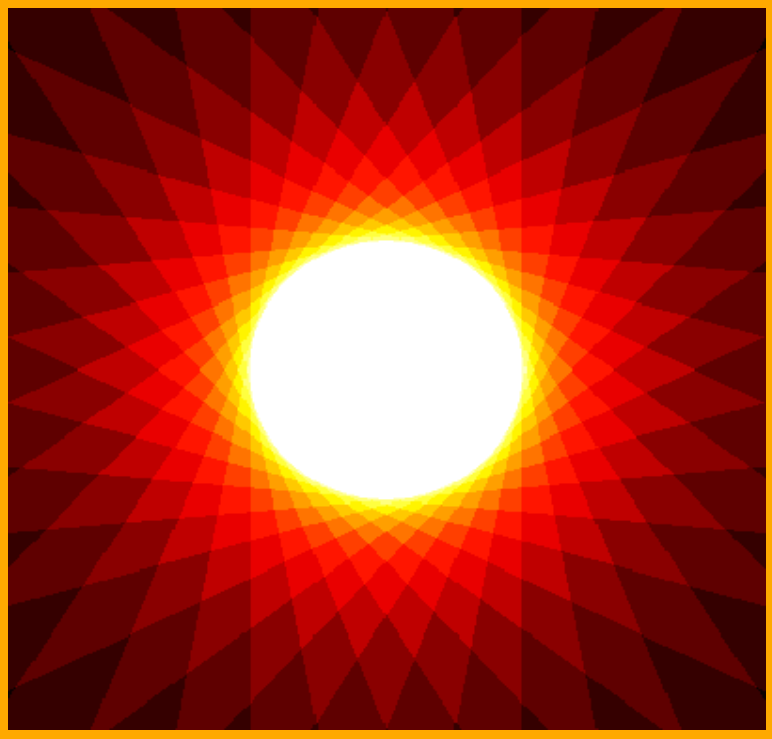
Adaptive Multiresolution Frame Classification of Biomedical Images



Amina Chebira

bimagicLab
Center for Bioimage Informatics
Dept. of Biomedical Engineering
Carnegie Mellon University

Adaptive Multiresolution Frame Classification of Biomedical Images

Amina Chebira

Advisor: Prof. Jelena Kovačević

Department of Biomedical Engineering
Carnegie Mellon University, Pittsburgh, PA 15213

Thesis Manuscript

Submitted in partial fulfillment of the requirements towards the Ph.D. degree awarded by the
Department of Biomedical Engineering, Carnegie Inst. of Tech., Carnegie Mellon University.

Thesis Committee Members

Jelena Kovačević (Advisor)
Carnegie Mellon University

José F. Moura
Carnegie Mellon University

Gustavo K. Rohde
Carnegie Mellon University

Martin Vetterli
École Polytechnique Fédérale de Lausanne

Stefan Zappe
Carnegie Mellon University

To my beloved parents,
for always giving me the freedom to choose and be.

À mes chers parents,
pour m'avoir toujours donné la liberté de choisir et d'être.

Contents

List of Figures

List of Tables

Abstract

This thesis presents a mathematical framework and an algorithm for the classification of biomedical image data sets based on adaptive and redundant multiresolution representations—frames. We illustrate the results on several different biomedical applications.

Classification is a ubiquitous problem in image processing; many biomedical tasks are in essence classification problems. Examples of such problems include determining a specific protein from its subcellular location pattern, determining the developmental stage of Drosophila embryos, recognizing tissue types in histological images of stem-cell teratomas, as well as determining otitis media stages. Though cumbersome, some of the above tasks, and many similar ones, are performed simply by visual inspection. As our eyes are not trained to extract statistical measures or time-frequency behavior of the signal across scales, these characteristics often pass unnoticed, resulting in poorer performance. We hypothesize that classifying adaptively in multiresolution subspaces will increase classification accuracy. We develop a new classifier, based on adaptive multiresolution ideas, by adding a multiresolution block in front of a generic classifier. The system is completed with a weighting block at the end, which plays the role of an arbiter; it decides how to combine the "subspace" decisions into a common one. The classifier achieves remarkable results, with most of the applications having classification accuracy in the mid- to high 90s.

In all of the applications, redundant multiresolution transforms performed the best. This led us to ask the following question: Why do frames perform better than bases? This question is nontrivial in scope, to begin to answer it we propose a classification scheme which uses finite frames and introduce a measure-theoretic framework for the analysis of classification errors. We then use this framework to examine those classes of signals for which a bases-based classification scheme is sufficient, and those for which a frame-based scheme is superior. We also show the proposed classification scheme performs well in the presence of noise.

Finally, as there are very few frame families available in the literature, we embarked on developing our own. To that end, we introduce a new class of frames we call lapped tight frame transforms, obtained by seeding from higher-dimensional orthonormal bases. We prove several properties of such frames, such as tightness, equal norm and maximal robustness.

Acknowledgments

My deepest gratitude goes to my amazing advisor Jelena Kovačević. Her enthusiasm, intuitions and dedication to work have guided my steps from the start of this adventure till the very end. She has been a constant source of inspiration. She made work fun and her magical touch always made the difference! Our meetings were always fun and a source of great motivation. Somehow, no matter what mood I was in before entering her office, I would always leave with a smile and feeling much better about the work, life and myself. I would like to thank her for caring so much, for the beautiful atmosphere of our group, for all her encouragements and for believing in me. Above all, she became more than an advisor and honored me with her friendship. For that, I thank her from the bottom of my heart.

Next, I would like to thank my thesis committee members José Moura, Gustavo Rohde, Martin Vetterli and Stefan Zappe, for accepting to be part of this work, for their support, invaluable advice and guidance, for their encouragements and for having my best interest at heart.

I would like to express my gratitude to all the collaborators with whom I have had the chance to work and from whom I learned so much. In particular, I would like to thank Carlos Castro, Justin Crowley, Matthew Fickus, Alejandro Hoberman, John Minden, Robert Murphy, John Ozolek and Markus Püschel. They shared their knowledge, wisdom and for some of them, their data. Working with them was exciting, a real pleasure and a great learning experience. In particular, I am indebted to Matthew Fickus for the work accomplished on frames classification, for welcoming me in his home, driving me around on icy roads in Ohio and for proofreading this manuscript.

I am deeply grateful to Martin Vetterli for guiding my first steps in the research world, for welcoming me in his "family", for being such a visionary and for his famous "gut feelings", for truly caring and always listening attentively, and of course for introducing me to Jelena! I would like to thank Pier Luigi Dragotti for his tremendous patience, his help and continuous support and for teaching me so much.

During my thesis, I have had the opportunity to supervise many talented students. Their contributions to this work are significant. In particular, I would like to thank the following people for their contributions to the different projects:

- Protein subcellular location project: Yann Barbotin and Alexia Mintos
- Recognition of developmental stages in *Drosophila*: Ryan Kellogg

- Identification of Histological stem cell teratomas: Mukta Gore and Garrett Jenkinson
- Otitis media Diagnosis: Irina Khaimovich and Shauna Ormon
- Fingerprint recognition: Luis Coelho, Stephen Lin, Garrett Jenkinson, Jeremiah MacSleyne, Christopher Hoffman and Philip Cuadra
- Lapped tight frame transforms: Christina Milo

My fellow Ph.D students have been instrumental to the accomplishments of this work. In particular, Ramu Baghavatula participated in the stem cell project; Charles Jackson came up with the idea of the new Haralick feature set; Thomas Merryman for the seminal work done with Gowri and I on the classification system, for coming up with many ideas including using neural networks and implementing large amounts of the code for the first two versions of the system; Aliaksei Sandryhaila for all the experiments performed with various MR bases for the fingerprint recognition project, for coming up with the proofs of maximal robustness for the lapped tight frame families and the —simple— proofs for equal-norm; Finally, Gowri Srinivasa for all her work on the first two versions of the classification system, for always being there and ready to help with a smile. I loved all the brainstorming sessions we had throughout the years! I would also like to thank the colleagues from CBI for their help and support. Elvira Garcia Osuna helped with the initial work on the protein subcellular location project by providing necessary information and the data set itself. Thomas Gulish, Paul Lucci and Daniel Willard were extremely helpful in various areas.

Many key people contributed to this work in different ways, if not directly, they did so by making me laugh, making sure I was alive and well and keeping me motivated! To all of them, thanks for their continued friendship throughout the years despite the geographical distance between us. I list some of them here according to where I met them.

1. Pittsburgh: To my gang, for all the fun and the tea times. This team consists of Sidi Bencherif, for his generosity and "craziness"; Sanna Gaspard, my partner in crime, for everything we have been through together in Pittsburgh; Estelle Glory for her tremendous kindness; Yannick Heintz for saving my life so many times and for being the happy guy around; and Warren Ruder for all the inspiring and fun discussions and the practical jokes. To Lydie Ngo Um, cocotte, for being the greatest roommate and for being such a fighter. To Ayorkor Mills-Tettey for being such a wonderful person. To my twin sister Gowri Srinivasa, this strong woman taught me a lot; going through everything with her made it that much easier! To Kevin Schnelli and Lionel Coulot for all the great trips we had.
2. Lausanne: Emanuel Corthay and Gabriel Walt for welcoming me so warmly on my first day at EPFL and then in their group, and making me feel at home; Sylviane Dal Mas for her friendship, Olivier Roy for all the laughter and the great memories, Ruben Merz and Luciano Sbaiz for their continued friendship.
3. Paris: To Kahina Abdeli for being this amazing little woman, and Souhila Abdeli and their family. To Sihem Amer-Yahia for her support and all the new years eve parties in Paris and New York! To Céline Bichard for her continued friendship, Christelle Boudonnat-Blavette for all of our shared meals at Coubertin and all

of our deep discussions. To Claudine Boudikian and her husband for taking care of me when I was sick and hosting me in their home as a member of their family.
4. Algiers: To Soraya Mokdad, Isma Hamidi and Anissa Zerouki for their continued friendship since high-school and for being amazing women. To my neighbors and second family with whom I grew up: the M'sili family. To Saliha, Faïza, Lyès, Mohammed, Rabiâa, Redouane, Toufik and Zoubida. On top of all the "twayèches", they taught me how to share even when you have very little. To Mounira Amer-Yahia Temim, my best friend, for the pureness of her heart, for nights and nights of laughter, confidences, and heart-to-heart discussions

I am grateful to various members of my family for their support and help. These are Amor Chebira, Abdelkader Bouchentouf, Mourad Gaham and Karim Gaham.

My aunt, Adra Gaham, has been continuously encouraging me and helping in any way she could. I would like to thank her for always treating me like her own daughter, for teaching me to never give up, for her immense generosity, for being there and for proofreading this manuscript despite all the "undecipherable math stuff around"!

