soft decisions, and incorporate the reliability of each source, to define a reasonable measure of distance in the feature space. Although the most common measure used in the literature is Euclidean distance, as a more reasonable measure we suggest a weighted mean absolute/square distance, with the weights based on the discriminatory powers. In other words,

$$\delta_u(\phi, c) = \frac{\Delta_u(\phi, c)}{\sum_{c \in \widetilde{C}} \Delta_u(\phi, c)} \tag{3.3.29}$$

$$D(\phi, c) = \sum_{u \in U^{(m)}} (\delta_u(\phi, c) \times \alpha_u) \tag{3.3.30}$$

Therefore, for a given input ϕ the best match c^o and its confidence measure is

$$c^o = \text{argmin}_{c \in \widetilde{C}} \{D(\phi, c)\} \tag{3.3.31}$$

$$\text{Conf}(\phi, c^o) = 1 - \frac{D(\phi, c^o)}{D(\phi, c')} \tag{3.3.32}$$

where c' is the second best candidate. In this framework, incorporating collateral information or prior knowledge and expectations from context becomes very easy and logical. All we need to do is to consider each of them as an additional source of information corresponding to a decision axis with a certain reliability and include it in the decision process.

3.3.1 Incorporating Spatial/Temporal Context Information

Many signal/image processing tasks consist of local processing of data followed by a combination of results obtained from the local windows. The windowing

approach is sometimes used because of hardware limitations when considering large data sets or because of non-stationarity of the data. Local decisions made over small windows are myopic and are not reliable on their own. So one needs to devise methods of resolving the ambiguity and fuzziness of local decisions in a consistent way. In many segmentation/recognition tasks, sliding windows form a set of ordered observations about the signal/pattern. These sliding 1-D or 2-D windows may partially overlap each other. Also, in a multiresolution analysis there are sliding windows of variable sizes and scales that cover various parts of the signal. Based on the common coverage area of the windows one can define degrees of relevance and interrelationship among a set of observations in a neighborhood.

In our approach soft "decision vectors" computed for each block are integrated through weighted combination of decisions/votes obtained independently from neighboring blocks. The alternative, viz. using large windows, is not recommended because over larger windows signals are highly non-stationary and the corresponding features result from averaging over heterogeneous microstructures and therefore are less reliable. Large windows provide less spatial resolution, which is of great concern in segmentation of signals and images.

Our decision integration scheme combines context information from various sources based on their degrees of relevance R. For example, in terms of temporal/spatial context we can write

$$\forall \omega \in \Omega \qquad D(\omega) = \sum_{\omega' \in N_\omega} R(\omega, \omega') D(\omega') \tag{3.3.33}$$

where N_s is a neighborhood around the point s and $D(s)$ represents the decision vector at s. For temporal processing this degree of relevance may correspond to the overlap of intervals covered by adjacent time windows. Likewise, in terms

of spatial context, assuming the information contained in each block about a region is proportional to the area of their overlap, one can write

$$R(\omega, \omega') = \frac{A(W_\omega \cap W_{\omega'})}{A(W_\omega)} \tag{3.3.34}$$

where $A(.)$ is a function representing the area of its argument. As the image is analyzed by windows of size W in window shift steps of size w, the area contained in each block W_ω centered at point ω is partially covered by neighboring blocks $\{W_{\omega'} : \omega' \in N_\omega\}$ and contributes to their classifications. In the case of a 2-D sliding window on an image, it can be shown that

$$\begin{aligned} R(\omega, \omega') &= 1 - (|i| + |j|)w/W + |ij|(w/W)^2 \text{ for } -W/w < i, j < W/w \\ D(\omega) &= D(\omega) + R(\omega, \omega') \times D(\omega') \end{aligned} \tag{3.3.35}$$

where $(i, j) = \omega - \omega'$. Thus, after one complete scan of the image, the contributions of all neighboring blocks are added, and a combined vote for each macro-pixel of width w is obtained. Note that, following the principle of least commitment, thus far we have expressed all "decisions" as real vectors and no hard decision has been made.

Multi-resolution analysis of data (images) combines the results of classifications obtained at several scales. Classification is typically done from coarse to fine. We start with the low-resolution data to perform classification and use higher-resolution data when the confidence level obtained is not satisfactory. The combination of decisions can be performed based on our assumption about the spatial relevance function, using the fact that the windows on the low-resolution signal are actually projections of larger areas on the high-resolution view. In other words, one can combine decisions obtained at different scales based on their discrimination power and relevance to each block. The final ma-

jority votes and their confidence measures are based on the accumulation of soft decisions within and across scales and the closeness of the best class candidates.

The combination of weighted soft decisions is less susceptible to error than is each individual local vote. Note that the idea of soft decision propagation and integration within and across scales is dual to the lateral inhibition between decision units involved in one or several scales. The role of decision propagation profiles is similar, but not identical, to the role of inhibition profiles; one is democratic while the other is competitive.

Chapter 4

Signal and Image Classification

4.1 Introduction

During the last three decades, there have been many studies of classification and segmentation of signals and images. A variety of descriptors based on statistical, structural and spectral properties of the single or multidimensional signals are utilized to form the best sets of discriminant features. Parametric methods based on hidden Markov models [42] and Markov random fields [16], time/spatial domain approaches based on higher order moments [62], co-occurrence and correlation matrices [17], and frequency domain/filtering methods [38, 44] are among the major suggested schemes.

Also, different families of multiscale decompositions including WT, WP and Gabor filtering have been successfully applied to various classification and recognition tasks [13, 51]. Most of the proposed multiscale approaches to classification problems are based on decompositions, either independent of signal characteristics or based on an energy or representation criterion.

Based on our analytical results in Chapter 2, our objective in this chapter is show how adaptive discrimination based WP features can be used to design efficient and yet simple signal and image classification systems with very small-dimensional feature vectors. To show the effectiveness of our ideas for real signal and image classification and segmentation tasks, we will apply them to Automatic Target Recognition (ATR) and texture segmentation tasks. In these tests a set of real-aperture radar returns are used as examples of 1-D signals and a set of standard textures are used as a framework for 2-D image

classification/segmentation.

In the tests described in this and the following chapters we have used different QMF filters given in [2]; the results show that the choice of filters may have minor effects on the intermediate results, but plays an insignificant role in the final performance.

4.2 Classification of Radar Signatures

To show the effectiveness of the suggested feature extraction process in the discrimination of one-dimensional signals, we applied it to the classification of radar target signatures using the database provided as part of the ARPA University ATR initiative. In this section, without going into details about the theory of radar signatures, we use them as a framework for testing our scheme.

Millimeter Wave (MMW) Real-Aperture Radar (RAR) signatures play an important role in automatic target recognition. Due to their high range resolution, RAR signals can resolve tactical targets at ranges of several kilometers. On the other hand MMW radar range profiles are very noisy and their dominant peaks are sensitive to clutter and small changes in aspect angle. Therefore RAR contains valuable information which is difficult to extract. It has been argued that some of the difficulties may be overcome by using multiscale features.

The radar is transmitted in Right Hand Circular Polarization (RHCP) and received both in RHCP and LHCP, so each RAR return consists of two images, right-right (even) and right-left (odd) polarizations. The RAR data consists of FFT magnitude range profiles for each of five different targets. There are a total of 128 range profiles each with 128 resolution cells/samples. The targets are a T-72 Tank, a ZIL truck, an ASTRO multiple missile launcher, a TZM, and a BTR60 armored personnel carrier.

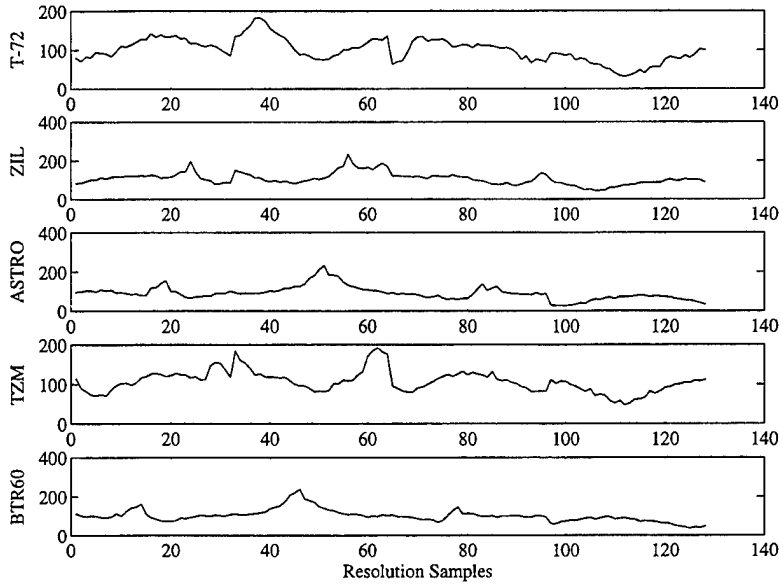


Figure 4.1: Example of radar target signatures for five different classes of targets.

There are four different views of each target: nose (0°), right side (90°), tail (180°), and left side (270°) views. 80 radar returns from five different stationary targets were used. For each target there are two radar returns for each of the two polarizations and the four view angles. Since the targets are assumed to be stationary, and in order to reduce noise, the average of every 32 channels was used. The data is divided into training and test sets. Figure 4.1 shows examples of such averaged signatures used in the classification test.

In these tests the idea of dimensionality reduction is applied to a two-level balanced wavelet packet tree. For each subband/node, second and third central moments are computed and $\{\mu_2, \mu_3/\mu_2\}$ is used as a feature vector. Figure 4.2 illustrates the separated clusters for five classes of radar targets where only the two most important features are used. All 16 radar signatures corresponding to one target are considered to be in one class. As Figure 4.2 shows, classification

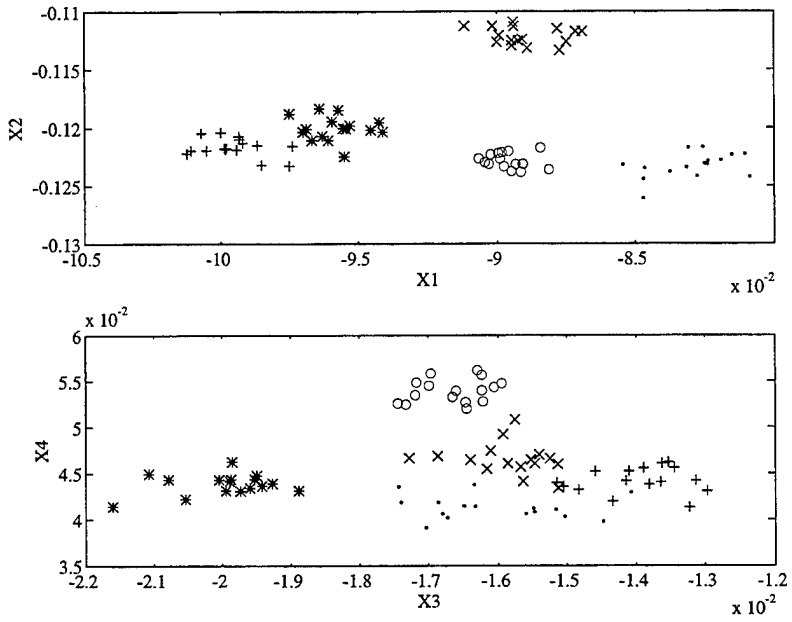


Figure 4.2: Clusters of feature points corresponding to five different classes of targets separated in the selected 2-D feature spaces: best two features (top), second best two features (bottom).

can be performed easily even with linear classifiers, and the distance between clusters allows us to achieve good classification results even in the presence of small Gaussian noise. For more details about this dataset see [27].

In this test a simple neural network is used as a "soft classifier". The network has two input, three hidden and five output units for five classes of targets. Results show about 1% error on the training set and about 2% on the test set. The confusion matrix is shown in Table 4.1. The training and test sets were similar because all targets were stationary and there were small changes across channels. This example shows how one can design a very simple and efficient classification system for a specific task.

Targets	T-72	ZIL	ASTRO	TZM	BTR60
T-72	20	0	0	0	0
ZIL	0	19	0	1	0
ASTRO	0	0	20	0	0
TZM	0	1	0	19	0
BTR60	0	0	0	0	20

Table 4.1: Confusion matrix in the radar signature classification test.

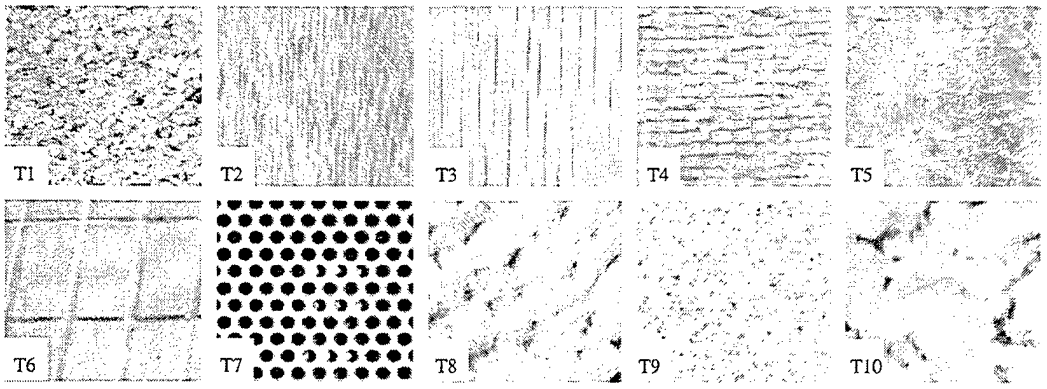


Figure 4.3: Some of the textures used in the classification experiments.