I would like to thank Lotfi Bacha, my "big brother" and childhood companion, for so many of my best memories ever, for constantly making me laugh to tears, for all the dancing and the music in my life, for all the scars on my body and simply, for the joy and happiness he brought to my life.

Finally, I would like to thank the two people who have been there every step of the way: my parents.

Papa: merci d'avoir toujours su dire les mots justes au bon moment, merci d'être le roc zen vers qui j'ai toujours pu me tourner, merci d'avoir toujours cru en ma bonne étoile, d'avoir remonté en courant le Blvd. St Michel avec moi pour de simples photos, d'avoir repeint ma chambre à Coubertin et de faire de notre famille une priorité de tous les jours. Merci pour tous les "bravo pour la maîtrise", et tous les regards et rires complices!

Maman: merci de m'avoir acheté des paniers de livres quand j'etais jeune, d'avoir refusé que je me retrouve a Bab Ezzouar, de m'avoir repété maintes fois qu'un 20/20 à la cité des Annassers ne valait rien dans le reste du monde. Merci d'être l'artiste et le pillier de notre famille, pour toutes les "chakhchoukha", et d'être le plus performant mutli-task, multi-processeur que je connaisse! Merci de me connaître et me comprendre si bien!

A vous deux: merci d'être généreux, aimants, honnètes, et toujours présents. Merci de m'avoir respecté, ainsi que mes choix. Merci de m'avoir soutenu dans tous les moments difficiles et dans les choix à faire. Enfin, merci d'avoir été mes guides et d'avoir tout sacrifié pour que je réalise mes rêves.

Acronyms and Notations

List of Acronyms

AOM	Acute otitis media
D	Dimensional (1D, 2D, 3D, ...)
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DNA	Deoxyribonucleic acid
DWT	Discrete wavelet transform
FB	Filter bank
ES	Embryonic stem
FFT	Fast Fourier transform
FIR	Finite impulse response
GFP	Green fluorescent protein
H&E	Hematoxylin and Eosin
HTF	Harmonic tight frame
LOT	Lapped orthogonal transform
LTFT	Lapped tight frame transform
MD	Multidimensional
MR	Multiresolution
MRI	Magnetic resonance imaging
NN	Neural network
OME	Otitis media with effusion
ONB	Orthonormal basis
PCA	Principal component analysis
PJB	Princen-Johnson-Bradley
RNA	Ribonucleic acid
RNAi	RNA interference
SVM	Support vector machine
SWT	Stationary Wavelet Transform
T_1	Original Haralick texture set
T_2	Modified Haralick texture set
T_3	Newly proposed Haralick texture set
TF	Tight frame
TM	Tympanic membrane
URI	Upper respiratory infection

List of Notations

A	Lower frame bound
A_j	Redundancy of a MR decomposition at level j
B	Upper frame bound
C	Number of classes
$\mathcal{C}$	Compact convex set in $\mathbb{R}^N$
$\mathcal{C}^c$	Complementary set of $\mathcal{C}$ in $\mathbb{R}^N$
$\hat{\mathcal{C}}$	Compact set approximating $\mathcal{C}$
$d(\hat{\mathcal{C}}, \mathcal{C})$	Distance between $\hat{\mathcal{C}}$ and $\mathcal{C}$
$D_{\tilde{\Phi}^*, \Omega}$	Decision function associate with $\tilde{\Phi}^*$ and Ω
$d_s^{(r)}$	Local decision made by subband s for an image r
$D_{c,s}^{(r)}$	cth element of vector $d_s^{(r)}$
$D^{(r)}$	Matrix of size $C \times S$ whose elements are $D_{c,s}^{(r)}$
$\delta^{(t)}$	Decision vector for test image t
Δ	Modulating window (diagonal matrix), $\Delta = \{\delta_k\}_{k=0}^{2N-1}$
$\mathcal{E}_p(\mathcal{C})$	Set of points at which classification errors due to the noise with probability density function p occur more than half the time
$\mathcal{E}_p(\mathcal{C}, (\tilde{\Phi}^*)^{-1}(\Omega))$	Set of points at which classification errors occur more than half the time due to the approximation of $\mathcal{C}$ by the frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$ in the presence of noise
F	Frame operator
$f_i^{(T_r)}$	ith Haralick texture feature (scalar) from the set T_r with $r = 1, 2$ or 3
g	Lowpass filter
$G(z)$	Polyphase expression of g
G	Grammian operator
h	Highpass filter
$H(z)$	Polyphase expression of h
$\mathbb{H}$	Hilbert space
I	Identity matrix
J	Anti-diagonal matrix
$\mathcal{J}$	Depth of a decomposition tree
$\mathbb{J}$	Index set of retained columns when seeding
$\chi_{\mathcal{C}}$	Characteristic function of $\mathcal{C}$
L	Length of filters
$m(\mathcal{C})$	Lebesgue measure of the set $\mathcal{C}$
M	Number of frame vectors
N	Dimension of the space ($\mathbb{R}^N$ or $\mathbb{C}^N$)
Ω	Subset of $\mathbb{R}^M$. In particular, we consider $\Omega = \prod_{m=1}^{M} [a_m,\ b_m]$ with a_m, b_m scalars in $\mathbb{R}^M$
p	Probability density function of a radially symmetric noise

List of Notations

ψ_m	Basis vector
Ψ	Collection of basis vectors or matrix whose columns are basis vectors
$\tilde{\Psi}$	Dual basis
φ_m	Frame vector
Φ	Collection of frame vectors or matrix whose columns are frame vectors
Φ^*	Hermitian transpose of the matrix Φ
$\Phi_p(z)$	Polyphase version of matrix Ψ
$\tilde{\Phi}$	Dual frame
$(\tilde{\Phi}^*)^{-1}(\Omega)$	preimage of Ω via $\tilde{\Phi}^*$, named frame set
S	Number of subbands (used in the MR classification algorithm)
R	Number of training images
x	signal: real, complex or of finite energy
X	Transform coefficients of x
W_N	Nth root of unity
w	weight vector
w_s	weight of subband s
W	Weight matrix of size $S \times C$. Each column is a class-specific weight vector

Part I

Introduction

Chapter 1

From Biomedical Applications to Frames and Back

Systems biology entails the study of the interactions between the components of a biological system as well as how these interactions give rise to function and behavior of that system. Thanks to the genome projects, we are witnessing an explosion in the "omics" areas, such as genomics and proteomics. At the same time, advances in biochemistry, probes, and microscopy gave the biologists the opportunity to observe cells and cell processes at a level never seen before, which lead to the collection of huge amounts of two-, three- and even higher-dimensional data. As a result, visual inspection of these data sets, always error-prone, nonreproducible and subjective, became impractical as well. A pressing need has therefore arisen for automatic systems which can extract knowledge from collected date in an accurate and efficient way.

In medicine, imaging techniques such as x-ray, ultrasound, tomography and magnetic resonance imaging have existed for many years. In the recent years, the field has undergone a shift in the methodologies and instrumentations used for diagnosis and surgeries, relying more and more on computerized tools. New medical instruments allow for the collection of large volumes of new data sets that are in need of automated processing, accurate and fast interpretation. For instance, many computer-aided and/or automated diagnosis tools have emerged in oncology and are sometimes used for early detection of cancers. These can assist, complement and help physicians in the decision making process.

Classification is a fundamental task in image processing. With regard to biomedical applications, classification was identified as the underlying problem to determine protein subcellular location patterns [13]. That was true for the project of determination of developmental stages in fly embryos, as well as the development of teratomas in stem cells, where multiple tissues are present and need to be recognized, and in the identification of middle ear infection stages. An accurate and efficient algorithm for classification would be of great use to biologists and physicians, motivating the developments in this work.

In some of the above developments, excellent results were achieved with the introduction of the simplest multiresolution (MR) tools, leading us to postulate that

using more sophisticated ones would lead to more accurate classification. Nonredundant MR tools—MR bases, in their adaptive incarnation, have been used with great success in fingerprint recognition. In the same problem, the authors observed that the translation variance of these bases might pose a problem and suggested to consider redundant MR techniques—frames, leading to the overall goal of this work:

To develop a mathematical framework and an accurate and efficient classification algorithm for the classification of biomedical images, based on adaptive and redundant multiresolution techniques.

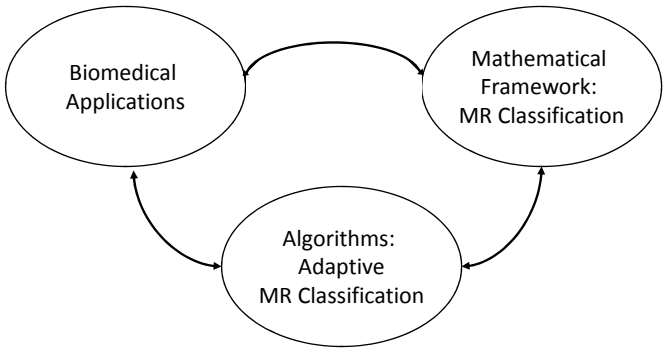


Figure 1.1: The big picture: From biomedical applications to mathematical framework to algorithms and back.

Given the considerations above, our work is focused on two main themes: The first is the design of an adaptive multiresolution classification algorithm for biomedical image data sets, while the second is a the development of a theoretical framework for redundant multiresolution classification and the design of new frame families. We now discuss each of these themes in more detail.

Algorithm: Multiresolution Classification for Biomedical Applications. Motivated by several emerging problems of biomedical imaging, we design a new, accurate and efficient classification algorithm. This algorithm takes advantage of the adaptivity and localization properties of multiresolution techniques. The philosophy behind using MR tools is to exploit the space-frequency information that lie in the biomedical data sets and be able to extract a faithful representation of these images. The extracted features would then provide discriminatory information that helps the classifier distinguishing between the various classes. We derive different instances of this MR classification algorithm drawing our inspiration from the specific biomedical applications of interest. In each case, the outstanding performance

of the classification algorithm when using redundant MR tools enable us to investigate and question the foundations of frame classification. This leads to the second main theme of this thesis:

Mathematical Framework: Theory of Frame Multiresolution Classification. In all of our biomedical applications, redundant multiresolution transforms performed the best in terms of classification accuracy. This led us to the following question: Why do frames perform better than bases? The answer to this question enables a better understanding of the role of redundancy and specific frame properties in the classification context. It also allows to predict those classes of signals for which a frame-based approach would be advantageous.

Given their success in this work, we also develop new redundant frame families. To better serve applications, these need to be efficient transforms in which the amount of redundancy is controllable. The design of such families enriches the frame toolbox dedicated to applications, and thus, offers a larger choice on the menu for the frame family that will best suit the application at hand.

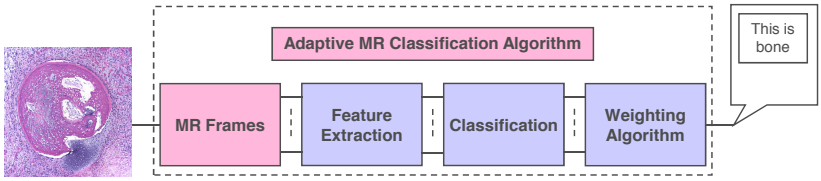


Figure 1.2: A diagram of the adaptive multiresolution classification algorithm developed in this work.

1.1 Thesis Contributions and Outline

Contributions.

1. **Multiresolution Classification Algorithm.** We revisit the classification problem and develop an accurate and adaptive classification algorithm as well as a toolbox made freely available online. The classification system adds multiresolution decomposition in front of a generic classifier: features are computed in each multiresolution subspace, yielding local classification decisions, which are then combined into a global decision using a weighting algorithm (see Fig 1.2). This approach is new and differs from what is traditionally done in multiresolution classification. In our work, we consider each multiresolution subspace itself an image to be classified.

   Given the very high accuracies we obtain with this algorithm for most of the applications, we demonstrate that the space-frequency localized information in the multiresolution subspaces significantly improves the discriminative power of a classification system. Moreover, we show that a small number of features is sufficient. Finally, we prove that frames are the class of multiresolution techniques that performs the best in the context of these biomedical applications.

2. **Biomedical Applications.** For all of the biomedical applications under study, our algorithm proves very accurate on those applications on which we have been working for a while. It gives promising results for the ones we just started investigating. We improve the classification accuracy of subcellular protein location pattern images to 95.4%. For the determination of ventral furrow formation of *Drosophila* embryo stages, we achieve a high accuracy of 98%. We reach an accuracy of 87.72% in the recognition of tissue types in histological stem-cell teratomas images. The diagnosis of otitis media infection is an application we recently started working on and our initial efforts lead us to an accuracy of 73.43%. Finally, we reach an accuracy of 99.5% for recognition of fingerprint images.

3. **Theory of Frame Classification.** When using the adaptive MR classification algorithm, MR frames always outperformed MR bases and afforded the best classification results in all applications without exception. This prompted us to ask two fundamental questions, the first one being: Why do MR frames perform better than MR bases in a classification context? The second question is the subject of our next item, namely: Can we design new frame families custom-tailored to the problem of biomedical image classification? As our first question is nontrivial in scope, we focus on the more tractable problem of establishing a rigorous mathematical framework for the analysis of frame-based classification. In particular, we propose a classification scheme using finite frames in $\mathbb{R}^N$. We consider a special case of classification, designing maps from a space of signals to a space of class labels that determine whether or not a given point in $\mathbb{R}^N$ belongs to a given compact convex set. We also introduce a measure-theoretic framework for the analysis of classification errors,

and apply it to the study of our proposed classification scheme. In particular, we show that this scheme performs well in the presence of noise.This mathematical framework allows to set the foundations for a theory of frame classification and provides rigorous tools to permit the development of more powerful classification algorithms. It also gives initial results for the characterization of those classes of signals which may be accurately classified using frames.

4. **Lapped Tight Frame Transforms.** As the success of the adaptive MR algorithm lends insight into the important role of frames in classification tasks of biological and medical image data sets, we sought to design new frame families we term lapped tight frame transforms. These can be seen as a redundant counterpart to bases known as lapped orthogonal transforms as well as an infinite-dimensional counterpart to harmonic tight frames. In four specific cases, we show that in addition to being tight, lapped tight frame transforms possess many desirable properties, such as equal norm, maximal robustness and efficient implementation. In the MR classification algorithm, the frame representation that is the most accurate is also the most expensive in terms of computational cost. This new family of frames has the advantage of being simple to design and affords control over the amount of its redundancy. This allows the user to customize the trade-off between efficiency and accuracy. In addition to providing custom-tailored frame transforms, the design of this new family enriches the frame toolbox and offers a larger choice in the pool of redundant MR representations.

5. **Broader Impact.** While we do concentrate on a few specific applications, we stress that the tools we develop serve multiple purposes: Classification is a fundamental image processing task and seems to be ubiquitous in biological as well as medical imaging. More broadly, these tools will be useful for automated analysis and interpretation of generic biomedical image databases. In the long term, this work will contribute towards the deployment and widespread use of complex and integrated, biomedical imaging systems. Finally, the underlying mathematical framework might allow us to look beyond the sole task of classification and possibly benefit other applications.

Outline. We divided this manuscript into three main parts. The first part presents some necessary background. In the second part, we detail our MR classification algorithm and study its performance in the various applications we consider in this work. The third and final part presents our work on the theory of frame multiresolution classification. We detail the outline of this thesis as follows.

- Part II consists of Chapters 2, 3 and 4, and presents all the necessary background and fundamental concepts used in our work. We being in Chapter 2 by introducing the specific biomedical applications we shall consider in this work. We present each application and motivate the need for an automated and accurate classification algorithm. The first two applications are biological

in essence, whereas the next two are biomedical. The last application is in biometrics, and while the focus of this work is biomedical applications, we use fingerprint recognition as a proof of concept of the universality of our work. Chapter 3 is an overview of classification, and presents the two main components of standard classification systems namely, feature extractors and classifiers, along with well-established examples for each. The last chapter in this part present multiresolution techniques. We first present nonredundant MR techniques which are bases, look at them via filter banks, and study important examples. We then motivate the need for redundancy through a fingerprint recognition and present redundant MR techniques, which are frames. We look at frames through filter banks, study their important properties and provide examples of frame families that we later use in our classification system.

- Part III consists of Chapters 5 and 6. In Chapter 5, we detail our adaptive multiresolution classification algorithm. Then, in Chapter 6, we experimentally evaluate the performance of our algorithm in each of the biomedical applications that were introduced in the previous part.

- Part IV of this manuscript presents our theoretical results in frame classification. In Chapter 7, we introduce a new mathematical framework for the study of frame-based classification, and provide results showing that for specific classes of signals, frames indeed outperform bases in a classification context. Finally, in Chapter 8, we design new frame families we term lapped tight frame transforms. We prove equal norm and maximal robustness of four lapped frame families.

We conclude this thesis by summarizing our work and proposing new venues and ideas for the future.

Over the course of this thesis, our philosophy has been to start from biomedical applications, evaluate their underlying algorithmic needs and provide solid and sophisticated solutions supported by a mathematical framework. Hence, we go from biomedical applications to redundant multiresolution tools and back, closing the loop and bridging disciplines that inspire, challenge and enrich each other.

Part II

Background

Chapter 2

Biomedical Applications

Contents

In this chapter, we show that, by looking into five different biomedical/biometric applications, the underlying problem is classification and thus, an accurate and efficient algorithm would be of great use.

Systems biology entails the study of the interactions between the components of a biological system and the mechanisms by which these interactions give rise to the function and behavior of that system. It can be viewed as a "macro" approach that encompasses mathematical and computational modeling based on quantitative data collected within each component of the biological system. Thanks to the genome projects, we are witnessing an explosion in the *"omics"* areas. These pertain to the study of a data of a particular type, for instance proteins (proteomics), from a specific biological system.

The Need for Automated Processing Advances in biochemistry, probes, and microscopy gave the biologists the opportunity to observe cells and cell processes at an accuracy never seen before, which led to the collection of huge amounts of 2D, 3D and even higher-dimensional data. By the same token, physicians have had the chance to collect very large amounts of new data. Medical images have typically been in use for much longer than biological ones through numerous imaging modalities such as X-rays, hence the medical community has addressed the problem

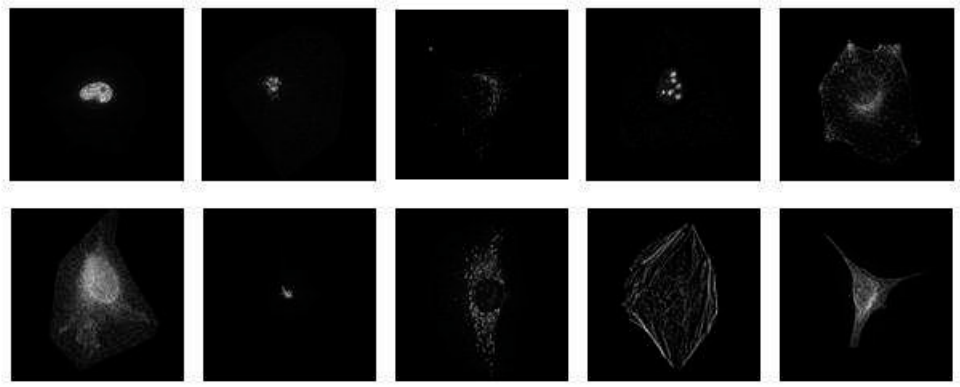


Figure 2.1: Typical images from the 2D HeLa collection. Top row, left to right: DNA, giantin, lysosomal, nucleolar, endosomal (Tfr). Bottom row, left to right: endoplasmic reticulum, gpp130, mitochondrial, actin, tubulin. (Images courtesy of Dr. R. F. Murphy [87].)

of automation of the processing much more so than biologists. However, thanks to the evolution of medical instrumentation and modernization of clinical tools and protocols, the sheer volume of collected medical images and their novelty make automation a crucial step in medical practice. As a result, visual inspection of these data sets, always error-prone, nonreproducible and subjective, became impractical as well. Hence the need for automated, accurate and efficient systems to extract knowledge contained in the collected data.

Classification of Biomedical Images Such automated knowledge extraction requires the expertise developed in signal processing, machine learning and mathematics. In the project of determination of protein subcellular location patterns described in Section 2.1, Murphy et al. identified classification as the underlying problem [13, 62]. Similarly, in the project of determination of developmental stages in fly embryos described in Section 2.2 [65, 49], we realized that the problem is again that of classification. Not surprisingly then, in several other projects, such as the development of teratomas in stem cells, where multiple tissues are present and need to be recognized (Section 2.3), the diagnosis of otitis media stages (Section 2.4) as well as fingerprint recognition (Section 2.5), the need for classification emerged. Thus, an accurate and efficient algorithm for classification would be of great use to biologists and physicians, motivating the developments in this work.

2.1 Determination of Protein Subcellular Location Patterns

The field of proteomics entails the study of proteins and their role and function in various cellular mechanisms. One of the critical characteristics of a protein is its subcellular location, that is, its spatial distribution within the cell. Knowledge of the location of all proteins will be essential to build accurate models that capture

and simulate cell behavior, and eventually can be expected to be useful for early diagnosis of disease and/or monitoring of therapeutic effectiveness.