4.3 Texture Classification

The effectiveness of the suggested basis selection is further illustrated by applying it to image texture classification tasks. The input data consists of ten textured images shown in Figure 4.3. Feature vectors are computed from the second and third central moments (μ_2 and μ_3/μ_2 respectively) of the image subbands. Each of the training and test sets consists of about 100 image samples of each texture, selected randomly from 512×512 texture images. Each texture sample is a 64×64 pixel image. Figure 4.4 shows the class separation

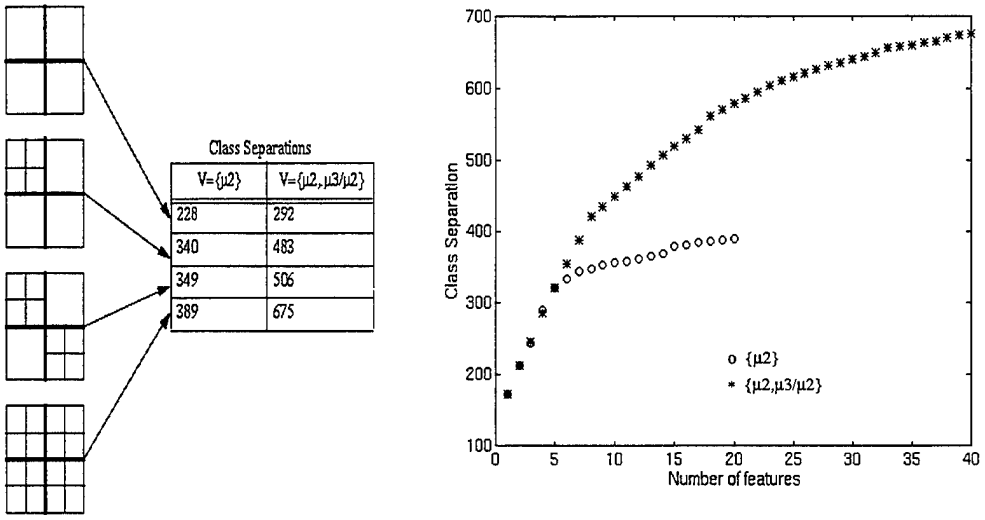
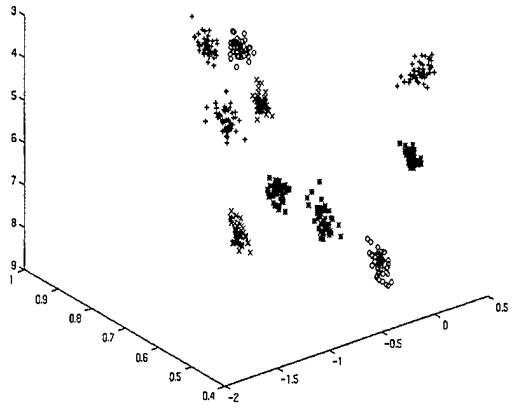


Figure 4.4: Decomposition results: selected subbands and computed class separabilities (left); increase in CCS with the number of features (right).

obtained at each level of the selected WP decomposition. The figure also shows the improvement obtained because of using both μ_2 and μ_3. The significant effect of using these features on classification performance also suggests that tree selection should not be based only on local energies (or second moments).

In Figure 4.5 some of the classification results for the ten textures in Figure 4.3 are given. Also their corresponding clusters in the best 3-D feature space based on the suggested dimensionality reduction idea are shown. Classification results are obtained based on class separation analysis and the suggested algorithm. The four most important features are selected. A simple feed-forward neural network [71, 49] with four input, eight hidden, and ten output units is used for classification. In some stages of building the wavelet packet tree (Figure 4.4), energy and separation based criteria suggest different strategies for extending the decomposition.

Textures	Correct Classification Rate
T1	100%
T2	98%
T3	99%
T4	100%
T5	100%
T6	98%
T7	97%
T8	100%
T9	100%
T10	99%



Window Size	Class Separation
64x64	675
32x32	270
24x2	140
16x16	70

Figure 4.5: Some of the classification results (left); clusters in the selected 3-D feature space (right).

Finally, some tests on 90° and 180° rotated input textures are performed. The classification errors increase by between 1% and 14%, depending on the textures. This is partly due to the separability of the filters and partly because of the basis selection algorithm. The algorithm by its nature looks for common features among all examples of the same class as well as features that discriminate examples in different classes. So if all examples of a directional texture are selected from the same image, it is expected that the algorithm will pick up some directionally sensitive features. In general, depending on the task, different rotated versions of a directional texture may or may not be "defined" as the same texture, and this has to be considered in the class separation analysis. To test this idea, for each texture we included some rotated examples defined as being in the same class, and we applied the same feature selection algorithm. Although the rotated examples were not included in the training of the classifier network, the resulting classification performance on rotated input textures improved significantly, e.g. from about 96% to 98% for 64×64 windows.

Despite the simplicity of the system, the results are comparable to other

recently published texture classification schemes [13]. Note that in this approach the selection of basis/features in performed once for all intended classes whereas in [13] for every input example the decision about the tree structure has to be made on-line and separately. Also, since our suggested basis selection is based on observations over a group of examples for each, class the resulting tree structure is less susceptible to errors in the noisy examples.

In order to test the effect of windowing on class separability, we tested four different window sizes, as shown in Figure 4.5. As expected, whenever we reduce the window size for better localization, we lose class separation, which results in less accurate or less certain local decisions. Thus the need for soft local classification and context-dependent decisions is apparent.

4.4 Texture and Image Segmentation

For the texture segmentation tests the features are based on segmentation windows, i.e. the central moments are computed over small windows on the decomposed image. Because of the down-sampling involved in the transform, the corresponding window sizes for the sub-bands at the k^{th} level of the tree are $W/(2^k)$. Therefore the depth of the tree is limited by the size of the input window and the nature of the signals to be classified. Also, the order of the filters in filter bank implementations should be smaller than the window size to avoid the dominance of window boundary effects on the resulting feature computation.

Figure 4.6 shows the segmentation results for three visually similar textures. In the test a window size of 16×16 pixels, with 8-pixel overlap, is chosen and decision integration is used. In this test we used a simple two-layer neural network with just three input, four hidden, and three output units to build our soft classifier. As this figure shows, results comparable to those of other texture

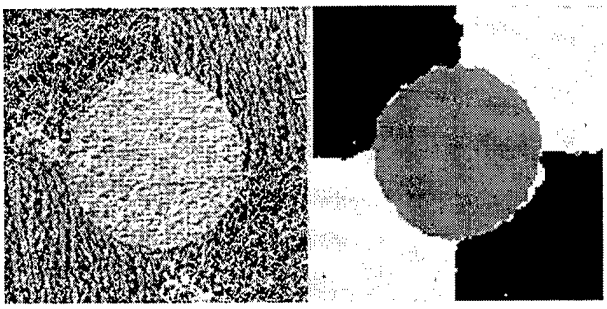


Figure 4.6: Example of a texture segmentation using a reduced two-dimensional feature space: (left) original image, (right) segmentation result.

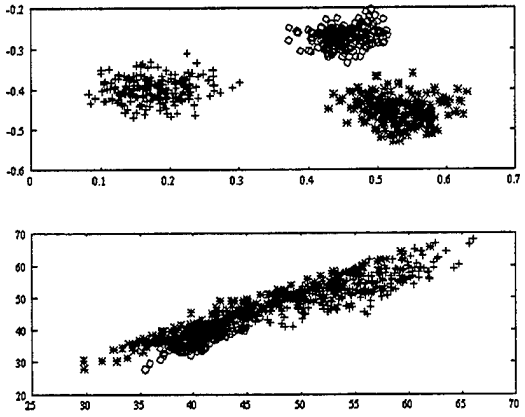


Figure 4.7: The clusters corresponding to three textures in the segmented image, based on the separability criterion (top) and based on an energy criterion (bottom).

segmentation schemes, including wavelet-based systems [38, 55], are obtained, using a generic scheme of low complexity and with a small number of features.

Figure 4.7 compares the cluster separations in the feature space when the best feature vectors are selected according to the suggested class separability based linear map to those obtained using dominant energy based approaches. As this example illustrates, with the same feature size the suggested method provides a very good separation of classes.

Chapter 5

Layout-Independent Document Page Segmentation

5.1 Introduction

Recent advances in information and communications technologies have increased the need for, and therefore the interest in, automated processing of documents. Efficient storage and transmission of documents as well as archiving and information retrieval for document databases and "digital libraries" have become important research issues.

Two important tasks of most document processing systems are page decomposition and optical character recognition (OCR). For coding or understanding a document it is essential to identify text, image and graphics regions, as a physical segmentation of the page, in order to be able to process it appropriately. For example, one must identify the text regions before applying OCR algorithms, and identify graphics regions before attempting to interpret or vectorize them. Physical page segmentation may also be required for the task of functional layout analysis, to identify the document's type (e.g. journal, memo, check, etc.) or to generate hypotheses as to the components' roles and logical functions (title, abstract, footnote, caption, signature, table, etc.). As part of a source compression scheme one may consider a document image as a composite source, decompose it into text, image and graphics sub-sources where each sub-source has more "homogeneous" outputs, and design separate coding schemes for each sub-source, based on appropriate fidelity criteria [14]; see Figure 5.1.

Page segmentation and layout analysis methods described in the literature make use of well-known image processing tools which can be broadly classified

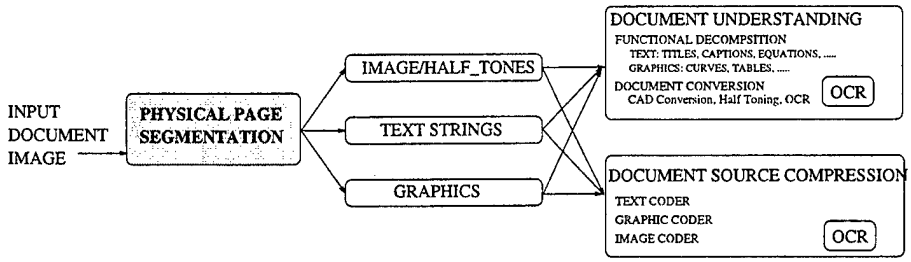


Figure 5.1: The role of page segmentation in document image processing.

as bottom-up and top-down [81]. Bottom-up tools, such as connected component analysis [32], start from the pixel level and merge regions together into larger and larger components (e.g. first characters, then words, text lines, paragraphs, etc.). Top-down techniques apply a priori knowledge about the page to hypothesize and split the page into blocks which are subsequently identified and subdivided further. For example, one may first locate major columns and then split them further into paragraphs, text lines, and eventually words. Examples of algorithms which use a top-down approach include recursive projection profile cuts [87, 84], run length smoothing and constrained run length [85]. In general, most approaches use a combination (or hybrid) of top-down and bottom-up techniques.

One method, described in [84, 50], uses projection profiles and an *X-Y tree representation* of documents to exploit the fact that the components of printed pages (e.g. text blocks, tables, figures) can often be bounded by rectangular blocks. The root of the tree is the entire page and after iterated subdivision, based on changes in the projection profiles, each rectangular block in the page is represented by a node in the tree. This results in a hierarchical block segmentation of the page.

The *constrained run length* algorithm starts from the binary image and re-

places every string of contiguous 0's (corresponding to white pixels) of length less than a predetermined constant by a string of 1's (i.e. black pixels) of the same length [14, 85]. This binary smearing process is performed in both horizontal and vertical directions. The final bit map is obtained from the logical "AND" of the two outputs. The vertical and horizontal constraint lengths are determined from anticipated inter-component spacing. Clearly these methods are dependent on assumptions about component sizes, component proximity, and page orientation. A survey of the most common techniques is contained in [81].

Recently, a more flexible method of page segmentation based on *analysis of background white space* has been explored by several authors [3, 64, 73]. The scheme is based on tracking major white spaces between printed components to identify region boundaries. This method is based on relatively few assumptions and provides good results even for skewed pages or documents with complex layouts. For identification of component type, some approaches use simple statistical tests to classify detected major blocks as text or non-text regions [84]. Black pixel density, black/white ratio or transitions, average vertical or horizontal run lengths, and row-by-row cross-correlations [65] are some of the features used in these post-classification stages.

Each of the above techniques relies to a different extent on prior knowledge about the generic document layout structure, such as rectangularity of major blocks, consistency in horizontal and vertical spacing, and independence of text, graphic and image blocks, and/or assumptions about textual and graphical attributes such as font size and text line orientation. Utilizing knowledge about the layout and structure of documents results in simple, elegant and efficient page decomposition systems but also limits the range of applicability of the al-

gorithms. For example, methods based on projection profiles fail if the page layout is complex, the page is skewed, or text strings on the page have different orientations. There are methods of estimating and correcting the skew angle [61, 4], but they have limited ranges and add to the complexity of the system. Methods based on smearing or white spaces are sensitive to character sizes as well as line and character spacing. They may also fail when text regions touch images or are embedded in them.

In some applications it is desirable to have segmentation methods that do not assume a priori knowledge about the content and attributes of text, or about the boundaries of major blocks. Such approaches should be robust to skew, noise and other degradation. Some of the difficulties, shown in Figure 5.2, which are common in general classes of documents, and which make these goals hard to attain include:

- Noise and degradation caused by copying, scanning, transmission or aging.
- Page skew and text lines with different orientations on the same page.
- Text touching or overlapping with image and graphics components.
- Combinations of varying text and background gray levels (e.g. inverted text).
- Complex and irregular layout structures that are common especially in non-technical documents. Document objects may not have rectangular or even convex boundaries and may be embedded in one another.
- Curved lines or multi-column pages where text lines in the two columns are not of the same size and/or are not aligned.
- Differences in language, font size and other textual attributes.

Figure 5.2: Examples of difficult cases for document page decomposition

See Figure 5.3, for example, where despite the fairly simple page layout, the projection profile based system has difficulty because no line across the page can separate text and image regions. Any of these problems may cause failure of the previously described techniques, and it is not uncommon to see combinations of the cases described above, on the same page. Based on these observations, a texture-based segmentation method for extracting text has been suggested by Jain et al. [45]. The approach uses multi-channel Gabor filters

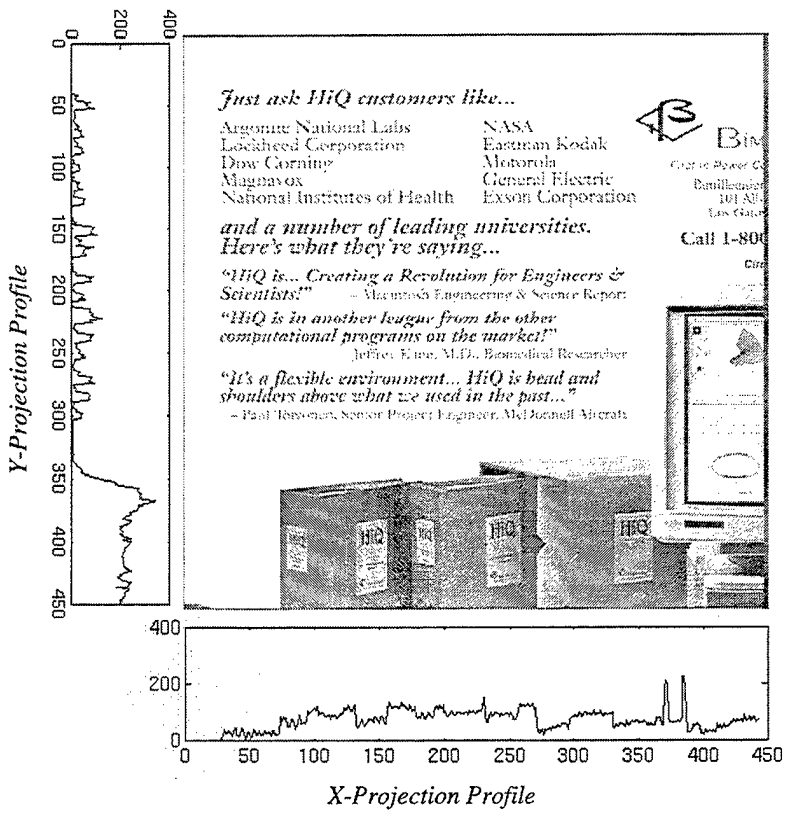


Figure 5.3: An example of a document with simple layout for which projection profile based methods fail.

as the input features of a classifier whose outputs are directly used to identify text. This method is computationally expensive and does not provide a means for incorporating context information.