Cancer cells are often used in studies of this nature. These are abnormal cells showing temporally unrestricted growth preference over their normal counterparts. Cancer cells are able to survive in crowded conditions and thrive in an anaerobic environment [47]. These properties make the culturing of these cells easier than their normal counterparts. Among the collections of cell cultures available are Chinese Hamster Ovary (CHO) cells, Colon cancer (CoCa2) cells, 3T3 cells (developed from a strain of Swiss mice) and HeLa cells (cervical cancer cell cultures prepared from malignant cells donated by Henrietta Lacks in 1951). The HeLa collection, the data set used in this work, is a well-established testbed for evaluating subcellular pattern analysis approaches.

The most widely used method for determining protein subcellular location is fluorescence microscopy, its success due in part to the advent of a range of new fluorescent probes. The first step in the acquisition process of fluorescence microscope images is the culture of cells. This involves transferring pre-harvested cells onto a substratum that emulates their natural growth environment. The second step is preparing the cells for imaging. The cells are stained with specific dyes. While staining is convenient and fast, the stains are chemicals beset with the disadvantage of nonspecificity. A technique that addresses the issue of specificity is tagging. Tags are special protein sequences or antibodies that are introduced into the DNA sequence. They are transcripted along with the sequence coding a specific protein [47]. Of particular interest is a recent fluorescent probe that is nontoxic, called green fluorescent protein (GFP). The third step is imaging. This is done using different microscopes, either the spinning disk or the confocal scanning laser microscope. The spinning disk is faster but has lower resolution than the confocal microscope. Both microscopes have excitation and emission filters that excite and capture the fluorescent light emitted by the markers introduced in the cells. Finally, the images produced are stored as a multidimensional data set.

The Need for Classification Given that mammalian cells are believed to express tens of thousands of proteins, comprehensive analysis of protein location requires acquisition of images whose numbers are beyond our ability to analyze visually. Moreover, these data sets very often present a challenge in terms of recognition of the type of proteins they depict. For example, the Golgi proteins giantin and gpp130 are indistinguishable by the human eye. Finally, as today's microscopes allow for imaging of high-dimensional data sets, both the enormous volume as well as the high dimensionality of the data render human analysis time-consuming, prone to error and ultimately, impractical, leading to the "holy grail" of protein subcellular location image interpretation and analysis: a system for fast, automatic, and accurate recognition of proteins based on their subcellular location. Murphy et al. have pioneered the use of automated systems for protein identification based on their subcellular location patterns [13, 62].

2.2 Detection of Developmental Stages in *Drosophila* Embryos

The genome projects have brought unprecedented opportunities to understand molecular mechanisms of development and disease. *Drosophila* (small fruit fly) sequences are of special interest because the fly serves as an important model organism for developmental and cellular processes common to higher eukaryotes, including humans. Comparative genomics studies have revealed that D. melanogaster, for example, has orthologs to 177 out of 289 examined human disease genes [105]. The genome sequence of D. melanogaster was published in 2000 [2], followed by the sequence of *Drosophila* pseudoobscura in 2005 [101].

While the *Drosophila* genome projects provide us with a wealth of data, the determination of the functions of the genes that are inferred from these sequences (approximately 13,600 genes for D. melanogaster) is an arduous task that requires novel, highly efficient and high-throughput screening methods [19] and methods for automated phenotype analysis [133].

RNA interference (RNAi) is one such method that can be used to silence a specific gene in a cell or an organism [43], as RNAi pathways play a major role in regulating development and genome maintenance. Analysis of a change in phenotype due to gene silencing indicates the function of the silenced gene. Silencing a gene in an entire fly embryo through RNAi requires injection of embryos with designed, double-stranded RNA (dsRNA) early in embryonic development, prior to the formation of the syncytial blastoderm. A powerful MEMS-based system for automated, high-throughput injection of *Drosophila* embryos has been recently proposed [132]. Phenotype analysis after gene silencing is greatly facilitated through genetic engineering of *Drosophila* embryos that express, for example, green fluorescent protein (GFP) in a tissue of interest [130].

In the project led by Minden at CMU, the formation of the ventral furrow is observed in early embryonic development. Ventral furrow formation is a key morphogenetic event during *Drosophila* gastrulation that leads to the internalization of mesodermal precursors [49]. Once a gene is silenced using RNAi, the fly embryos are tagged with GFP and imaged using a confocal laser fluorescence microscope. This acquisition process allows for the collection of z-stack images of entire embryo volumes in time [65, 49], that is four-dimensional data sets.

The Need for Classification *Drosophila* gastrulation involves four major morphogenic events, the first one being ventral furrow formation, which can be divided into three steps: (1) The initial stage is when the ventral furrow is yet to form. (2) The beginning of Stage 2 consists of cells migrating basally as a result of their nuclei losing their apical attachment. During this step, about half of the cells in the central patch undergo shape changes over a 10- to 12-minute period. This is sufficient for the entire ventral furrow to collapse inward, bringing a band into the interior of the embryo over a period of several minutes (end of Stage 2). (3) Stage 3 consists of having a closed and formed ventral furrow. To study this process, Minden at al. acquire 3D volumes of the formation of the ventral furrow in time.

As manual processing of huge amounts of these 4D (space + time) data sets is impractical, cumbersome and error-prone, we believe that reliable, accurate, flexible, and efficient algorithms for automated 4D image and phenotype analysis are crucial to enable high-throughput functional genomics screens such as this one.

2.3 Identification of Histological Stem-Cell Teratomas

The study of stem cells is one of the most exciting and promising research areas in the biomedical field. Embryonic stem cells (ES) and cells derived from them hold great promise, both as therapeutic agents in clinical medicine, as well as biological windows into the early stages of development. The range of therapeutic options includes repair of damaged or injured tissue (tissue regeneration after stroke, heart attack, cartilage renewal in arthritis), restoring defects in genetic, biochemical, and metabolic pathways, as well as drug testing and discovery [77, 95, 119]. Studying the sequence of genetic events within ES cells as they develop and differentiate into tissue will have a significant impact on explaining and ultimately defining the therapy for a wide range of developmental syndromes. ES cells possess certain inherent characteristics that set them apart from any other cell type. They have the ability to self-renew, perpetuate indefinitely, and produce all three germ layers from which all tissue types are derived (pluripotency). Typically, in the laboratory, ES cells are defined by their expression of specific proteins and their behavior in cell culture. However, human and nonhuman primate cells isolated and cultured cannot be considered ES cells until they show the ability to produce a teratoma tumor when injected into immunocompromised mice. A teratoma is a tumor that is strictly defined by histological evidence of tissue types contributed by each of the original three germ layers. These include ectoderm (neuroepithelium, mature neuroglial tissue, skin), mesoderm (smooth and skeletal muscle, connective tissue, bone, and cartilage), and endoderm (lung and intestinal mucosa, pancreas, liver). While at first glance, most teratomas derived from ES cells appear as disorganized tissue masses with recognizable germ layer elements, little is known about the contribution of each germ layer to the lesion, and this information may hold important clues to normal and abnormal development.

In the project led by Castro and Ozolek at the University of Pittsburgh Medical School, the aim is to answer some of these questions using high-field magnetic resonance imaging (MRI) and histological staining methods. First, stem cells are implanted in the testis of a rat that was genetically altered to have a deficient immune system. Implanting the cells into the testis further ensures that the immune system will not interfere with the division and growth of the cells into a tumor. With no control over the differentiation process, the implanted stem cells develop into various tissue types: muscle, epithelial, brain matter, cartilage and bone, to name a few. This tissue-rich tumor is then surgically removed intact from the animal, imaged using MRI, and subsequently stained with Hematoxylin and Eosin (H&E) and then imaged.

MRI images are taken of the tumor in the coronal, sagittal, and axial planes. The field is on the order of 10[T], allowing for cellular-level resolutions. While MRI permits tracking of stem cells *in vivo*, its nature often leads to poorly defined

contours, low contrast in the 3D volumes and prohibits resolutions which would match histological images. Therefore, MRI is currently inadequate as a stand-alone imaging procedure.

Histology studies tissues that have been thinly sliced, and is an important diagnostic tool. First, the tissue is mechanically and biomechanically stabilized in a fixative. It is then embedded and sectioned into very thin (2 to 8 μm) sections using a microtome. The slices are then stained using one or more pigments to give contrast to the tissues. The most common stains used are H&E. Hematoxylin gives a blue color to the nuclei whereas Eosin gives a pink color to the cytoplasm (see Fig. 2.2). The images obtained through this process have a great advantage of being highly detailed and show distinctive features of the tissue at different resolutions. However, *in vivo* imaging is not possible and the tissue sample is destroyed preventing the biologists from observing the development of the tissue over time.

An important step towards understanding stem-cell division and differentiation would be to coregister cellular class identifications from the histological images to the difficult-to-read MRI data. One of the goals here is to enhance the diagnostic capabilities of MRI using histology images as ground truth.

The Need for Classification While sophisticated image analysis and bioinformatics are burgeoning fields within pathology, most of the imaging applications have focused on the automation and digitization of the tissue processed for histological examination. New techniques are in development that would allow the pathologist to manipulate high-resolution images and sign out cases at the computer rather than examine tissue under a microscope. Image analysis currently allows segmentation of tissue areas defined by specific immunohistochemical stains to highlight the tissue of interest. However, the ability to automate the recognition of specific and varied tissue types from the routine H&E stained tissue sections (used almost exclusively to make the vast majority of diagnoses) is not available. As this applies to ES cell biology and teratoma analysis, advanced digital imaging applications would help answer the following questions: How much of each specific tissue type is present? How are these specific tissues arranged in space with respect to one another? How are the present tissues affected in type and quantity when derived from ES cells that have been manipulated genetically, biochemically, and environmentally (that is, by drugs or toxins)? The ability to accurately detect and quantify specific tissue types will allow detection of species-specific differences in developmental programming and enable accurate three-dimensional reconstruction of teratomas and precise correlation with high-resolution MRI. Since the teratoma contains many varied tissue types, the first step in this direction is to actually recognize these specific tissues types, a classification problem.