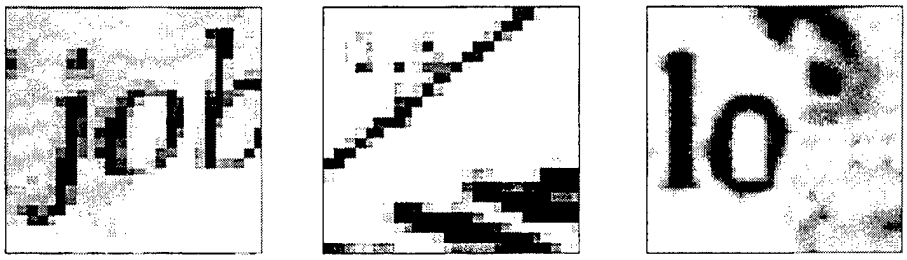
In this chapter, text, image and graphics regions in a document image are described as three classes of textures. The idea can be justified by the fact that humans can identify document objects easily even from low-resolution images or from distant views of a document page. This shows that the physical segmentation of a document is not detail- or content-sensitive, and like texture segmentation, is a low-level vision process. Given the following considerations, some of the existing texture segmentation techniques [38, 43, 17] can be modified and used to identify these regions on the page. One distinctive feature of this task,

compared to texture segmentation problems, is that there are large inter-class, as well as intra-class, variations in the textural features. Text and graphics are texturally quite different, but different images may also contain scenes of significantly different textural structure, as is the case for texts of different fonts, sizes, and even languages. The variabilities are even more pronounced for graphics.

An important observation about the human audio/visual recognition system, which is the backbone of most artificial neural network models, is the improved recognition power gained through interactions of simple computational units. Each local process can be as simple as projection or filtering, passing through simple non-linearities, etc. In addition, all decisions, at least in low-level vision, are non-binary, highly context-dependent, and based on multi-scale representations of the input signals/images [83, 93]. With these motivations we search for consistent multi-scale context dependent schemes based on soft local decisions.

Our method is based on the fact that there is some uncertainty associated with the local decisions over small windows, due to the limited view of the signals and/or to the randomness and ambiguity inherent in the problem, or even to the presence of multiple classes, overlapped or adjacent, in the same window. Using large windows is not recommended, because over larger windows the signals are highly non-stationary and features computed based on heterogeneous microstructures are less reliable. Also, larger windows provide less spatial resolution, which is of great concern in segmentation schemes.

In the document domain, image sub-blocks may contain text, image and graphic sub-regions adjacent to, or overlapping, one another (Figure 5.4). Such situations may occur on boundaries, where, for example, text lines come close to or touch image regions, or when major text regions occur in an image or on a graph. In such cases it is not appropriate, even for an optimally designed



(a) (b) (c)

Figure 5.4: Examples of image blocks for which there is no correct hard decision: (a) text overlapped on an image, (b) both text and graphics in a block, (c) both image and text in a block.

classifier, to make hard (binary) local decisions. These cases exhibit an inherent fuzziness of class membership which does not come from the noise or randomness, and they support our claim that soft local decisions are more realistic and efficient. The uncertainties reflected in soft decisions are then reduced by propagating and integrating decisions made independently in the neighborhoods within and across scales.

In this chapter we propose the utilization of multiscale representations in a soft decision framework for the task of layout-independent physical page segmentation. In an attempt to handle even the most difficult cases of segmentation, we make few assumptions about the document's textual and graphical attributes and layout structure. The system is designed so that as hypotheses about document components are generated and verified, more domain-specific processing may occur.

The organization of the chapter is as follows: In Section 5.2 the pyramidal wavelet transforms and their generalized form, wavelet packets, are introduced. These transforms are used to compute the input feature vectors at different scales/resolutions. In Section 5.3 we describe how multi-scale feature vectors are used for "soft classification" of small windows and how the "propagation"

and "integration" of those soft decisions, within and across scales, can improve the overall classification and segmentation performance. Some issues about incorporating prior knowledge of structure (or a model of the document) and the notion of "biased voting" are then addressed. Some comments about the post-processing stages are given in Section 5.4. Page segmentation experiments, showing the performance of the method, are then described in Section 5.5. Finally the results are discussed and some suggestions about possible variations and future directions are made.

5.2 WP Decomposition of Document Pages

The fact that document objects (e.g. characters and lines) appear at multiple scales and our belief that physical segmentation is a low-level vision process similar to texture analysis suggest that the use of multi-resolution representations is appropriate. There are several classes of multi-scale decompositions that seem to be biologically plausible and that have been successfully employed in modern signal processing schemes. In this paper we use wavelet-based decompositions (Figure 5.5) because they provide perfectly reconstructible decompositions through fast algorithms [56, 22]. For a given class of signals, wavelet packets can be adaptively designed to obtain compact representations that meet a predetermined objective criterion [19]. Also the perfectly reconstructible multiscale representation, employed in our system, can be used as part of a multi-scale document compression scheme.

Following our discussions about efficient discriminant feature extraction in Chapter 2, we build the tree in such a way that the spread of feature points in each class becomes smaller and at the same time clusters become farther apart. The feature vectors consist of central moments computed over local windows on

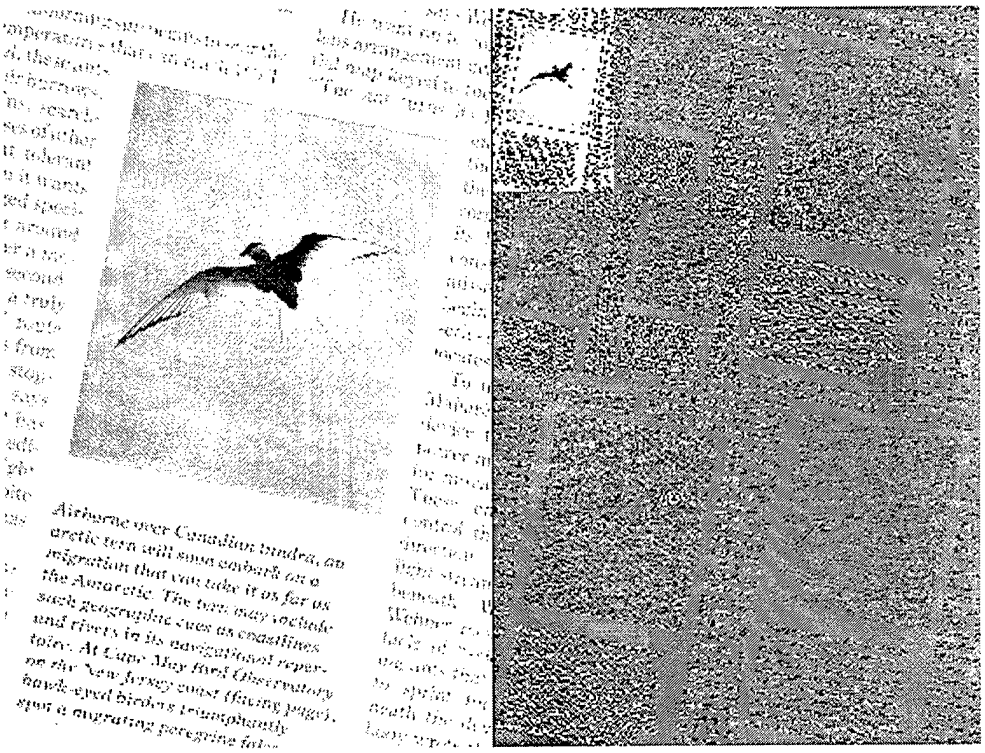


Figure 5.5: An example of a pyramidal WT on a document page: the original image (left); the wavelet decomposition (right).

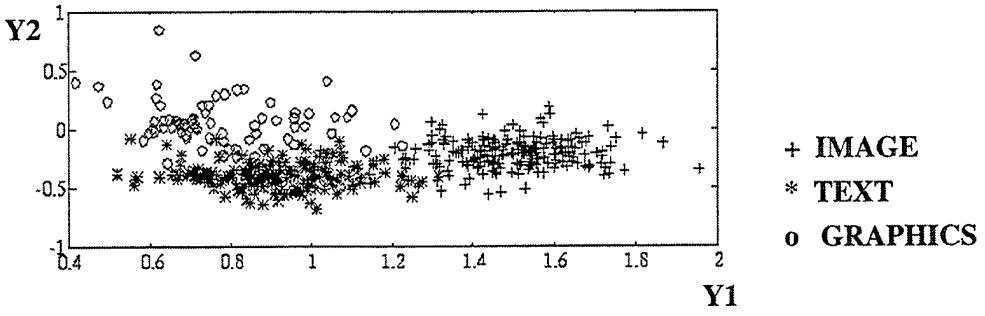


Figure 5.6: Clusters in the feature space may overlap; the three clusters shown correspond to text, image and graphics subblocks in the database.

different subbands.

Figure 5.6 shows three clusters of feature points corresponding to text, image and graphics blocks in a database. The features are extracted based on the maximum class separability criterion. Considering the nature of mixed classes,

overlap of clusters may be inevitable. In the following experiments, a pyramidal wavelet transform and separability-based wavelet packet trees are used.

5.2.1 Knowledge-based Post-processing

In some applications we may wish to incorporate constraints into our decision based on a priori or derived knowledge about the domain. For example, we may observe patterns of data that result from rules subject to physical constraints. In the case of document segmentation into text, graphics and image components, regions are typically rectangular, text symbols are arranged along straight lines, and small graphics and images within text regions are unlikely. For more structured classes of documents, blocks such as the title, abstract and page number are expected to be in specific regions of the page, and even to have specific attributes and formats. In general these constraints are task-dependent.

Although such constraints are often considered in higher levels of processing, one may also utilize them in the early stages of classification to get more reliable results. In the context of the described majority vote method, this idea can easily be fit into the system without increasing its complexity by a biased voting scheme. Our expectation about observing a certain class of patterns in a certain part of the scene is reflected in a biased vote in favor of a particular class over that region. In this case the system does not start from an all-zero vote matrix V_{tot}, but at each position a small non-zero initial vote is already given to the class(es) that have been frequently observed in that location. The "biased voting" can be viewed as not starting from the middle of the fuzzy decision cube (i.e. the most fuzzy point), but deviating from it in favor of one of the classes (in the corners); see Figure 5.7.

The prior vote or decision bias for each macro-pixel can be computed from

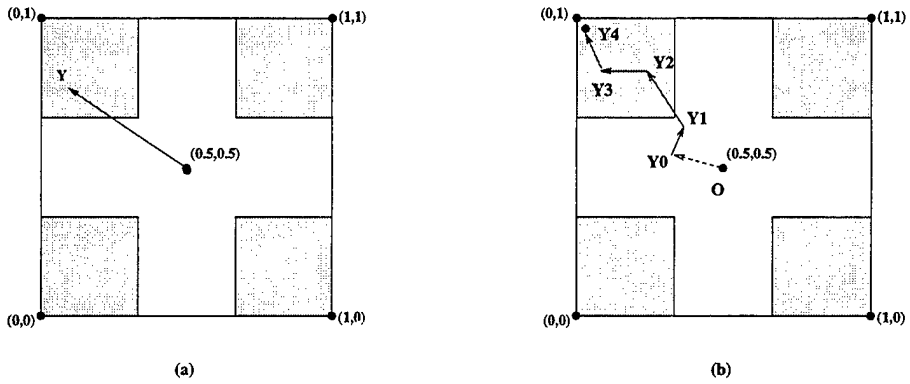


Figure 5.7: Fuzzy decision square: (a) when we make a one-shot decision; (b) when a weighted sum of soft decisions is used. A final decision outside the gray area has a low level of confidence.

the empirical distributions of document objects in the labeled documents of the training set Γ:

$$\begin{aligned} \forall \omega \in \Omega \qquad Y_0^{(c)}(\omega) &= Pr\{C = c|\omega\} \\ &\approx \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} I\{L(\gamma, \omega) = c\} \text{ for } c = 1, 2, ..., L \end{aligned} \tag{5.2.1}$$

where $L(\gamma, \omega)$ is the label of macro-pixel ω derived from ground truth data, L is the number of classes, and $I\{.\}$ is the indicator function. This initial vote can be used if the type or class of document is already determined by other means or is known a priori.

The spatial patterns of combined soft decisions in vote matrices directly reflect the locations, shapes and classes of major blocks. In some cases, however, obtaining a more precise segmentation requires some knowledge-based post-processing to incorporate additional knowledge about the structure. In such cases the spatial pattern of votes also provides a convenient starting point to apply constraints and performs further analysis. The local nature of texture-based segmentation of documents sometimes results in sparse mis-classified regions. For example, the textural characteristics of the leaves on a tree in the

image part of a page may locally resemble, and therefore be classified as, text elements. Some of these sparse mis-classifications are corrected after the decision integration process, but in some cases higher-level and knowledge-based post-processing may be needed. Certain rules and constraints can be considered, with different ranges of generality and therefore applicability. These rules put restrictions on the absolute locations of document objects on the page and their positions relative to each other. A stroke embedded in a text region should be interpreted as a character, but the exact same pattern in the margin should be interpreted as noise. Similarly, one might hypothesize that "small" blobs labeled as graphics, as well as small text-like regions in an image block with no collinearity between them, have been misclassified and therefore can be relabeled. Depending on how likely they are to be encountered in an application, one can put restrictions on the shapes and minimum sizes of the labeled regions. For example, one may make use of the fact that text elements tend to be organized into lines, and are typically left-justified in groups that fall into columns of rectangular shape. Although restrictive, assumptions about the minimum and maximum character sizes, as well as minimum sizes of image and graphical objects, may be derived and utilized.

By applying these structural rules, one can hypothesize and remove logically undesirable gaps and noise-like small blobs of misclassified regions. Similarly one may complete and rectify region boundaries and fit them with polygons or rectangles, to obtain a parametric layout representation consistent with derived knowledge of the domain.

As an example, and without going into the details of imposing layout-specific constraints in our experiments, we establish a structural hierarchy. It suggests that a text region is typically uniform and contains no graphics or image com-

ponent, but a graphic or image region may have subordinate text. The text must, however, have a uniform background and extend over a relatively large area with respect to the font size. These properties are enforced by applying morphological operations of each type to filter out noise which is too small to constitute a document component. Constraints on the sizes of regions, as described in Section 5.1, are implemented using 3×3 morphological kernels. Given a constraint on the size of a text region, we perform a closing operation on the image, to eliminate regions which appear locally as text. For the image resolution and window size used in the experiments described in the next section, a six-step closing operation eliminates a majority of the noise regions. Since a six-step closing operation is approximately the size of a capital "M" in a 9pt font, we do not have to be concerned that larger text regions will be eliminated. For text which actually appears as part of the image, higher-level constraints must be used, if possible, to associate the text with the image.

5.3 Experiments

To show the effectiveness of the suggested soft decision integration method it has been applied to document page segmentation.

5.3.1 Input Representation and Training Set

In the following experiments both wavelet transform and wavelet packet decompositions are used as input signal representations. In the first two examples, features are computed from a two-level wavelet transform. At each level, only detail subbands are used and there is one classifier for each scale. The result of classification at the two scales are combined as described in Section 3.3. In the other examples, features are selected using a separability measure on the wavelet packet decomposition. For these experiments, six features that contain

the highest classification information, based on the previously used separability measure, have been selected and used.

The input data consist of several gray-scale document pages scanned at 200 dpi and the input features are the second and third central moments (μ_2 and μ_3) of the image subbands computed over small windows W on the decomposed image.

The training set consists of about 200 samples from each of the text, image and graphics sub-blocks. These 16×16 pixel sub-blocks are extracted randomly from several document pages. In order to avoid over-training, a "validation" set is used to test the performance of the network, after every ten iterations, during the training stage. As training proceeds, errors on both the training and validation sets decrease. Training is suspended as soon as the error in the validation set starts increasing. If the desired performance is achieved, the process stops; otherwise, part of the validation set is included in the training set and training proceeds on the augmented training set.

5.3.2 Network Description and Training

In all of the experiments, multi-layer feed-forward neural networks are used as the soft classifiers. The network consists of six input, eight hidden, and three output units. The input units are linear, whereas the hidden and output units have sigmoid nonlinearities. A conjugate gradient method is used for fast convergence of the supervised learning algorithm [88].

The three outputs correspond to text, image, and non-text non-image classes. In other words, any sub-block not identified as text or image is considered as "graphics". Blank regions are detected separately in a straightforward way. The outputs can take values in $[0, 1]$ and the network is trained in such a way that

these outputs provide soft non-binary decisions about the class memberships of the input image blocks. This is essential because, as mentioned above, small regions may locally resemble more than one class, or the image sub-block may be composed of text, image or graphics subregions. In such cases, during the training, outputs corresponding to text, graphics and images are required to take target values roughly in proportion to the fraction of block area they occupy. Including such composite blocks in the training set results in better performance on the boundaries. If a decision integration stage is used the result will be much less sensitive to these adjustments.

Despite its significance, the effect of a suitable output representation is sometimes overlooked. In fact in some cases, such as design and training of soft decision based classifiers, the choice of output representation can be as important as that of input representation. In this experiment, in order to provide the learning algorithm with a consistent set of input-output pairs the following procedure has been implemented: Assuming that the data in the training set is labeled correctly and consistently, for any macro-pixel ω and any class c one can compute the desired soft decision for class membership as

$$\forall \omega \in \Omega \qquad L_c^{(\text{Target})}(\omega) = \frac{1}{|W|} \sum_{x \in W} I(\text{Lab}(x) = c) \tag{5.3.2}$$

i.e., the relative number of pixels in the window labeled as c. This form of target value computation is consistent with our assumption about spatial relevance. It is also a suitable means of determining soft local decisions when mixed classes are present in the window, e.g. overlapped and adjacent text and image components in the area covered by W. These labeled examples are the basis for learning the fuzzy membership functions in our multidimensional feature space.

5.4 Results and Discussion

The decision integration scheme described in this chapter has been used to identify text, images, and graphics regions. We have tested our approach on a number of document images which are difficult for other approaches due to multiple scales and complex (but not unusual) layout of the components.

Figure 5.8: For this example feature vectors are computed from the wavelet transform with two levels of decomposition. The example shows the advantage of using decision integration in identifying major document blocks. In this image we have a skewed page with multiple columns where the text lines of the different columns are not aligned, the image is surrounded by text, and there are two text fonts/sizes on the page.

Figure 5.9: This is the same example that was shown in Figure 5.3. In this test we have used only two features extracted from the wavelet packet decomposition of the document images in our database; the features are selected based on the aforementioned separability criterion. This is an example of a page with different font sizes and non-rectangular object boundaries.

Figure 5.10: This example shows the effectiveness of the suggested scheme for cases where image and text regions are very close to each other and regions do not have rectangular or even convex boundaries. For this example the results of prescribed post-processing based on morphological operations are also illustrated.

Figure 5.11: This example shows a very difficult scenario where text is embedded in the image, i.e. where different classes of objects are overlapped. Even in this case our method provides good results.

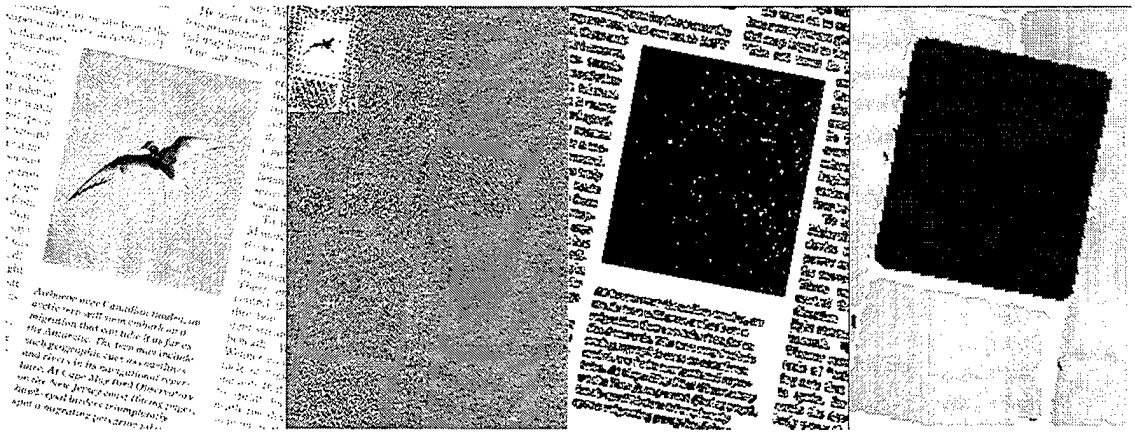


Figure 5.8: Page segmentation results for a document image. (From left to right): original image, two-level wavelet decomposition, segmentation without decision integration, segmentation with decision integration. Dark gray and light gray represent image and text areas respectively.

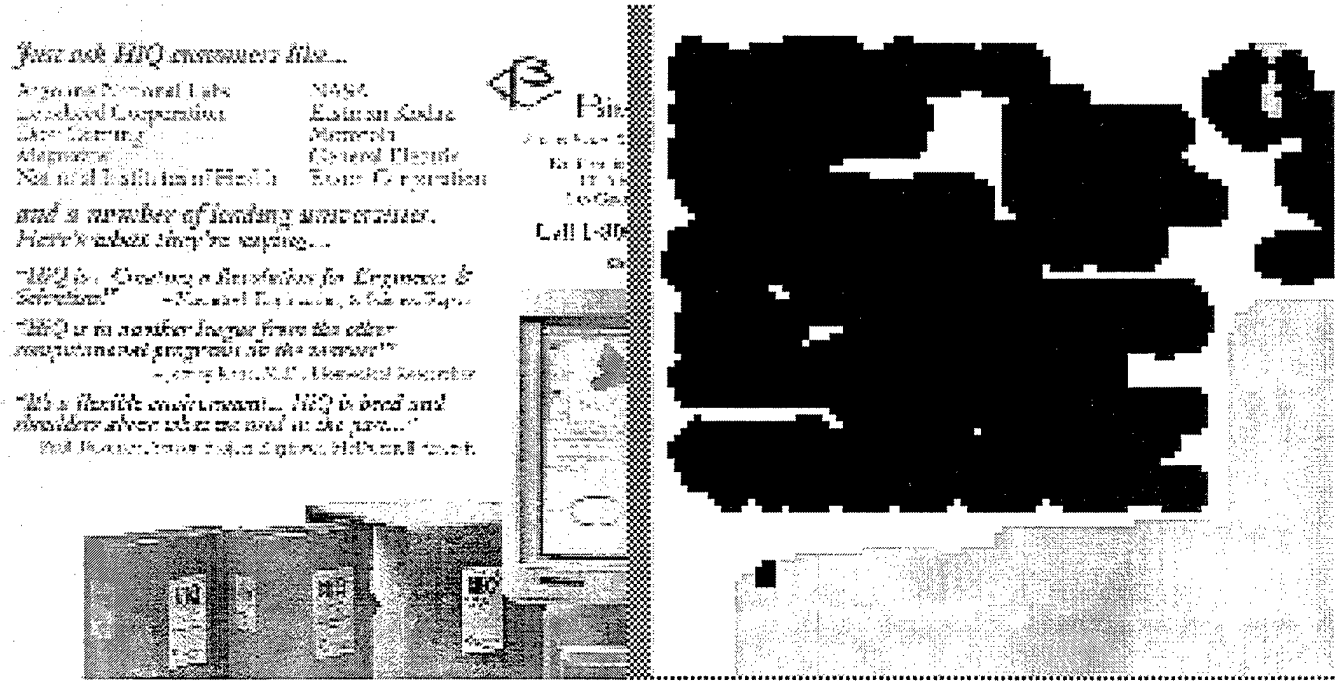


Figure 5.9: Page segmentation results for a complete page, with multiple font sizes

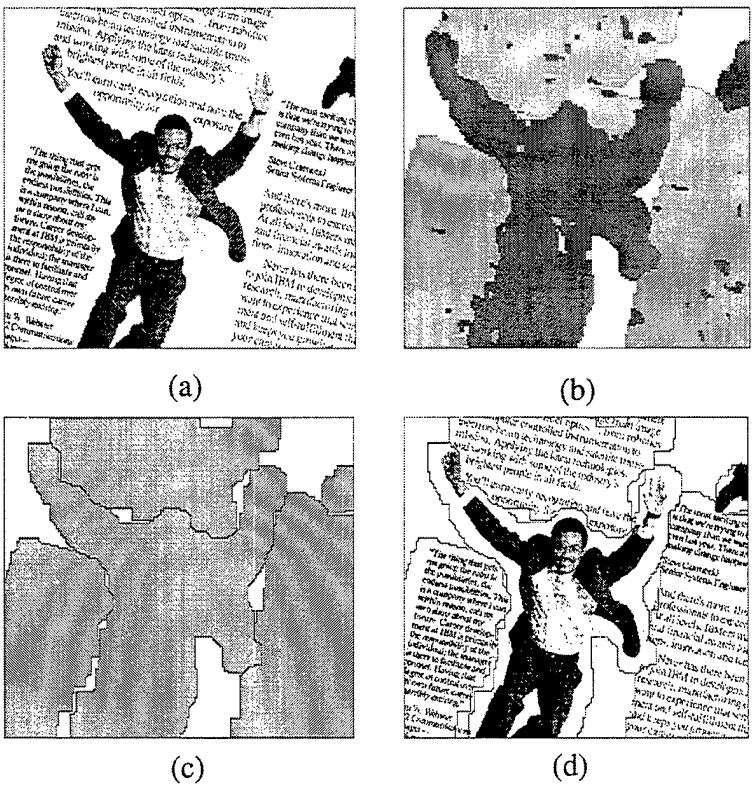


Figure 5.10: An example of a difficult a segmentation: irregular and non-convex image boundary very close to text. (a) Original image; (b) segmentation without post-processing; (c) result after post-processing; (d) final segmentation.

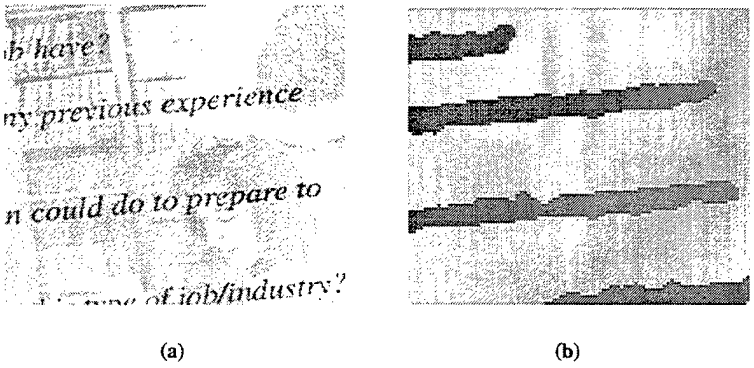


Figure 5.11: An example of a difficult segmentation: text embedded in an image.

5.5 Conclusions

Our experiments have shown that good page segmentation results can be obtained using context information through propagation and integration of soft decisions based on multiscale features. The improvement resulting from decision integration is significant when confident hard local decisions cannot be made because of poor features, poor resolution, windowing considerations, noise and/or the inherent fuzziness of the classification task. Very good performances have been obtained on complex document layouts, using simple feature sets and classifiers. A majority of the calculations and decisions are made independently and in parallel without any iterative stages. They are therefore well adapted to distributed and parallel algorithms and architectures, which promise robust and fast implementation.

As mentioned earlier, the physical segmentation process is typically part of a larger system and, depending on the application, it may be followed by a functional decomposition module in a document understanding or source encoder system. This work can be extended in a number of directions. For example, one may estimate the text font and text line orientation for all blocks labeled as text, in order to prepare them for OCR algorithms. Also, one can find parametric representations of labeled regions referenced to the page so that their logical identities can be defined or searched for in a database. Other feature vectors such as multi-channel Gabor filters, co-occurrence matrices, or sets of document-specific features, such as black pixel densities or black/white transitions, can be explored to produce accurate local decisions. The basic idea of incorporating context information through integrating soft local decisions can be applied to other image and signal segmentation tasks.