2.4 Classification of Otitis Media Stages

Otitis media is an inflammation of the middle ear. It occurs in the area between the ear drum (the end of the outer ear) and the inner ear, including a duct known as the Eustachian tube. It is one of the two categories of ear inflammation that can underlie what is commonly called an earache, the other being otitis externa. Depending on

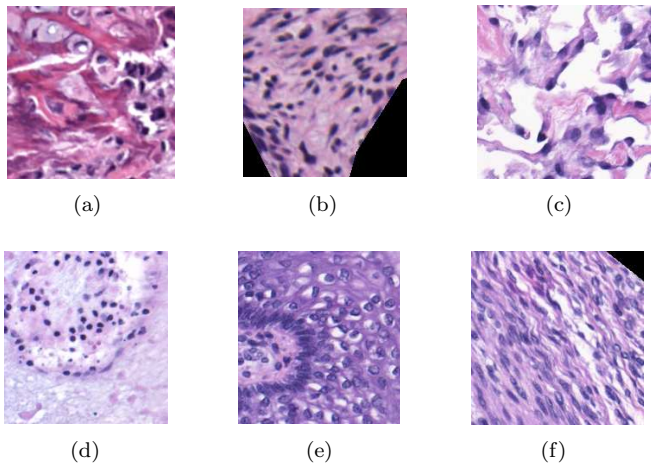


Figure 2.2: Sample H&E-stained images from each tissue class: (a) bone, (b) mesenchyme (embryonic connective tissue), (c) myenteric plexus, (d) necrotic tissue, (e) skin, and (f) striated muscle. (Images courtesy of Dr. J. A. Ozolek and Dr. C. A. Castro, University of Pittsburgh Medical Center [92].)

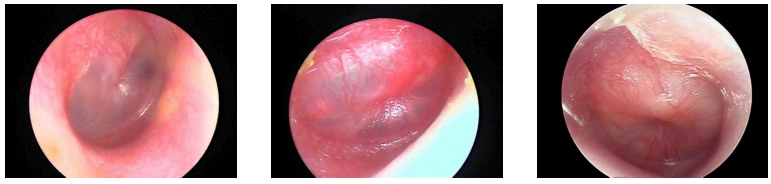


Figure 2.3: Otitis media sample images: (a) Normal ear (no infection), (b) otitis media with effusion (OME) and (c) acute otitis media (AOM). (Images courtesy of Dr. A. Hoberman, University of Pittsburgh Medical Center [56].)

its severity, otitis media can be divided into two categories (see Fig. 2.3): acute otitis media (AOM) and otitis media with effusion (OME). AOM, the most severe form of otitis media, usually arises as a complication of a preceding viral upper respiratory infection (URI) such as a cold or a sore throat. AOM is most often purely viral and self-limited, as it usually accompanies viral URI. If the middle ear, which is normally sterile, becomes contaminated with bacteria, fluid and pressure build up in the middle, resulting in bacterial AOM. Viral AOM can sometimes quickly lead to bacterial AOM ,especially in children. Symptoms of bacterial AOM include the classic earache, severe and continuous pain, and fever. Complications of bacterial cases can be severe, such as perforation of the ear drum, infection of the mastoid space (mastoiditis) and, in very rare cases, meningitis [57].

OME, also called serous or secretory otitis media, is defined as the presence of fluid in the middle ear without signs or symptoms of acute ear infection. This

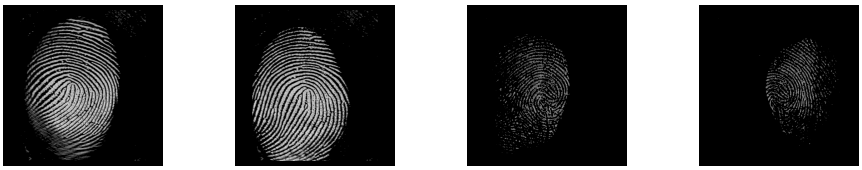


Figure 2.4: Example fingerprint images from an easy class (left) and a difficult class (right). (Images courtesy of NIST [127].)

fluid builds up as a result of the negative pressure produced by altered Eustachian tube function. The tube is blocked by the swelling of its lining or plugged with mucus due to a cold (or some viral URI), and is unable to open to ventilate the middle ear. This lack of ventilation is what causes the fluid to accumulate. If the tube remains plugged, the fluid collects in the normally air-filled middle ear. Continuous presence of middle-ear fluid from OME results in decreased mobility of the tympanic membrane and becomes a barrier to sound conduction, leading to hearing impairment. OME can precede and/or follow a bacterial AOM [103].

Distinguishing between AOM and OME is an important but difficult task. Although OME is more common than AOM, it is often mistaken for AOM. When this is the case, antibiotics are prescribed unnecessarily.

The Need for Classification Otitis media is very common, especially in children. In the United States, 50% of children have an episode before their first birthday and 80% of children are affected by their third birthday. An estimated $5 billion is spent each year on care of patients with AOM and related complications [57, 109]. Beyond the cost, the prescription of antibiotics in otitis media cases has been a subject of controversy. The initial diagnosis of otitis media is usually performed by a primary-care physician and is based on otoscopy and symptomatology. Very often, AOM is over-diagnosed, that is, most physicians prefer to be on the safe side and prescribe an antibiotic treatment when they observe an infection of the ear, which in many cases is in fact OME [57, 103]. The problem with that is the resistance developed to such treatments, lessening the effect of the currently available medication. This later leads to the necessity of combining multiple antibiotics to make for an effective treatment. As a result, multi-drug-resistant bacterial pathogens spread, making current drugs ineffective. Moreover, it is hard to diagnose accurately the stages of otitis media (in particular in infants due to the language barrier), even for well-trained physicians. The certainty of diagnosis of AOM is only 58-73% [109]. This clearly calls for automated classification systems that are accurate and upon which physicians can rely.

2.5 Application in Other Domains: Fingerprint Recognition

While fingerprint recognition is not a biomedical application and is traditionally considered a security application, we use it in this work, as a proof of concept of the

flexibility of our classification system. One could even argue that it is a biomedical application since it concerns biometric characteristics of human beings. In our work, we use it rather as a proof of concept for the universality of our classification system.

Personal identification has been a topic of interest for some time, with various solutions proposed. Accessing buildings or facilities, withdrawing money or using a credit card, and gaining access to electronic information on a local computer or over the Internet, are all examples of situations which require accurate and reliable methods of personal identification, and solutions vary greatly. There are hundreds of modalities for personal identification, from items one might keep in one's possession (for example, identification cards or keys) to combinations of numbers and information one might memorize (for example, Social Security numbers and passwords). Using human biometric characteristics (fingerprints, irises, faces, etc) has great advantages over other techniques: the information cannot be lost or forgotten, and forgery requires greater skill. Most prominent amongst biometric characteristics are fingerprints. Because of their uniqueness, consistency over time and ease of acquisition, fingerprints have been the most widely used and researched area of biometrics. Using fingerprints for recognition of individuals started in the late 19th century. Sir Galton defined the characteristics from which fingerprints can be identified—"Galton points". Later, with the advent of computers, a subset of these points, now termed minutiae, were used in automated fingerprint recognition systems. In 1969, overwhelmed by its growing database and the manual processing required, the Federal Bureau of Investigation led a major effort in the development of new automated and accurate fingerprint recognition systems. Two decades later, this lead to the famous Integrated Automated Fingerprint Identification System (IAFIS)[88]. With this came the need to develop acquisition and sensing systems. There are two main categories of sensing: off-line (ink based) and live-scan. The second technique is the most widely used nowadays and almost all sensing tools belong to one of the three following families: optical sensors, solid-state (or silicon) sensors and ultrasound sensors [80].

The Need for Classification In a world where an ever increasing need for identification is present and identity theft is a problem authorities and consumers face daily, fingerprint recognition systems have become a necessity. A 2003 survey sponsored by the US Federal Trade Commission estimates the annual total loss to businesses due to identity theft approached $50 billion. MasterCard and Visa fraud losses related to identity theft in 2000 equaled $114 million, an increase of 43% from about $80 million in 1996[31]. While identity theft is usually associated with the financial industry, many other system access points require reliable recognition of the person trying to access them. Examples are numerous: office buildings, secure access to data, etc. In each case, the person trying to access the system needs to be either confirmed or identified prior to being allowed access based on that persons biometric characteristics.

Depending on the application context, a fingerprint-based biometric system may be called either a verification system or a recognition (identification) system. The former outputs a binary answer yes/no to the question "is this person X?",

whereas the latter answers the question "who is it?". Both are a type of classification problem in which one individual corresponds to one class. A crucial goal in processing this biometric data is to do so automatically, accurately and quickly.

Chapter 3

Classification

Contents

Pattern recognition is a task that human beings perform a thousand times a day without even thinking about it. Whether to recognize familiar faces, objects and their function, words and their meaning or even the colors of the rainbow, we are in fact very accurate classifiers!

Many applications, such as face recognition and tracking, are in need of accurate recognition systems. As we have seen in the previous chapter, the need for classification of new —biomedical— data sets is pervasive, and ever better solutions are needed. In these biomedical applications, biologists and physicians often rely mostly on their training and experience to perform recognition tasks. They do so via manual and visual inspection of the images. Their years of experience, visual observations of salient characteristics and previously acquired knowledge allow them to decide which class label to assign to an image. This visual processing quickly showed its limitations.

First, the complexity of biomedical images makes it difficult to visually recognize the salient characteristics representing a class (for example, a protein type). Next, the sheer volume of these data sets, due to the emergence of high-throughput systems, make any manual processing cumbersome. Finally, subjectivity makes the decision process unreliable and nonreproducible. Therefore, building an artificial recognition machinery is crucial. In this chapter we present an overview of various classification methods. For more details on the theory of classification and pattern recognition, we refer the reader to the excellent books by Duda, Hart and Stork [40] and Bishop [10].

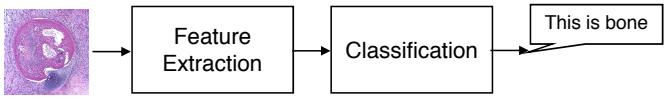


Figure 3.1: Generic classification system.

3.1 Overview of Classification Methods

Let us consider a popular pattern recognition problem, that of optical character recognition. This technique is used by the United States Postal Service to route mail by ZIP codes. It consists of recognizing handwritten numerical digits (0–9) from images of these digits that are of size 28×28. The goal here is to build an artificial "black box" that takes as input a vector representing one such image and outputs the identity of the digit depicted in the image. The presence of noise as well as the variability of handwriting make this task nontrivial. To aid in tackling this problem, samples of input vectors can be provided, forming the *training set*. This set can be associated with a *target set*. For each digit from the training set, there is a target vector that represents the identity of the digit. Target sets are constructed using prior information when the "classes" of the digits are known in advance. Then this set is built through a (usually manual) pre-labeling of the training samples. Using the training and target set, a learning algorithm can be used to tune the parameters of an adaptive model of the classes. This is called the *training phase*. In the *testing phase*, a new input *test* vector is fed into this newly-clever "box", then the output of the machine learning algorithm consists of the class that seems the most probable based on the previously learned patterns or characteristics of each class.