Chapter 6

Automatic Face Recognition

6.1 Introduction

Inspired by humans' ability to recognize faces as special objects, and motivated by the increased interest in commercial applications of automatic face recognition as well as the emergence of real-time processors, research on automatic recognition of faces has become very active. Studies about the analysis of human facial images have been conducted in various disciplines. These studies range from psychophysical analysis of human recognition of faces and related psychovisual tests [5, 23] to research on practical and engineering aspects of computer recognition/verification of human faces and facial expressions [91] or race/gender classification [9, 35].

The problem of Automatic Face Recognition (AFR) is a composite task that involves detection and location of faces in a cluttered background, facial feature extraction, subject identification, and verification [74, 15]. Depending on the nature of the application, e.g. image acquisition conditions, size of database, clutter and variability of the background/foreground, noise, occlusion, and finally cost and speed requirements, some of the subtasks become more challenging than others.

Detection of a face or group of faces in a single image or a sequence of images, which has applications in face recognition as well as video conferencing systems, is a challenging task and has been studied by many researchers [40, 15, 92]. Once the face image is extracted from the scene, its gray level and size are usually normalized before storing or testing. In some applications, such as identification

of passport pictures or mug-shots, conditions of image acquisition are usually so controlled that some of the preprocessing stages may not be necessary.

One of the most important components of an AFR system is the extraction of facial features, which attempts to find the most appropriate representation of face images for identification purposes. The main challenge in feature extraction is to represent the input data in a low-dimensional feature space in which points corresponding to different poses of the same subject are "close" to each other and "far" from points corresponding to instances of other subjects' faces. However, there is a lot of within-class/subject variation due to differing facial expressions, head orientations, lighting conditions, etc., which makes the task more complex.

Closely tied to the task of feature extraction is the intelligent and sensible definition of similarity between test and known patterns. The task of finding a relevant distance measure in the selected feature space, and thereby effectively utilizing the embedded information to accurately identify human subjects, is one of the main challenges in face identification. In this chapter we focus on the feature extraction and face identification processes.

Typically, each face is represented using a set of gray-scale images/templates, a small-dimensional feature vector, or a graph. There are also various proposals for recognition schemes based on face profiles [90] and isodensity or depth maps [36, 60]. There are two major approaches to facial feature extraction for recognition in computer vision research: holistic template matching based systems, and geometrical local feature based schemes and their variations [15].

In holistic template matching systems each template is a prototype face or face-like gray-scale image or an abstract reduced-dimensional feature vector which has been obtained through processing the face image as a whole. Low-dimensional representations are highly desirable for large databases, fast

adaptation, and good generalization. Based on these needs, studies have been performed about the minimum acceptable image size and the smallest number of gray levels required for good recognition results [74]. Reduction in dimensionality can also be achieved using various data compression schemes. For example, representations based on Principal Component Analysis (PCA) [18, 48, 80, 66] and Singular Value Decomposition (SVD) [77] have been studied and extensively used for various applications. It has also been shown that the nonlinear mapping capability of multilayer neural networks can be utilized and the internal/hidden representations of face patterns which, typically, are of much lower dimensionality than the original image, can be used for race/gender classification [9, 35]. Some of the most successful AFR schemes are based on the Karhunen-Loeve Transform (KLT) [48, 66], yielding so-called eigenfaces. In these methods the set of all face images is considered as a vector space and the eigenfaces are simply the top principal components of this "face space"; they are computed as eigenvectors of the covariance matrix of the data.

In geometrical feature-based systems one attempts to locate major face components or feature points in the image [21, 58, 69, 76]. The relative sizes of and distances between the major face components are then computed. The set of normalized size and distance measurements constitutes the final feature vector for classification. One can also use the information contained in the feature points to form a geometrical graph representation of the face that directly shows the sizes and relative locations of major face attributes [58]. Most geometrical feature-based systems involve several steps of window-based local processing, followed by iterative search algorithms, to locate the feature points. These methods are more adaptable to large variations in scale, size and location of the face in an image but are more susceptible to errors when face details are occluded by

objects, e.g. by glasses, by facial hair, due to facial expressions, or by variations in head orientation. Compared to template/PCA based systems, these methods are computationally more expensive. Comparative studies of template versus local feature-based systems can be found in [15, 9, 67]. There are also various hybrid schemes that apply the KLT and/or template matching idea to face components and use correlation-based search to locate and identify facial feature points [9, 66]. The advantage of performing component-by-component matching is improved robustness against head orientation changes, but its disadvantage is the complexity of searching for and locating the face components.

The human audio/visual system, as a powerful recognition model, takes great advantage of context and auxiliary information. Inspired by this observation one can devise schemes that can consistently incorporate context and collateral information, when and if they become available, to enhance its final decisions. Incorporating information such as race, age and gender, obtained through independent analysis, improves recognition results [66]. Also, since face recognition involves a classification problem with large within-class variations, caused by dramatic image variations in different poses of the subject, one has to devise methods of reducing or compensating such variability, e.g.

1. For each subject store several templates, one for each major distinct facial expression and/or head orientation. Such systems are typically referred to as view-based systems.

2. Use deformable templates along with a 3-D model of a human face to synthesize virtual poses and apply the template matching algorithm to the synthesized representations [94].

3. Incorporate such variations in the process of feature extraction.

In our experiments, we take the third approach and keep the first method as an optional stage that can be employed depending on the complexity of the specific task. Our approach is to use holistic LDA-based feature extraction for human faces followed by evidential soft decision integration for multisource data analysis. This method is a projection-based scheme of low complexity that avoids any iterative search or computation. In this method both off-line feature extraction and on-line feature computation can be done at high speeds and recognition can be done almost in real time. Our experimental results show that very reliable recognition performance can be achieved with very low complexity and small numbers of features.

The organization of this chapter is as follows. In Section 6.2, we provide an objective study of multi-scale features of face images in terms of their discriminating power. In Section 6.3 we propose a holistic method of projection-based discriminant facial feature extraction through LDA of face images. We also make a comparative study of the features obtained using the proposed scheme and the ones employed in compression-based methods such as PCA/KLT. In Section 6.4 we address the task of classification/matching through multi-source data analysis and combining soft decisions from multiple imprecise information sources. Finally, based on the reliability of the basic decisions, we propose a task-dependent measure of similarity in the feature space, to be used at the identification stage. All the experiments in this chapter are based on the application of LDA to the original image, but the ideas can be extended to multiscale representations.

6.2 Linear Discriminant Analysis of Facial Images

As highly structured 2-D patterns, human face images can be analyzed in the spatial and/or the frequency domain. These patterns are comprised of components that are easily recognized at high levels but are loosely defined at low levels of our visual system [59, 23]. Each of the facial components/features has a different discriminatory power for identifying a person or the person's gender, race or age. There have been many studies of the significance of such features using subjective psychovisual experiments [5, 23].

Using objective measures, in this section we propose a computational scheme for evaluating the significance of different facial attributes in term of their discriminatory potential. The results of this analysis can be supported by subjective psychovisual findings. To analyze any representation V, where V can be the original image, its spatial segments, or transformed images, we provide the following framework.

First, we need a training set composed of a relatively large group of subjects with diverse facial characteristics. The appropriate selection of the training set directly determines the validity of the final results. The database should contain several examples of face images for each subject in the training set and at least one example in the test set. These examples should represent different frontal views of subjects with minor variations in view angle. They should also include different facial expressions, lighting and background conditions, and examples with and without glasses. It is assumed that all images are already normalized to $m \times n$ arrays and they only contain the face regions and not much of the subjects' bodies.

Second, for each image/subimage, starting with the two-dimensional $m \times n$ array of intensity values $I(x, y)$, we construct the lexicographic vector expansion

$\phi \in R^{m \times n}$. This vector corresponds to our initial representation of the face. Thus, the set of all faces in the feature space is treated as a high-dimensional vector space.

Third, by defining all instances of the same person's face as being in one class and the faces of different subjects as being in different classes, for all subjects in the training set, we establish a framework for performing a cluster separation analysis in the feature space. Also, having labeled all instances in the training set and having defined all the classes, we compute the within- and between-class scatter matrices, i.e. S_w and S_b respectively. Then we can use any of the class separability measures of Chapter 2. For example

$$J_V^2 = \text{Sep}(V) = \text{tr}(S^{(V)}) \tag{6.2.1}$$

$$J_V^3 = \text{tr}(S_b)/\text{tr}(S_w) \tag{6.2.2}$$

can be considered. In this test $J_V = J_V^2$ is our measure of the Discriminatory Power (DP) of a given representation V. As mentioned above, the representation may correspond to the data in its original form (e.g. a gray-scale image), or it can be based on a set of abstract features computed for a specific task.

For example, through this analysis we are able to compare the DP's of different spatial segments/components of a face. We can apply the analysis to segments of the face images such as the areas around the eyes, mouth, hair, chin, or combinations of them. Figure 6.1 shows a separation analysis for horizontal segments of the face images in the database. The results show that the DP's of all segments are comparable, and that the area between the nose and the mouth has more identification information than other parts. Figure 6.2 shows that the DP of the whole image is significantly larger than the DP's of its parts.

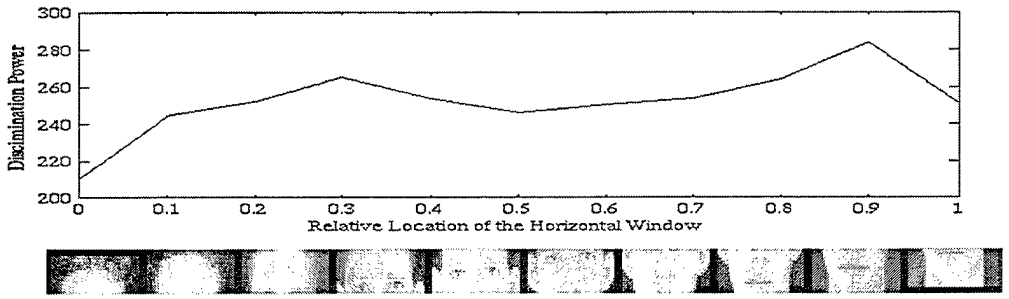


Figure 6.1: Variation of the discriminatory power of horizontal segments of the face defined by a window of fixed height sliding from top to bottom of the image.

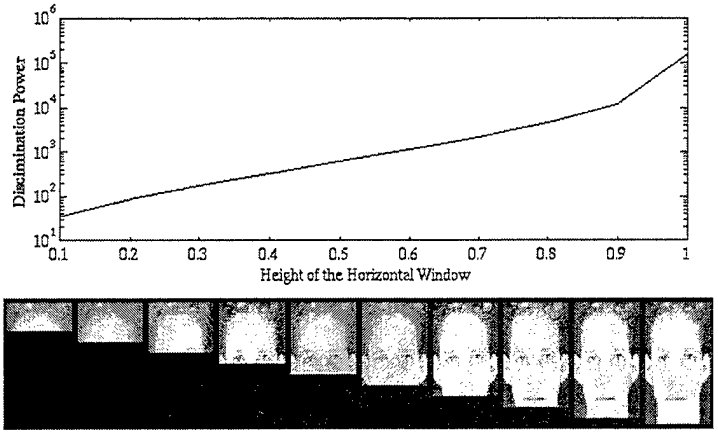


Figure 6.2: Variation of the discriminatory power of a horizontal segment of the face that grows in height from the top to the bottom of the image.

Using wavelet transforms [19, 56, 22] as multi-scale orthogonal representations of face images, we can also perform a comparative analysis of the DP's of subimages in the wavelet domain. Different components of a wavelet decomposition capture different visual aspects of a gray scale image. As Figure 6.3 shows, at each level of decomposition there are four orthogonal subimages corresponding to

- **LL:** The smoothed, low-frequency variations.
- **LH:** Sharp changes in the horizontal direction, i.e. vertical edges.

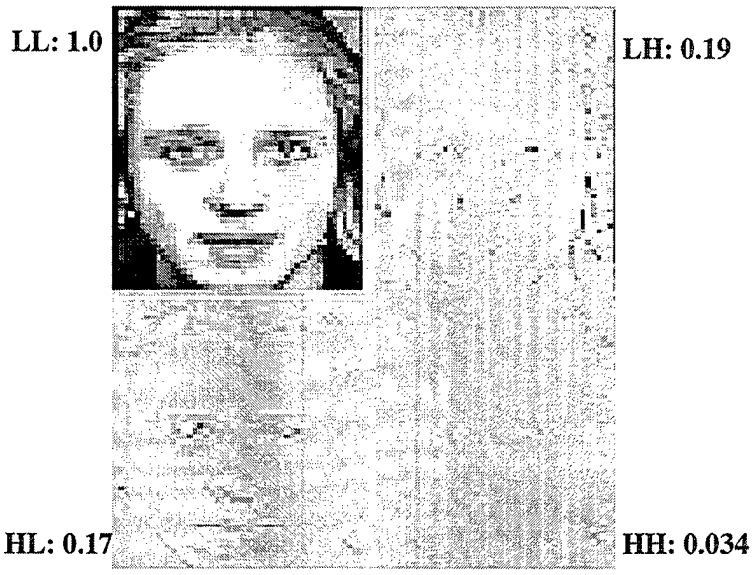


Figure 6.3: Different components of a wavelet transform, capturing sharp variations of the image intensity in different directions, have different discriminatory potentials. The numbers represent the relative discriminatory power.

- **HL:** Sharp changes in the vertical direction, i.e. horizontal edges.
- **HH:** Sharp changes in non-horizontal/non-vertical directions, i.e. other edges.

We applied the LDA to each subimage of the WT of the face and estimated the discriminatory power of each subband. Figure 6.3 compares the separations obtained using each of the subbands. Despite their equal sizes, different subimages carry different amounts of information for classification; the low-resolution component is the most informative. The horizontal edge patterns are almost as important as the vertical edge patterns, and their relative importance depends on the scale. Finally, the least important component in terms of face discrimination is the fourth subband, i.e. the slanted edge patterns. These results are consistent with our intuition and also with subjective psychovisual experiments.

One can also apply this idea to study the importance of facial components

for gender or race classification from images.

6.3 Discriminant Eigenfeatures for Face Recognition

In this section we propose a new algorithm for face recognition that makes use of a small, yet efficient, set of discriminant eigentemplates. The analysis is similar to the method suggested by Pentland et al.[66, 80], which is based on PCA and KLT. The fundamental difference is that in our system eigenvalue analysis is performed on the separation matrix rather than the covariance matrix.

Human face images as two-dimensional patterns have a lot in common and are spectrally very similar. Therefore, considering the face image as a whole, one expects to see important discriminant features that have low energies. These low-energy discriminant features may not be captured in a compression-based feature extraction scheme like PCA, or even in multi-layer neural networks, which rely on minimization of average Euclidean error. In fact, there is no guarantee that the error incurred by applying the compression scheme, despite its low energy, does not carry significant discriminatory information. Also, there is no reason to believe that for a given compression-based feature space, feature points corresponding to different poses of the same subject will be closer (in Euclidean distance) to each other than to those of other subjects. In fact it has been argued and experimentally shown that ignoring the first few eigenvectors, corresponding to the top principal components, can lead to a substantial increase in recognition accuracy [66, 63]. Therefore the secondary selection from the PCA vectors is based on their discriminatory power. But one could ask, why do we not start with a criterion based on discrimination rather than representation from the beginning, to make the whole process more consistent?

The KLT/PCA approach provides us with features that capture the main

directions along which face images differ the most, but it does not attempt to reduce the within-class scatter of the feature points. In other words, since no class membership information is utilized, examples of the same class or different classes are treated in the same way. LDA, however, uses the class membership information and allows us to find eigenfeatures and therefore representations in which the variations among different faces are emphasized, while the variations of the same face due to illumination conditions, facial expression, orientation etc. are de-emphasized.

According to this observation, and based on the results that follow, we believe that for classification purposes, LDA-based feature extraction seems to be an appropriate and logical alternative to PCA, KLT, or any other compression-based system which tries to find the most compact representation of face images. Concurrently, but independently of our studies, LDA has been used by Swet and Weng [78, 79] to discriminate human faces from other objects.

In order to capture the inherent symmetry of basic facial features and the fact that a face can be identified from its mirror image, we can use the mirror image of each example as a source of information [48]. Also, by adding noisy but identifiable versions of the given examples, we can expand our training data and improve the robustness of the feature extraction against small amount of noise in the input. Therefore, for each image in the database we include its mirror 4 of its noisy versions, as shown in Figure 6.4. We thus have

$$\Phi = \{\Phi_s : s = 1, 2, ..., N_S\} \tag{6.3.3}$$

$$\Phi_s = \{\phi_i^s, \widetilde{\phi_i^s}, (\phi_i^s + \nu) : i = 1, 2, ..., N_E, \ \nu = [N(0, \sigma^2)]^{m \times n}\} \tag{6.3.4}$$

where $\widetilde{\phi_i^s}$ and $\phi_i^s + \nu$ are mirror images and noisy versions of the i^{th} example of subject s in the data base Φ, respectively. Also N_S is the number of subjects and

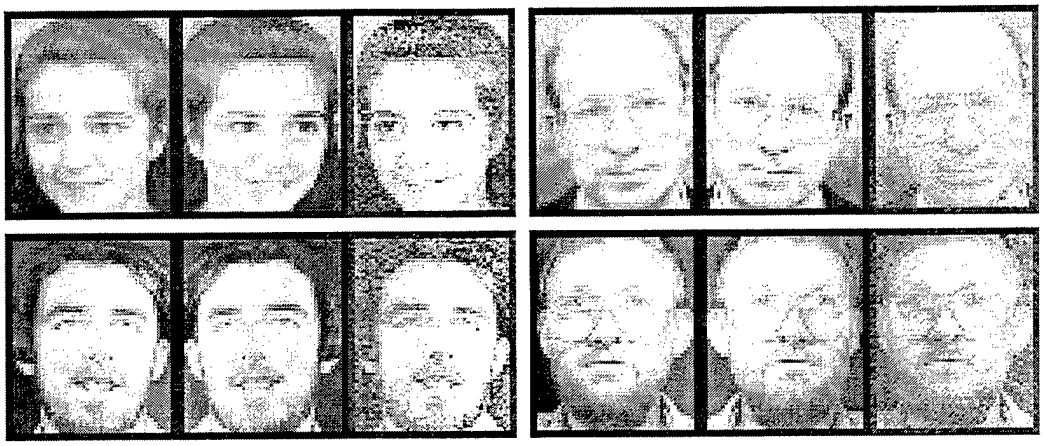


Figure 6.4: For each example in the database we add its mirror image and some noisy versions.

N_E is the number of examples per subject in the initial database. Following our earlier observations, and having determined the separation matrix, we perform eigenvalue analysis of the separation matrix $S^{(\Phi)}$ on the augmented database:

$$\text{eig}\{S^{(\Phi)}\} = \{(\lambda_i, u_i), i = 1, ..., N_S - 1,\ \lambda_i > \lambda_{i+1}\} \tag{6.3.5}$$

Now let $\Lambda^{(m)}$ and $U^{(m)}$ represent the set of m largest eigenvalues of $S^{(\Phi)}$ and their corresponding eigenvectors. As discussed in Chapter 2, $U^{(m)}$ minimizes the drop $|\text{Sep}(X) - \text{Sep}(U^T X)|$ in classification information incurred by the reduction in the feature space dimensionality, and no other $\mathbf{R^n}$ to $\mathbf{R^m}$ linear mapping can provide more separation than $U^{(n)}$ does.

Therefore, the optimal linear transformation from the initial representation space in $\mathbf{R^n}$ to a low-dimensional feature space in $\mathbf{R^m}$ based on our selected separation measure results from projecting the input vectors ϕ onto m eigenvectors corresponding to the m largest eigenvalues of the separation matrix $S^{(\Phi)}$. These optimal vectors/direction can be obtained from a sufficiently rich training set and can be updated if needed.

The columns of $U^{(m)}$ are the eigenvectors corresponding to the m largest

Figure 6.5: Some of the top eigenpictures based on PCA (top) and LDA(bottom).

eigenvalues; they represent the directions along which the projections of the face images within the database show the maximum class separation. As Figure 6.5 shows, unlike the KLT-based eigenfaces of the discriminant eigenvectors, these vectors do not typically have face-like patterns and are not directly related to our intuitive notions of isolated features of human faces such as eyes, hair, chin, etc.

Each face image in the database is represented, stored and tested in terms of its projections onto the selected set of discriminant vectors, i.e. the directions corresponding to the largest eigenvalues of $S^{(\Phi)}$:

$$\forall {\phi^s}_i \in \Phi_s, \forall u \in U^{(m)} : \psi_i^s(u) =< \phi_i^s, u > \tag{6.3.6}$$

$$\Psi^s = \{\Psi_i^s(u) : \forall u \in U^{(m)}, I = 1, ..., N_S\} \tag{6.3.7}$$

Although all images of each subject are considered in the process of training, only one of them needs to be saved, as a template for testing. If a view-based

approach is taken, one example has to be stored for each distinct view. Since only the projection coefficients need to be saved, for each subject we retain the example that is closest to the mean of the corresponding cluster in the feature space. Storing the projection coefficients instead of the actual images is highly desirable when large databases are used. Also, applying this holistic LDA to multi-scale representations of face images, one can obtain multiscale discriminant eigentemplates. For example one can apply LDA to each component of the WT of the face images and select the most discriminant eigentemplates obtained from various scales. This approach is more complex because it requires the WT computation of each test example, but in some applications it may be useful, for example when the DP of the original representation is not captured in the first few eigenvectors, or when the condition of $m < N_{\text{classes}} - 1$ becomes restrictive, e.g. in gender classification.

After extracting our projection-based discriminant features, we apply the multisource decision integration scheme of Chapter 3. In the process of decision integration we will use the DP of each decision axis resulting from a projection as a measure of its reliability. Then, for each presented face, we apply our simplified distance measure of equation (3.3.30) to the resulting feature vector to obtain a sorted list of the top candidates. Figure 6.6 illustrates distributions of projection coefficients along various axes for a four-class case.

6.4 Experiments and Results

In our experiments, in order to satisfy the requirements mentioned above, we used a mixture of two databases. We started with the database provided by Olivetti Research Ltd. [75]. This database contains 10 different images of each of 40 different subjects. All the images were taken against a homogeneous

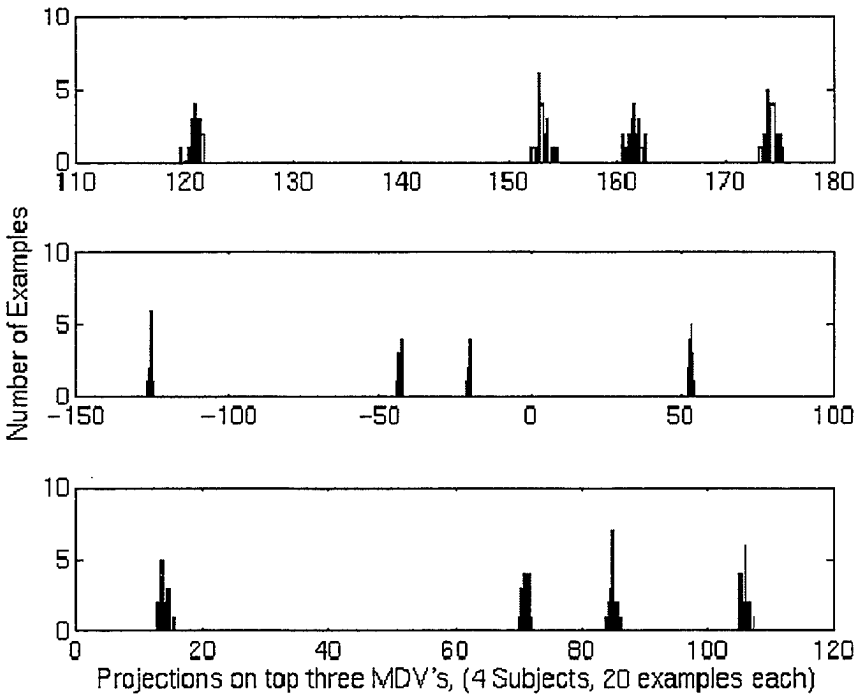


Figure 6.6: The distribution of projection coefficients along three discriminant vectors with different levels of discriminatory power for several poses from four different subjects.

background and some were taken at different times. The database includes frontal views of upright faces with slight changes in illumination, facial expression (open/closed eyes, smiling/non-smiling), facial details (glasses/no-glasses), and some side movements. Originally we chose this database because it contains many instances of frontal views for each subject. Then, to increase the size of the database, we added some hand-segmented face images from the Ferret database [31]. We also included mirror-image and noisy versions of each face example in order to expand the data set and improve the robustness of recognition performance to image distortions. The total numbers of images used in training and testing were about 1500 and 500 respectively. Each face was represented by a 50×60 pixel 8-bit gray-level image, which for our experiments was reduced to

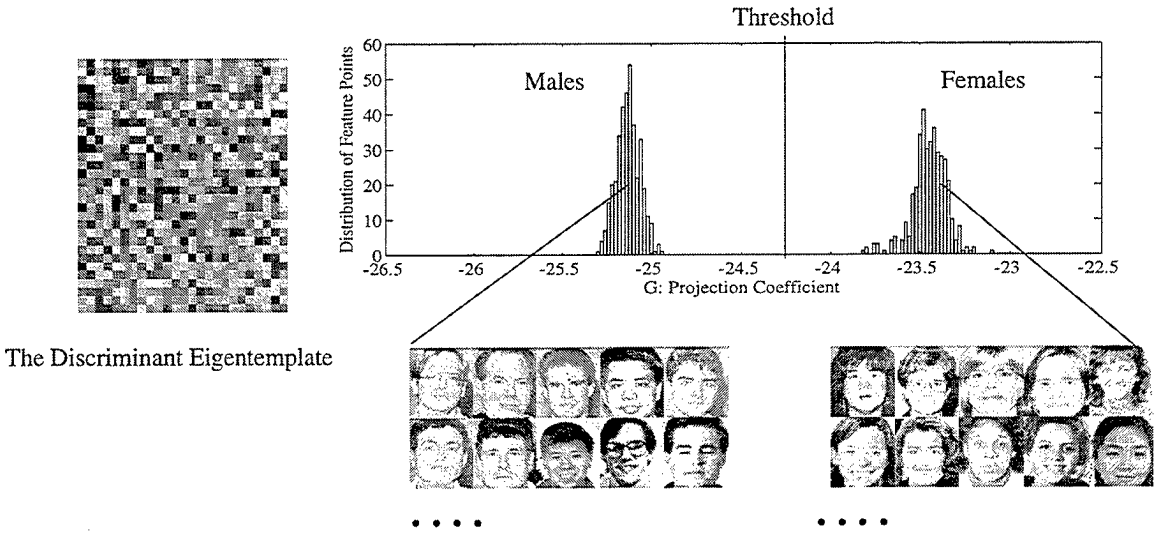


Figure 6.7: Distribution of feature points for male and female examples in the database.

25×30. The database was divided into two disjoint training and test sets. Using this composite database we performed several tests on gender classification and face recognition.

The first test was on gender classification using a subset of the database containing multiple frontal views of 20 males and 20 females of different races. The LDA was applied to the data and the most discriminant template was extracted. Figure 6.7 shows this eigentemplate and the distribution of projection coefficients for all images in the set. As Figure 6.7 shows, with only one feature very good separation can be achieved. Classification tests on a disjoint test set also gave 95% accuracy. As mentioned above, one can also apply LDA to wavelet transforms of face images and extract the most discriminant vectors of each transform component and combine multiscale classification results using 4 the proposed method of soft decision integration.

We then applied LDA to a database of 1500 faces, with 60 classes corre-

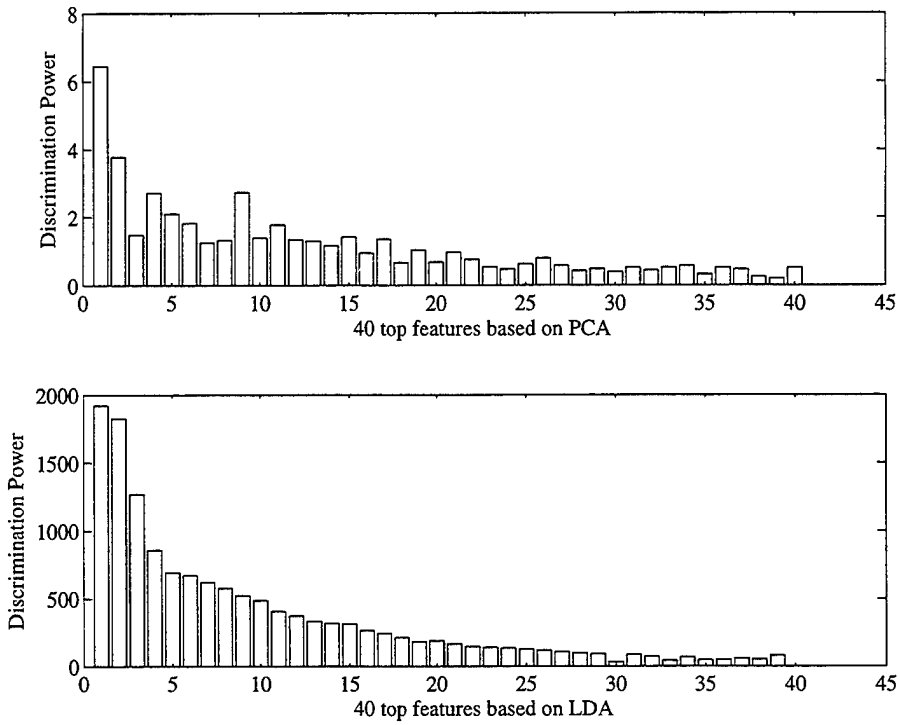


Figure 6.8: A comparison of DP's of the top 40 selected eigenvectors based on PCA and LDA.

sponding to 60 individuals. Figure 6.8 shows the discriminatory power of the top 40 eigenvectors chosen according to PCA and LDA. As Figure 6.8 shows, the classification information of the principal components does not decrease monotonically with their energy; in other words, there are many cases where a low-energy component has a higher discriminatory power than a high-energy component. The figure also shows that the top few discriminant vectors from LDA contain almost all the classification information embedded in the original image space.