The input to machine learning algorithms does not always consist of the raw data. In fact, in most cases, it is impractical to use the raw data. In the example presented above, each raw vector would consist of $28^2 = 784$ real numbers. Depending on the size of the training set and the number of classes (here 10), the problem can rapidly become computationally infeasible. Indeed, in practical applications, one very often deals with larger images, or even higher-dimensional data. Moreover, the raw data may be too confusing in the sense that the algorithm only needs to see discriminatory information that will help build an accurate model for the data and will help distinguish between the different categories. For example, the digits may not fill the entire image and so many pixels surrounding the digit have value zero. It seems wasteful to use all of these pixels in the input vectors. Also, reducing the unnecessary variability (such as scale and position of the digits) in the data can only help the recognition process. Further, it is reasonable to assume that one only need to extract the "essence" of the data or the classes. In fact in practice, the true information content of the data is significantly less than its dimension would indicate. By representing classes by their most salient characteristics, we achieve two goals. First, only the important and helpful information is given to the algorithm

and second, we considerably reduce the dimensionality of the problem. This type of pre-recognition processing is called *feature extraction*. This is a sensitive step in the recognition process because it usually is difficult to identify exactly which characteristics are important (and how to compute them) without discarding crucial information.

Two important families of pattern recognition tasks are classification and clustering. Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as *supervised learning* problems. A subcategory of such problems is *classification*, where the aim is to assign each -test- input vector to one of a finite number of discrete categories—*classes*. In pattern recognition tasks where the data consists of a set of input vectors without corresponding target values are called *unsupervised learning* problems. When the goal is to find groups (clusters) of similar examples within the data, these tasks are called *clustering problems*. We present examples of classifiers and a clustering algorithm in Section 3.3

We now define more formally our classification problem.

3.1.1 Problem Statement

The problem we are addressing is that of classifying images from biomedical data sets. Assume that the images are of size $N \times N$ and let $\mathcal{R}$ denote the set of intensities covered by all the images in the given data set, compactly represented as an image belonging to $\mathcal{R}^{N \times N}$. Then, the problem can be formulated as designing a map from the *signal space* of the examined images $\mathcal{X} \subset \mathcal{R}^{N \times N}$, to a *response space* $\mathcal{Y} \subseteq \{1, 2, ..., C\}$ of class labels. Thus, decision *dec* is the map, $dec\colon \mathcal{X} \mapsto \mathcal{Y}$ that associates an input image with a class label [106]. To reduce the dimensionality of the problem, one sets up a feature space $\mathcal{F} \subset \mathcal{R}^f, f \leq N^2$, between the input space and the response space. The feature extractor θ is the map $\theta\colon \mathcal{X} \mapsto \mathcal{F}$, and the classifier ν is the map $\nu\colon \mathcal{F} \mapsto \mathcal{Y}$. The goal is to find a (θ, ν) pair that maximizes the classification accuracy.

3.2 Feature Extraction

As we mentioned earlier, feature extraction is an essential step in the classification process. We want to find features that are useful and fast to compute and yet that also preserve the useful discriminatory information in the data. Features are numerical descriptors that characterize the input data, usually in a lower-dimensional space. We focus here on the following feature sets:

3.2.1 Haralick Texture Features

Haralick texture features are calculated using four co-occurrence matrices [54, 53]. These matrices describe the way certain grey-levels occur in relation to other grey-levels. For example, an element of such matrix contains the number of times a pixel with grey-level i occurs at a certain distance from a pixel with grey-level j.

The four matrices are: 1) P_H (horizontal nearest neighbors), 2) P_V (vertical nearest neighbors), 3) P_{LD} (left diagonal nearest neighbors), and 4) P_{RD} (right

diagonal nearest neighbors). Haralick calculates 13 measures on each of these four matrices. For example, the first two features on P_H are:

$$f_{H,1} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(\frac{P_H(i,j)}{R_H}\right)^2, \qquad f_{H,2} = \sum_{k=0}^{N_g-1} k^2 \sum_{|i-j|=k} \frac{P_H(i,j)}{R_H}, \tag{3.1}$$

where N_g is the number of gray levels in the image and R_H is a normalizing constant equal to the sum of all the elements in P_H. Haralick features include entropy, contrast ($f_{H,2}$ in(3.1)) and angular second moment.

These other measures are computed in a similar fashion, giving us four sets of 13 measures: $f_{(H,1-13)}$, $f_{(V,1-13)}$, $f_{(LD,1-13)}$ and $f_{(RD,1-13)}$. Haralick's original method reduces these to a single set of 13 features by calculating the mean of each measure across the four sets (feature set T_1):

$$f_i^{(T_1)} = \frac{f_{H,i} + f_{V,i} + f_{LD,i} + f_{RD,i}}{4}, \tag{3.2}$$

for $i = 1, \ldots, 13$. An alternative method [54, 53] is to use both the mean and the range of the 13 measures, thus resulting in two sets of 13 features (26 features overall, feature set T_2).

3.2.2 Morphological Features

There are 16 of these features that visually describe distinctive aspects of images as discerned by the human eye. The main categories of morphological features are object, edge (boundary of the object), convex hull (a closed, convex contour of the object) and skeleton (a detailed grid of the image) features [63].

3.2.3 Zernike Moment Features

Moment features are based on the relationship between pixels based around a central point, usually the center of the image. They are computed by taking the inner product of the original image with a moment feature polynomial described radially outward around the central point [6]. A special type of moment feature, Zernike moments, are described by

$$Z_{hk} = \frac{h+1}{\pi}\sum_{i,j} im(i,j)\nu^*_{hk}(\rho,\theta),$$

where $\nu^*_{hk}(\rho,\theta)$ is the complex conjugate of a Zernike polynomial ν_{hk} and $im(i,j)$ is the original image bounded by the unit circle ($i^2 + j^2 \leq 1$). Hence, each of the Zernike moment features computed for an image is a similarity measure between the corresponding Zernike polynomial and the image. Zernike polynomials are distributions defined over the unit circle. Their behavior around the unit circle is controlled by two parameters, one that describes the number of times the polynomial rises and falls as it goes from the center of the circle to the perimeter (angular dependence k) and another that describes the fold of radial symmetry of the polynomial (degree h) [67]. The number of Zernike moments depends on k and h. In this work, we consider 49 moments. Zernike polynomials possess properties such as rotational invariance and orthogonality.

3.3 Clustering Algorithms and Classifiers

Most of our discussions focus on supervised learning. However, as our initial efforts in developing a pattern recognition procedure used a well-known clustering method, we briefly summarize it here.

3.3.1 K-means Clustering

K-means algorithm is the most popular and used clustering method. It is an iterative process that essentially tries to form clusters of —feature— vectors in a multidimensional space. These vectors are judged similar or close enough, according to some metric, to be grouped together. Namely, K-means operates over a fixed number of clusters K, while attempting to satisfy two properties. The first is that each cluster has a center which is the mean position of all the sample vectors in that cluster. The second is that each sample vector is in the class whose center it is closest to. More explicitly, suppose we have a data set $\{x_1, \ldots, x_T\}$ of N-dimensional vectors in an Euclidean space. The goal is to partition this set into K clusters. Assume $\{\mu_k\}_{k=1}^K$ is a set of vectors representing each cluster, these are in fact the centers of the clusters. The goal is to assign each data point to a cluster such that the sum of the distances of each point to its closest vector μ_k is minimized. That is, the goal is to find$\{\mu_k\}_{k=1}^K$ and $\{r_{tk}\}_{k=1}^K$ that minimize the objective function

$$J = \sum_{t=1}^{T} \sum_{k=1}^{K} r_{tk} \|x_t - \mu_k\|^2,$$

where r_{tk} is such that $r_{tk} = 1$ if x_t is assigned to cluster k and $r_{tk} = 0$ otherwise. The resulting optimal μ_k for cluster k is the mean of all points x_t assigned to cluster k, hence the name K-means algorithm.

We now go back to the main theme of this chapter: Classifiers. In general, classifiers can be divided into two categories: linear and nonlinear. Linear classifiers, such as linear discriminant functions, use a linear boundary function to discriminate between classes. This function is dependent on the training data and can thus introduce inaccuracies due to unseen data. Nonlinear classifiers, on the other hand, are more flexible, and are used when it is not possible to separate data into classes using linear functions.

There are several types of classifiers [40] such as Bayesian decision classifiers, linear discriminant functions, k-nearest neighbor classifiers, support vector machines (SVMs), and neural networks (NNs). While they work under different principles, their goal is the same. Bayesian decision theory is a probabilistic approach that attempts to divide data according to the probability density function that governs it. Linear discriminant functions use a linear function from a regression of training data as the boundary function between classes. k-Nearest neighbor classifiers return the most common class label among the k training examples nearest to x. SVMs try to find a surface that splits the data into regions. NNs operate as highly parallel simple processors that work under the principles of learning and adaptation. Below, we look into how the performance of a classifier might be evaluated then give a quick overview of some of well-established classifiers.

3.3.2 Evaluation of Classification Systems

An aspect of great importance in classification is the evaluation of a classifier's performance. The test data set is usually used to determine how "accurate" a classifier is. The classifier has never been exposed to the test data before, therefore, its ability to distinguish between classes using the test data is a good indicator of its functionality. The accuracy of an algorithm represents the conformity or closeness of the outcome of this algorithm, namely the class labels, to the ground truth or true values of the labels. Obviously, the goal in classification systems is to build classifiers with the highest accuracy possible. Note that here we assume that we have access to the ground truth or the gold standard.

There are two types of errors a classifier can make: The error of identifying the wrong class (a positive outcome when the reality is false) or the error of not identifying the true class (a negative outcome when in fact the reality is true). The first one is referred to as false positive (F^+) or Type-I error while the second is called a false negative (F^-) or Type-II error. Similarly, a classifier can be correct by agreeing with the truth in a positive or negative way: Have a positive outcome with the true class (T^+) or have a negative outcome with the wrong class (T^-) (see Table 3.1).

		Ground truth	
		True	False
Algorithm output	Positive	T^+	F^+
	Negative	F^-	T^-

Table 3.1: Outcome of a classification algorithm compared to the ground truth.

We define the accuracy of a classification algorithm as

$$\text{Accuracy} \;=\; \frac{T^+ + T^-}{T^+ + F^+ + T^- + F^-}. \tag{3.3}$$

In practical recognition settings, it occurs often that an algorithm has no negative outcome, and only outputs a specific class label. In that case, T^- and F^- do not exist.