Figure 6.9 shows the separation of clusters for ten poses of four different individuals using the two most discriminatory eigenvectors or eigenpictures. As Figure 6.9 indicates, the differences between classes (individuals) are emphasized

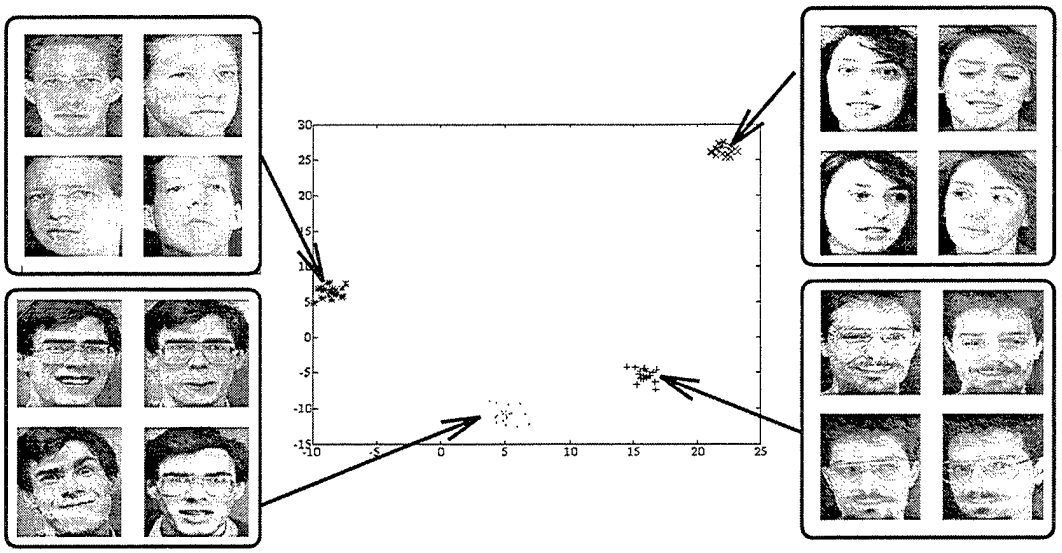


Figure 6.9: Separation of clusters in the selected 2-D feature space. Four clusters correspond to variations of the faces of four different subjects in the database.

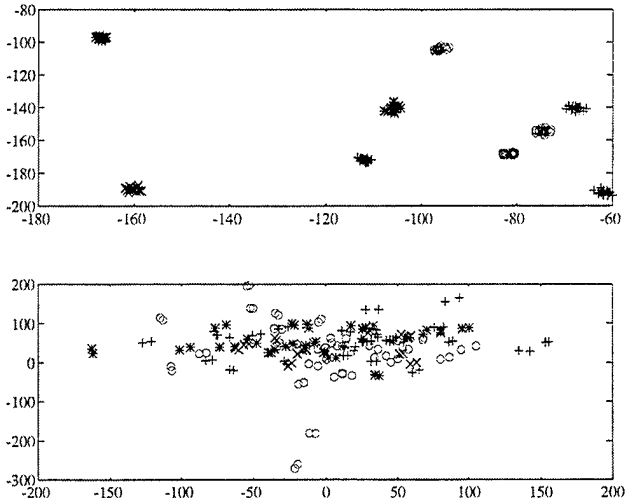


Figure 6.10: Cluster separation in the best 2-D feature space, based on LDA (top) and based on PCA (bottom).

while the variations of the same face in different poses are de-emphasized. The separation is achieved despite all the image variations resulting from the various poses of each subject. Figure 6.10 shows the distribution of clusters, for 200 images of 10 subjects, in the best two-dimensional discriminant feature space and

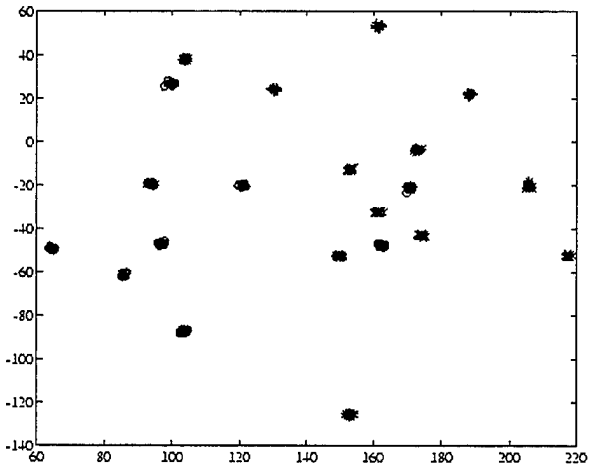


Figure 6.11: Cluster separation in the best 2-D discriminant feature space for 20 different subjects.

in the best two-dimensional PCA-based space. Figure 6.11 shows the clusters for 20 different subjects using the top two discriminant eigenvectors.

For each test face example, we first projected it onto the selected eigenvectors and found the distance from the corresponding point in the 4-D feature space to all of the previously saved instances. All distances were measured according to equation (3.3.30) and the best match was selected. For the given database excellent (99.2%) accuracy was achieved; see Table 6.1.

The simplicity of our system, the size of the database, and the robustness of the results to small variations of the pose or noise show that our suggested scheme is a good alternative approach to face recognition. It provides highly competitive results at much lower complexity using low-dimensional feature sizes.

Task	No. of Examples	No. of Features	Recogn. Rate (Training Set)	Recogn. Rate (Test Set)
Face Recognition	2000	4	100%	99.2%
Gender Classification	400	1	100%	95%

Table 6.1: Summary of recognition rates.

6.5 Conclusions

The application of LDA to study the discriminatory power of various facial features in the spatial and wavelet domains is presented. Also, an LDA-based feature extraction scheme for face recognition is proposed and tested.

A holistic projection-based approach to facial feature extraction is taken where eigentemplates are the most discriminant vectors derived from the LDA of face images in a rich enough database. The effectiveness of the proposed LDA-based features is compared with that of PCA-based eigenfaces. For classification a variation of evidential reasoning is used, in which each projection becomes a source of discriminating information with reliability proportional to its discrimination power. The weighted combination of similarity or dissimilarity scores suggested by all projection coefficients is the basis for the membership values.

Several results on face recognition and gender classification are presented, in which highly competitive recognition accuracies are achieved with very small numbers of features. The feature extraction can be applied to the WP representations of the images to provide a multiscale discriminant framework. In such

cases the system becomes more complex at the expense of improving separability and performance. The proposed feature extraction combined with soft classification seems to be a promising alternative to other face recognition systems.

Chapter 7

Conclusions

The combination of the theories of wavelet-based multiresolution analysis and discriminant analysis in statistical pattern recognition provides us with powerful and flexible frameworks for extracting discriminant features for pattern recognition/verification and segmentation systems.

Multi-scale features can be built in the process of selecting or linearly combining waveforms from a library of local basis functions with the objective of obtaining largest class separability in the feature space. The original library or dictionary of waveforms may be a combination of different classes of orthogonal wavelets, Gabor functions and local trigonometric functions. For the case of tree-structured orthogonal local bases such as wavelet packets and local trigonometric functions there are fast search algorithms to find the most discriminatory basis, whereas for redundant dictionaries only suboptimal greedy search algorithms are available. The resulting selection of local waveforms may not be a complete basis for the signal space, as is required in function approximation problems.

In many classification/recognition based systems decisions are based on multiple features or sources of information, where different features or sources have different levels of reliability or impreciseness. The results of classification based on incomplete, noisy or mixed data or sources of different reliabilities can be best represented using soft decision vectors. Following the principle of least commitment one needs to keep all intermediate results as soft decisions up to the last step when a crisp decision may be needed. Soft decision boundaries and thereby

fuzzy partitions of the feature space are obtained using the nonlinear mapping of multilayer neural networks or distance-based similarity/dissimilarity scores.

In some applications it is possible to take advantage of multiple observations in a spatial/temporal neighborhood or context area in the final decision making. We investigated soft decision integration using a consensus rule which is a variation of a linear or logarithmic opinion pool with discrimination-based weighting factors. Also, we enhanced the result by combining decisions in a context area based on a relevance pattern. Decision integration can be implemented in a probabilistic or evidential frame of reasoning.

We explored these ideas by testing them on a variety of applications, including

- Recognition of Real Aperture Radar returns
- Classification and segmentation of texture images
- Layout-independent segmentation of complex document pages
- Automatic face recognition

Despite the many differences among these applications, we consistently obtained promising results using fundamentally similar ideas. In some applications, such as face recognition and document page segmentation, we presented a completely new approach to the problem, while in other cases (e.g. texture and radar classification) we obtained competitive results using systems of lower complexity based on our alternative discriminant feature extraction and classification methodology.

Future Work: In this dissertation we have explored some aspects of context-dependent pattern recognition systems using a toolbox containing multi-scale

signal processing algorithms and multi-source decision integration methodologies. We hope that the results described in this dissertation will serve as a basis for further investigations. This work seems to be extendible in different directions from analytical and implementation points of view.

In the context of multi-scale discriminant basis selection further studies can be done on an appropriate choice of a composite and redundant dictionary and on developing/testing various greedy-type separation pursuit algorithms. Also, separation-based local basis selection may have applications in noise suppression and signal enhancement systems, where noise is defined as an additional class and through projection-based methods, the system tries to find a multi-scale representation in which there is maximum separation/difference between the signal components and the noise.

Also, more extensive research on discrimination and relevance-based decision integration using more sophisticated and efficient consensus rules is needed. This may involve various combinations of probabilistic, evidential, fuzzy and neural-based approaches in decision making.

Future work can also address new applications. One can test our proposed scheme on many other signal and image processing tasks, such as recognition of speech and acoustic signals or analysis of aerial or medical images. Also, in some of the applications that we have explored there is room for extensions and enhancements. For example, our results on page segmentation can be linked to higher levels of knowledge-based post-processing for document understanding and/or compression. In face recognition, our work can be extended to larger and richer databases, covering wider variations of race, age, gender, etc. Such a complete database can also be used for more detailed analysis of facial features that can be compared and linked with psychophysical findings.

Bibliography

[1] T. Aibara, K. Ohue, and Y. Matsuoka, "Human Face Recognition by P-type Fourier Descriptors," in *SPIE Proceedings, Vol. 1606: Visual Communications and Image Processing*, pp. 198-203, 1991.

[2] M. Antoniony, M. Barlaud, P. Mathieu and I.Daubechies, "Image Coding Using Wavelet Transform", in *IEEE Trans. on Image Processing*, Vol. 1, pp. 205-220, 1992.

[3] A. Antonacopoulos and R.T. Ritchings,"Flexible Page Segmentation Using the Background", in *Proceedings, International Conference on Pattern Recognition*, pp. 339-344, 1994.

[4] H.S. Baird, "The Skew Angle of Printed Documents", in *SPSE 40th Annual Conference and Symposium on Hybrid Imaging Systems*, pp. 21-24, 1987.

[5] R. Baron, "Mechanisms of Human Facial Recognition," *in International Journal of Man-Machine Studies*, Vol. 15, pp. 137-178, 1981.

[6] J. C. Bezdek and S.K. Pal (editors), *Fuzzy Models and Pattern Recognition*, IEEE Press, NewYork, 1992.

[7] J.A. Benediktsson and P.H. Swain,"Consensus Theoretic Classification Methods", in *IEEE Trans. on Systems, Man, and Cybernetics*, Vol.22, pp. 688-704, 1992.

[8] P.L. Bogler, "Shafer-Dempster Reasoning with Applications to Multisensor Target Identification Systems," in *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 17, pp. 968-977, 1987.

[9] R. Brunelli and T. Poggio, "HyperBF Networks for Gender Classification," in *Proceedings, DARPA Image Understanding Workshop*, pp. 311-314, 1992.

[10] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates," *in IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, pp. 1042-1052, 1993.

[11] S. Carey, "A Case Study: Face Recognition," in *Explorations in the Biological Language* (E. Walker, ed.), pp. 175-201, Bradford Books, New York, 1987.

[12] S. Carey, R. Diamond, and B. Woods, "The Development of Face Recognition — A Maturational Component?," in *Developmental Psychology*, Vol. 16, pp. 257-269, 1980.

[13] T. Chang and C.C.J. Kuo, "Texture Analysis and Classification with Tree Structured Wavelet Transform", in *IEEE Trans. on Image Processing*, Vol. 2, pp. 429-440, 1993.

[14] P. Chauvet, J. Lopez-Krahe, E. Taflin and H. Maitre, "System for Intelligent Office Document Analysis, Recognition and Description", in *Signal Processing*, Vol. 32, pp. 161-190, 1993.

[15] R. Chellappa, C.L. Wilson and S. Sirohey, "Human and Machine Recognition of Faces, a Survey," in *Proceedings of the IEEE*, Vol. 83, pp. 705-740, 1995.

[16] R. Chellappa, "Two-Dimensional Discrete Gaussian Markov Random Field Models for Image Processing", in *Progress in Pattern Recognition* (L.N. Kanal and A. Rosenfeld, eds.), Vol. 2, pp. 79-112, North Holland, Amsterdam, 1985.

[17] P.C. Chen and T. Pavlidis, "Segmentation by Texture using Correlation", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 5, pp. 64-69, 1983.