3.3.3 Bayesian Decision Theory

Bayesian theory is a fundamental approach to the classification problem based on the probability density functions governing the data being classified [40]. It makes the assumption that all probabilistic values associated with the problem are known. Given a problem with C classes, a classifier must decide which of these C classes a particular feature vector (or data point) x belongs to. First, the probability of observing x coming from the class $\mathcal{C}_k$ where $k = 1, 2, \ldots, C$ is

$$p(\mathcal{C}_k, x) \;=\; P(\mathcal{C}_k|x)p(x) = p(x|\mathcal{C}_k)P(\mathcal{C}_k).$$

Rearranging the terms, we obtain

$$P(\mathcal{C}_k|x) = \frac{p(x|\mathcal{C}_k)P(\mathcal{C}_k)}{p(x)}.$$

Now consider a two-class classification problem. For a given vector x and the observation that $P(\mathcal{C}_1|x) > P(\mathcal{C}_2|x)$, we would intuitively classify x as coming from class $\mathcal{C}_1$, assigning the input to the class that is most likely (maximum likelihood rule). Conversely, if $P(\mathcal{C}_2|x) > P(\mathcal{C}_1|x)$, we choose class $\mathcal{C}_2$. Given this decision scheme, the error associated with making a decision is defined as follows

$$P(error|x) \ = \begin{cases} P(\mathcal{C}_1|x) & \text{if we decide } \mathcal{C}_2 \\ P(\mathcal{C}_2|x) & \text{if we decide } \mathcal{C}_1 \end{cases},$$

and the average probability of error is defined as

$$P(error) = \int_{\mathbb{R}^N} P(error, x)\mathrm{d}x = \int_{\mathbb{R}^N} P(error|x)p(x)\mathrm{d}x$$

One can minimize this error by making $P(error|x)$ as small as possible. This equates to the following decision rule

$$\text{Decide } \mathcal{C}_1 \text{ if } P(\mathcal{C}_1|x) > P(\mathcal{C}_2|x) \text{ , otherwise, decide } \mathcal{C}_2.$$

By assigning costs to making each decision, one can minimize the error or cost by choosing the decision that produces the smallest cost given x. The basic ideas of Bayesian decision theory still apply when adding the concept of costs, but the approach differs. For more details, refer to [40].

3.3.4 Principal Component Analysis

The principal component analysis (PCA), also known as Karhunen-Loève transform, is a key element in a wide range of applications in signal processing and communications. We focus here on its recognition capabilities.

The PCA aims to represent a set $\{x_t\}_{t=1}^T$ of N-dimensional vectors with a —lower-dimensional— single vector x_0. Namely, the goal is to find a vector x_0 that minimizes the sum of square distances between x_0 and all sample points x_t. This can be represented by the following objective function

$$J(x_0) \ = \ \sum_{t=1}^{T} \|x_0 - x_t\|^2$$

The basic PCA approach begins with computing the N-dimensional mean vector μ and the $N \times N$ covariance matrix Σ for the entire data set. Then, the eigenvectors e_i and corresponding eigenvalues λ_i of Σ are computed. The K eigenvectors with the K largest eigenvalues are kept while the smaller ones are discarded. This is done because larger eigenvalues correspond to eigenvectors whose direction represents more of the variability of the original data than the other eigenvectors.

Capturing the variability of the data is essential while the remaining eigenvectors usually represent noise or less significant components of the data. The kept eigenvectors are arranged into the columns of a $N \times K$ matrix S. Finally, the data is projected onto the K-dimensional subspace spanned by these eigenvectors using

$$X \;=\; Px \;=\; S^{\top}(x - \mu).$$

Both the training and testing samples are projected onto this K-dimensional space. Since the label of each training sample is known, it is then easy to assign a class to a test sample by choosing the same label as the closest (in Euclidean distance) training sample. Some variations of PCA are independent component analysis, nonlinear component analysis, and kernel PCA.

3.3.5 Linear Discriminant Functions

Linear discriminant functions are classifiers that do not require prior knowledge of the underlying probability distributions. The general form of a linear discriminant function is

$$g(x) \;=\; w^{\top}x + w_0$$

where x is the feature vector, w is the weight vector and w_0 is the bias. Assuming a two-class problem, we use the decision scheme of choosing class 1 if $g(x) > 0$ and class 2 if $g(x) < 0$. When g(x) is linear, this scheme defines a hyperplane $\mathcal{H}$ that divides the feature space into two decision regions $\mathcal{R}_1$ and $\mathcal{R}_2$. If we have two feature vectors x_1 and x_2 which are both on $\mathcal{H}$, then

$$w^{\top}x_1 + w_0 \;=\; w^{\top}x_2 + w_0,$$

or equivalently

$$w^{\top}(x_1 - x_2) \;=\; 0.$$

This shows that w has to be normal to $\mathcal{H}$. Since $g(x) > 0$ if x is in $\mathcal{R}_1$, then w must point into $\mathcal{R}_1$. Therefore, a linear discriminant function divides a feature space into decision regions with a hyperplane whose orientation and position are determined by w and w_0. However, sometimes the original training data cannot be partitioned with a hyperplane. When this happens, we embed the data onto a higher-dimensional space and find a hyperplane that divides the space into proper regions. The mapping which takes the lower-dimensional feature vectors and maps them to the higher-dimensional space depends on the nature of the data.

3.3.6 Support Vector Machines

Support vector machines (SVMs) are similar to linear discriminant machines but use preprocessing to project data onto a higher-dimensional space using a nonlinear mapping υ. Typically, this dimension is much higher than that of the original feature space. Assuming a two-class problem and a set of T feature vectors $\{x_t\}_{t=1}^{T}$, the goal is to make that data from these two classes separable by a hyperplane in the higher-dimensional space. More specifically, find in the augmented space, the optimal hyperplane with the maximal distance from the nearest training samples

(the support vectors) of all classes. Given a feature vector x_t, it is transformed to y_t using

$$y_t \;=\; \upsilon(x_t).$$

Define a linear discriminant in the augmented space as

$$g(y) \;=\; a^\top y.$$

Next, we also define $r_t = \pm 1$ according to whether the sample x_t is in the first or the second class. Then, the separating hyperplane ensures

$$r_t g(y_t) \geq 1,$$

for all $t = 1, \ldots, T$.

We observe that the solution to choosing a is a region of infinite size but of known boundaries. But since the goal is to find the hyperplane with largest margin, then we can introduce a positive parameter b that represents this margin. In fact, the larger the margin b, the better generalization of the classifier. To find the separating hyperplane, we need to find a unique a that maximizes b using the following:

$$\frac{r_t g(y_t)}{\|a\|} \geq b, \text{ given } b\|a\| = 1,$$

for all $t = 1, \ldots, T$. The support vectors are the transformed training samples for which the equation above is an equality. They are the training samples that define the hyperplane that we seek and are the hardest samples to classify.

SVMs are important classifiers because the complexity of the classifier is not dependent on the dimensionality of the transformed space. Rather, it depends on the number of support vectors. As a result, SVMs are less prone to problems of over-fitting than other methods. Moreover, although the training involves nonlinear optimization, the objective function is convex, hence making the solution relatively straightforward.

3.3.7 Correlation Filters

The idea behind correlation filters is that by analyzing the frequency domain representation of a signal, one can create filters that when correlated with signals, give strong responses for the class they were trained to recognize and very weak responses for all other classes. There are two possible ways of training such filters that give rise to two distinct types of correlation filters. Given a set of training data with multiple sample points for each class, one has two choices. The first is to train the filters using one sample from each class. This type of training results in matched spatial filters. The second is to train the filters using a synthesis of the samples from each class. This type of training creates synthetic discriminant functions . Each approach results in a set of filters that can be used for both recognition and verification, but synthetic discriminant functions are of more interest because they have the potential to accommodate for noise and variation in the data. Once the filters are trained, one can create a simple and effective classifier.

If the filters are designed properly, the correlation plane for a filter from a given class will contain a -strong- peak when correlated with data from this same class. Whereas, when correlated with other classes, it will result in a low-energy correlation plane. Synthetic discriminant functions have given rise to many types of correlation filters which attempt to minimize the effect of noise and variation in the data, maximize the peak of the "true" class, and minimize the overall energy of "impostor" classes. Examples of such correlation filters comprise minimum variance synthetic discriminant functions, maximum average correlation height filters, minimum average correlation energy filters. In Chapter 4, we will see how these filters were used in combination with multiresolution tools for fingerprint recognition.

3.3.8 Neural Networks

In most of our work, we use NN classifiers as they act like highly parallel simple processors that work under the principles of learning and adaptation. NNs are simple to use as well as efficient, and above all, they have the ability to separate classes in a nonlinear fashion. NNs are closely related to linear discriminant functions except they are capable of generating arbitrary decision regions that provide more general solutions. Similarly to linear discriminant functions, NNs use mapping functions to map points in a lower-dimensional space into a higher-dimensional one. However these mapping are nonlinear and can lead to arbitrary decision regions. NNs are classifiers based on grouping the input vectors (features) into intersections of clusters of one type while the union of all such intersections yields the entire feature space.

A NN is composed of an input layer, one or more hidden layers and an output layer. Figure 3.2 shows a simple NN with one of each layer type. Each layer contains a basic unit called the *perceptron*, represented by a node and its associated edges. The input layer has as many nodes as the dimension of the feature vector. The outputs of the input layer can be weighted and/or biased depending on the chosen design. These outputs are weighted, biased, and summed to become the input or *net activation* of the hidden layers according to the following:

$$net_j \;=\; \sum_{n=1}^{N} x_n w_{jn} + w_{j0} \;=\; \langle w_j, x \rangle,$$

where net_j is the net activation at the hidden unit j, N is the number of input units, w_j are the weights at the hidden unit j and the input vector x has been augmented by appending a feature value $x_0 = 1$.

A layer is called hidden when its inputs are not the initial inputs to the NN and its outputs are not the overall output of the NN. The outputs of a hidden layer act as net activation to the subsequent layers. Each hidden unit emits an output *hnet* that is a nonlinear function of the net activation of that unit, namely,

$$\mathrm{hnet}_j \;=\; f(net_j).$$

This nonlinear function can be a simple threshold, a sign function or a tansigmoïd function. The number of hidden layers and the number of nodes in each hidden

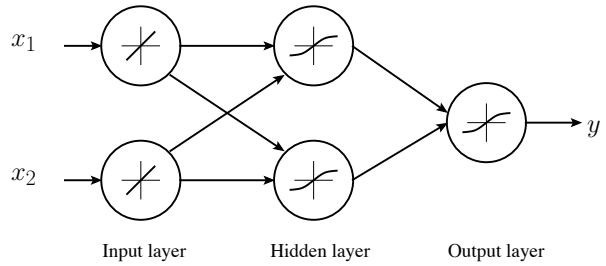


Figure 3.2: A simple neural network with one hidden layer.

layer correspond to the number of arbitrary intersections and arbitrary unions of the feature vectors, respectively. The output is obtained as another multi-dimensional vector that corresponds to the class the input belongs to. A "crushing function" (such as f) is usually applied at the output layer to map the output to a restricted range of values [40]. This can be described as follows:

$$y \;=\; f(net_k),$$

where y is the output and net_k is defined as the net activation function of the output layer (k indexes the units in the output layer) and can be written as

$$net_k \;=\; \sum_{j=1}^{b_h} \mathrm{hnet}_j w_{kj} + w_{k0} \;=\; \langle \mathrm{hnet}, w_k \rangle,$$

with b_h being the number of hidden units, and hnet_j are the outputs of the hidden units (see Fig. 3.2).