[18] Y. Cheng, K. Liu, J. Yang, and H. Wang, "A Robust Algebraic Method for Human Face Recognition," in *Proceedings, International Conference on Pattern Recognition*, pp. 221-224, 1992.

[19] R.R. Coifman and M.V. Wickerhauser, "Entropy Based Algorithms for Best Basis Selection", in *IEEE Trans. on Information Theory*, Vol. 38, pp. 713-718, 1992.

[20] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 1992.

[21] I. Craw, D. Tock, and A. Bennett, "Finding Face Features," *in Proceedings, European Conference on Computer Vision*, pp. 92-96, 1992.

[22] I. Daubechies, "Orthonormal Basis of Compactly Supported Wavelets", in *Comm. Pure Appl. Math.*, Vol. 41, pp. 909-996, 1988.

[23] G. Davies, H. Ellis, and E. J. Shepherd, *Perceiving and Remembering Faces*, Academic Press, New York, 1981.

[24] P.A. Devijver and J. Kittler, *Pattern Recognition, a Statistical Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[25] M. Desai and D. Shazeer,"Acoustic Transient Analysis Using Wavelet Decomposition", in *Proceedigns, IEEE Conference on Neural Networks for Ocean Engineering*, 1991.

[26] A.J. Devaney and B. Hiscomez, "Wavelet Signal Processing for Radar Target Identification: A Scale Sequential Approach," in *SPIE Proceedings, Vol. 2242: Wavelet Applications*, pp.389-399, 1994.

[27] K.B. Eom and R. Chellappa,"Classification of Millimeter Wave Radar Signatures by Hierarchical Modeling", Technical Report CAR-TR-769, Center for Automation Research, University of Maryland, April 1995.

[28] K. Etemad and R. Chellappa, "Separability Based Tree Structured Basis Selection for Texture Classification", in *Proceedings, First International Conference on Image Processing*, pp. 441-445, 1994.

[29] K. Etemad, R. Chellappa and D. Doermann, "Page Segmentation Using Wavelet Packets and Decision Integration", in *Proceedings, International Conference on Pattern Recognition*, pp. 345-349, 1994.

[30] K. Etemad, R. Chellappa and D. Doermann, "Document Page Decomposition by Integration of Distributed Soft Decisions", in *Proceedings, IEEE International Conference on Neural Networks*, pp. 4022-4027 , 1994.

[31] A.T. DePersia and P.J. Phillips, "The Ferret program, overview and accomplishments", Technical Report, George Mason University, Fairfax, VA, 1995.

[32] L.A. Fletcher and R. Kasturi,"A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 10, pp. 910-918, 1988.

[33] S. Furui, "Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques", in *Speech Communication*, Vol.5, 1986.

[34] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, New York, 1989.

[35] B. A. Golomb and T. J. Sejnowski, "SEXNET: A Neural Network Identifies Sex From Human Faces," in *Advances in Neural Information Processing Systems 3* (D. S. Touretzky and R. Lipmann, eds.), pp. 572-577, Morgan Kaufmann, San Mateo, CA, 1991.

[36] G. Gordon, "Face Recognition Based on Depth Maps and Surface Curvature," in *SPIE Proceedings, Vol. 1570: Geometric Methods in Computer Vision*, pp. 234-247, 1991.

[37] G. G. Gordon and L. Vincent, "Application of Morphology to Feature Extraction for Face Recognition," in *SPIE Proceedings, Vol. 1658: Nonlinear Image Processing*, pp. 151-164, 1992.

[38] J. Ghosh and A.C. Bovik, "Neural Networks for Textured Image Processing", in *Progress in Artificial Neural Networks and Statistical Pattern Recognition* (I.K. Sethi and A.K. Jain, eds.), pp.133-154, North-Holland, Amsterdam, 1991.

[39] A. Goshtasby, "Description and Discrimination of Planar Shapes Using Shape Matrices," in *IEEE Trans. on Patterns Analysis and Machine Intelligence*, Vol. 7, pp. 738-743, 1985.

[40] V. Govindaraju, S. N. Srihari, and D. B. Sher, "A Computational Model for Face Location," in *Proceedings, International Conference on Computer Vision*, pp. 718-721, 1990.

[41] L. Harmon, M. Khan, R. Lasch, and P. Ramig, "Machine Identification of Human Faces," in *Pattern Recognition*, Vol. 13, pp. 97-110, 1981.

[42] X.D. Huang, Y. Akiri and M.A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, U.K., 1990.

[43] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[44] A.K. Jain and F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters", in *Pattern Recognition*, Vol. 24, pp. 1167-1186, 1991.

[45] A.K. Jain and S. Bhattacharjee,"Text Segmentation Using Gabor Filters for Automatic Document Processing", in *Machine Vision and Applications*, Vol. 5, pp. 169-184, 1992.

[46] L. Kanal, ed., *Uncertainty in Artificial Intelligence*, Vol. 1, North Holland, Amsterdam, 1986.

[47] G. J. Kaufman, Jr. and K. J. Breeding, "The Automatic Recognition of Human Faces from Profile Silhouettes," in *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 6, pp. 113-121, 1976.

[48] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 12, pp. 103-108, 1990.

[49] B. Kosko, *Neural Networks and Fuzzy Systems*, Prentice Hall, Englewood Cliffs, NJ, 1992.

[50] M. Krishnamoorthy, G. Nagy, S. Seth and M. Viswanathan, "Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, pp. 1464-1480, 1990.

[51] R.E. Learned, W.C. Karl and A.S. Willsky, "Wavelet Packet Based Transient Signal Classification", in *Proceedings, IEEE Conference on Time Scale and Time Frequency Analysis*, pp. 109-112, 1992.

[52] T. Lee, J.A. Richards, and P.H. Swain, "Probabilistic and Evidential Approaches for Multisource Data Analysis", in *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 25, pp. 283-293, 1987.

[53] H. Li, P. Roivainen, and R. Forchheimer, "3-D Motion Estimation in Model-Based Facial Image Coding," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, pp. 545-555, 1993.

[54] R.P. Lippmann, "An Introduction to Computing with Neural Nets", in *IEEE Acoustics, Speech and Signal Processing Magazine*, pp. 4-22, April 1987.

[55] Y.C. Lin, T. Chang and C.C.J. Kuo, "Texture Segmentation Using Wavelet Packets", in SPIE *Proceedings, Vol. 2034: Mathematical Imaging: Wavelet Applications in Signal and Image Processing*, pp.277-287, 1993.

[56] S.G. Mallat, "A Theory for Multi-resolution Signal Decomposition, the Wavelet Representation", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 11, pp. 674-693, 1989.

[57] S. Mallat and Z. Zhang, "Matching Pursuit with Time Frequency Dictionaries", in *IEEE Trans. on Signal Processing*, Vol. 41, pp. 3397-3415, 1993.

[58] B. S. Manjunath, R. Chellappa, and C. v. d. Malsburg, "A Feature Based Approach to Face Recognition," in *Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 373-378, 1992.

[59] D. Marr, *Vision*, W. H. Freeman, San Francisco, CA, 1982.

[60] O. Nakamura, S. Mathur, and T. Minami, "Identification of Human Faces Based on Isodensity Maps," in *Pattern Recognition*, Vol. 24, pp. 263-272, 1991.

[61] Y. Nakano, Y. Shima, H. Fujfisawa, J. Higashino and M. Fujinawa, "An Algorithm for Skew Normalization of Document Images", in *Proceedings, International Conference on Pattern Recognition*, pp. 8-13, 1990.

[62] C.L. Nikias, "Higher Order Spectral Analysis", in *Advances in Spectrum Analysis and Array Processing*, Vol. I (S. Haykin, ed.), Prentice-Hall, Englewood Cliffs, NJ., pp. 326-365, 1991.

[63] A. O'Toole, H. Abdi, K. Deffenbacher, and D. Valentin, "Low-Dimensional Representation of Faces in Higher Dimensions of the Face Space", in *Journal of the Optical Society of America* , Vol. A10, pp. 405-410, 1993.

[64] T. Pavlidis,"Page Segmentation by White Streams", in *Proceedings, International Conference on Document Analysis and Recognition*, pp. 945-953, 1991.

[65] T. Pavlidis and J. Zhou, "Page Segmentation and Classification", in *CVGIP: Graphical Models and Image Processing*, Vol. 54, pp. 484-496, 1992.

[66] A. Pentland, B. Moghaddam, T. Starner, and M. Turk, "View-based and Modular Eigenspaces for Face Recognition," in *Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 84-91, 1994.

[67] A. Rahardja, A. Sowmya, and W. Wilson, "A Neural Network Approach to Component Versus Holistic Recognition of Facial Expressions in Images," in *SPIE Proceedings, Vol. 1607: Intelligent Robots and Computer Vision X: Algorithms and Techniques*, pp. 62-70, 1991.

[68] K. Ramchandran and M. Vetterli, "Best Wavelet Packet Bases in a Rate-Distortion Sense", in *IEEE Trans. on Image Processing*, Vol. 2, pp. 160-175, 1993.

[69] D. Reisfeld and Y. Yeshurun, "Robust Detection of Facial Features by Generalized Symmetry," in *Proceedings, International Conference on Pattern Recognition*, pp. 117-120, 1992.

[70] O. Riol and M. Vetterli, "Wavelets and Signal Processing", in *IEEE Signal Processing Magazine*, pp. 14-38, 1991.

[71] D. Rumelhart and J. McClelland, *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, MA, pp. 322-328, 1988.

[72] N. Saito and R.R. Coifman, "Local Discriminant Basis", in *SPIE Proceedings, Vol. 2303: Mathematical Imaging: Wavelet Applications in Signal and Image Processing*, pp. 2-14, 1994.

[73] T. Saitoh and T. Pavlidis, "Page Segmentation Without Rectangle Assumption", in *Proceedings, International Conference on Pattern Recognition*, pp. 277-280, 1992.

[74] A. Samal and P. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey," in *Pattern Recognition*, Vol. 25, pp. 65-77, 1992.

[75] F. Samaria and A. Harter, "Parameterisation of a Stochastic Model for Human Face Identification", in *Proceedings, IEEE Workshop on Applications of Computer Vision*, 1994.

[76] M. Seibert and A. Waxman, "Recognizing Faces from their Parts," in *SPIE Proceedings, Vol. 1611: Sensor Fusion IV: Control Paradigms and Data Structures*, pp. 129-140, 1991.

[77] L. Sirovich and M. Kirby, "Low-Dimensional Procedure for the Characterization of Human Face," in *Journal of the Optical Society of America*, Vol. A4, pp. 519-524, 1987.

[78] D.L. Swets and J.J. Weng, "SHOSLIF-O: SHOSLIF for Object Recognition (Phase I)", Technical Report CPS 94-64, Michigan State University, East Lansing, MI, 1994.

[79] D.L. Swets, B. Punch and J.J. Weng, "Genetic Algorithms for Object Recognition in a Complex Scene", in *Proceedings, International Conference on Image Processing*, Vol. 2, pp. 595-598, 1995.

[80] M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces," in *Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586-591, 1991.

[81] Y.Y. Tang, C.Y. Suen, C.D. Yan and M. Cheriet,"Document Analysis and Understanding: A brief survey", in *Proceedings, International Conference on Document Analysis and Recognition*, pp. 17-31, 1991.

[82] M. Unser and M. Eden, "Multiresolution Feature Extraction and Selection for Texture Segmentation", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 11, pp. 717-728, 1989.

[83] R.L. De Valois and K.K. De Valois, *Spatial Vision*, Oxford University Press, New York, 1990.

[84] M. Viswanathan and G. Nagy, "Characteristics of Digitized Images of Technical Articles", in *SPIE Proceedings, Vol. 1661: Machine Vision Applications in Character Recognition and Industrial Inspection*, pp. 6-17, 1992.

[85] F.M. Wahl, K.Y. Wong and R.G. Casey,"Block Segmentation and Text Extraction in Mixed Text/Image Documents", in *Computer Vision, Graphics, and Image Processing*, Vol. 20, pp. 375-390, 1982.

[86] F. Wang, "Fuzzy Supervised Classification of Remote Sensing Images", in *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 28, pp. 194-201, 1990.

[87] D. Wang and S.N. Srihari, "Classification of Newspaper Image Blocks using Texture Analysis", in *Computer Vision, Graphics, and Image Processing*, Vol. 47, pp. 327-352, 1989.

[88] R.L. Watrous, "GRADSIM: A Connectionist Network Simulator using Gradient Optimization Techniques", Technical Report MS-CIS-88-16, University of Pennsylvania, 1988.

[89] J.W. Woods and S.D. O'Neil, "Subband Coding of Images", in *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 5, pp. 1278-1288, 1986.

[90] C. Wu and J. Huang, "Human Face Profile Recognition by Computer," in *Pattern Recognition*, Vol. 23, pp. 255-259, 1990.

[91] Y. Yacoob and L. S. Davis, "Computing Spatio-Temporal Representations of Human Faces," in *Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 70-75, 1994.

[92] G. Yang and T. S. Huang, "Human Face Detection in a Scene," in *Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 453-458, 1993.

[93] X. Yang, K. Wang and S.A. Shamma, "Auditory Representations of Acoustic Signals", in *IEEE Trans. on Information Theory*, Vol. 38, pp. 824-839, 1992.

[94] A. Yuille, D. Cohen, and P. Hallinan, "Feature Extraction From Faces Using Deformable Templates," in *Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 104-109, 1989.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1996	Technical Report

4. TITLE AND SUBTITLE
Multi-scale Discriminant Analysis and Recognition of Signals and Images

5. FUNDING NUMBERS
N00014-95-1-0521

6. AUTHOR(S)
Kamran Etemad

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Center for Automation Research
University of Maryland
College Park, MD 20742-3275

8. PERFORMING ORGANIZATION REPORT NUMBER
CAR-TR-821
CS-TR-3629

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 North Quincy Street, Arlington, VA 22217-5660
Advanced Research Projects Agency
3701 North Fairfax Drive, Arlington, VA 22203-1714

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release.
Distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This dissertation explores multiscale discriminant basis selection, as well as the improvement of classification reliability through context-dependent integration of soft decisions. These methods are applied to texture and radar signature classification, document image segmentation, and human face recognition.

14. SUBJECT TERMS
discriminant basis selection, integration of soft decisions, texture classification, radar signature classification, document image segmentation, human face recognition

15. NUMBER OF PAGES
131

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	PR	- Project
G	- Grant	TA	- Task
PE	- Program Element	WU	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- DOE - See authorities.
- NASA - See Handbook NHB 2200.2.
- NTIS - Leave blank.

Block 12b. Distribution Code.

- DOD - Leave blank.
- DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- NASA - Leave blank.
- NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.