The primary difficulty in implementing NNs is that of determining the complexity of the NN. Namely, determining how many hidden layers and nodes within each layer should the NN contain is a complex task. It depends on the training data at hand as well as the desired output. Many algorithms have been developed to build and train NNs, the backpropagation algorithm being the most popular [108].

Chapter 4

Multiresolution Tools

Contents

In the context of the classification of biomedical images, the feasibility of automated interpretation of subcellular patterns in fluorescence microscope images has been clearly demonstrated over the past ten years, initially by Murphy's group [11, 12, 13] and then by others [93, 35, 32]. Their work resulted in systems that can classify protein subcellular location patterns with well-characterized reliability and better sensitivity than human observers (for reviews, see [27, 48]). The heart of such systems is a set of numerical features—Subcellular Location Features—to describe the spatial distribution of proteins in each cell image. These features include Haralick texture features (T_1 with 13 features or T_2 with 26 features), morphological features (16 features), and Zernike moments (49 features). Of particular relevance to the work described here is the use of simple wavelet (30 features) and Gabor (60 features) features, as the addition of these simple MR features resulted in a significant improvement in classification accuracy, with the highest reported accuracy being 91.5% for the 2D HeLa data set [62].

As the introduction of the simplest MR features produced a statistically significant jump in classification accuracy, our hypothesis is that more sophisticated MR techniques would result in even more accurate classification. In particular, the three crucial characteristics of MR [79, 125] we wish to explore are:

- Localization: Fluorescence microscope images have highly localized structures

both in space and frequency. This leads us to MR tools, as they have been found to be the most appropriate tools for computing and isolating such localized structures [78].

- Adaptivity: As we are designing a system to distinguish between classes of proteins, it is clear that an ideal solution should be adaptive, a property provided by MR techniques. The reasoning is that if there is a different MR transform for each individual class, then the transform itself would help in distinguishing that class.
- Fast and Efficient Computation: It is well known that MR techniques such as wavelets have a computational cost of the order $O(N)$ (where N is the input size), as opposed to $O(N \log N)$ typical for other linear transforms including the FFT. This is one of the major reasons for the phenomenal success of MR techniques in real applications and one of the reasons to incorporate MR features into our system.

We now give a brief overview of multiresolution (MR) techniques, which have been extensively studied and used in signal and image processing over the past two decades [125]. MR processing means analysis and processing of data at different resolutions and/or scales. MR transforms decompose a signal into zooming spaces (coarse spaces and many detail spaces called *subbands*) and are implemented by *filter banks (FBs)*, through filtering and sampling.

We first begin by presenting nonredundant MR transforms—bases, as they are the most popular and common MR tools in use. Then, we look at bases implemented by filter banks and review famous nonredundant transforms such as the discrete wavelet transform. We then shift our focus towards redundant MR tools known as frames. Due to their performance in our classification system (see Chapter 6), and their subsequent importance in this work, we review frames in more detail and look at their properties and at some frame families. We finally review the state of the art of MR classification.

4.1 Nonredundant Multiresolution Techniques: Bases

Most of MR techniques in use are nonredundant—the underlying mathematical structures are *bases (MR bases)*.

Assume finite-dimensional spaces $\mathbb{R}^M$ or $\mathbb{C}^M$. Given a basis for such a space, $\Psi = \{\psi_i\}_{i=0}^{M-1}$, we associate to it a matrix (operator) which we will also call Ψ:

$$\Psi \;=\; \begin{pmatrix} \psi_{0,0} & \cdots & \psi_{M-1,0} \\ \vdots & \ddots & \vdots \\ \psi_{0,M-1} & \cdots & \psi_{M-1,M-1} \end{pmatrix}.$$

Matrix Ψ has basis vectors as its columns, and $\psi_{i,j}$ is the jth element of the ith basis vector. Given a pair of biorthogonal bases $(\Psi, \tilde{\Psi})$ dual to each other, a signal x belonging to $\mathbb{R}^M$ or $\mathbb{C}^M$ can be expressed as:

$$x \;=\; \Psi X \;=\; \Psi \tilde{\Psi}^* x, \tag{4.1}$$

where $X = \tilde{\Psi}^* x$ is the vector from $\mathbb{R}^M$ or $\mathbb{C}^M$ of so-called *transform* coefficients (inner products of x with respect to $\{\psi_i\}$), and where $\tilde{\Psi}^*$ denotes the Hermitian transpose of the dual basis $\tilde{\Psi}$.

If the expansion is into an orthonormal basis (ONB), then $\Psi = \tilde{\Psi}$ and the above becomes $\Psi\Psi^* = I$, which further implies that Ψ is a unitary matrix.

4.1.1 Filter-Bank View of Bases

The only infinite-dimensional class of MR decompositions we discuss here are those implemented by FBs, as these are the bases most used in applications and our only link to the real world. The vectors (signals) live in the infinite-dimensional Hilbert space $\ell^2(\mathbb{Z})$. In fact, we can investigate finite-dimensional MR decompositions within the FB framework as well. In other words, all cases we consider, both finite-dimensional and infinite-dimensional, we can look at as FBs. A filter bank is the basic signal processing structure used to implement most MR transforms. Fig. 4.4 depicts a FB with M channels and sampling by N. When $M = N$, we deal with critically-sampled FBs implementing bases. A thorough analysis of FB bases is given in [72].

A FB decomposition can be expressed as in (4.1) where x is now an infinite sequence belonging to $\ell^2(\mathbb{Z})$, X is an infinite sequence of transform coefficients (inner products), and Ψ is the basis expansion matrix given in a setting with finite impulse response (FIR) filters. The matrix Ψ is used in the synthesis FB (the reconstruction step) whereas its dual, $\tilde{\Psi}$ is used in the analysis FB (the decomposition step).

Assume that the nonzero support of the filter ψ_i, or, its length is $L = kM$ (if not, we can always pad with zeros), and write the basis operator as

$$\Psi \;=\; \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \cdots & \Psi_0 & 0 & \cdots & 0 & 0 & \cdots \\ \cdots & \Psi_1 & \Psi_0 & \cdots & 0 & 0 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & \Psi_{k-1} & \Psi_{k-2} & \cdots & \Psi_0 & 0 & \cdots \\ \cdots & 0 & \Psi_{k-1} & \cdots & \Psi_1 & \Psi_0 & \cdots \\ \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{4.2}$$

where each block Ψ_r is of size $M \times M$:

$$\Psi_r \;=\; \begin{pmatrix} \psi_{0,rM} & \cdots & \psi_{M-1,rM} \\ \vdots & \ddots & \vdots \\ \psi_{0,rM+M-1} & \cdots & \psi_{M-1,rM+M-1} \end{pmatrix}. \tag{4.3}$$

We can rephrase the basis decomposition in the z-domain as well using polyphase analysis. A polyphase matrix $\Psi_p(z)$ collects the subsequences modulo N. For bases, $\Psi_p(z)$ is of size $M \times M$ and can be written as:

$$\Psi_p(z) \;=\; \sum_{r=0}^{k-1} \Psi_r z^{-r}, \tag{4.4}$$

where Ψ_r are as defined in (4.3). A paraunitary polyphase matrix (representing an ONB) satisfies

$$\Psi_p(z)\Psi_p^*(z) \;=\; cI, \tag{4.5}$$

where c is a constant.

4.1.2 Block Transforms

When the filter length L is equal to the sampling factor M, we have a *block transform*. Then, in (4.2), only Ψ_0 is nonzero, making Ψ block-diagonal. In effect, since there is no overlap between processed blocks, this can be analyzed as a finite-dimensional case, where both the input and the output are M-dimensional vectors. This discussion shows how finite-dimensional bases can be analyzed in the FB context. Amongst the most famous block transforms used in signal processing are the *Discrete Fourier Transform (DFT)* and the *Discrete Cosine Transform (DCT)*.

The Discrete Fourier Transform The DFT is ubiquitous; however, it is not traditionally looked upon as a signal expansion or written in matrix form. The easiest way to do that is to look at how the reconstruction is obtained:

$$x_k \;=\; \frac{1}{N}\sum_{i=0}^{N-1} X_i W_N^{ik}, \quad k = 0, \ldots, N-1, \tag{4.6}$$

where $W_N = e^{-j2\pi/N}$ is an Nth root of unity. In matrix notation we could write it as

$$x \;=\; \underbrace{\frac{1}{N}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & W_N & \cdots & W_N^{N-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & W_N^{N-1} & \cdots & W_N \end{pmatrix}}_{\Psi = \mathrm{DFT}_N} \underbrace{\begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_{N-1} \end{pmatrix}}_{X}.$$

Note that the DFT matrix defined as above is not normalized, that is $(1/N)(\mathrm{DFT}_N)(\mathrm{DFT}_N)^* = I$. If we normalized the above matrix by $1/\sqrt{N}$, the DFT would exactly implement an orthonormal basis.

The decomposition formula is given as

$$X_i \;=\; \sum_{k=0}^{N-1} x_k W_N^{-ik}, \quad i = 0, \ldots, N-1, \tag{4.7}$$

and, in matrix notation:

$$X \;=\; \mathrm{DFT}_N^* x.$$

Note that in most signal processing texts, the decomposition would be given as $X = \mathrm{DFT}_N x$ and the reconstruction as $x = DFT_N^* X$, instead of the above formulas. Here, to fit our filter bank framework, we switch the roles of the usual analysis and synthesis operators.

Consider now the normalized version. In basis parlance, the basis would be $\Psi = \{\psi_i\}_{i=0}^{N-1}$ where the basis vectors are:

$$\psi_i \;=\; \frac{1}{\sqrt{N}} \left(W_N^0, W_N^i, \ldots, W_N^{i(N-1)} \right)^T, \quad i = 0, \ldots, N-1. \tag{4.8}$$

Then, the expansion formula (4.7) can be seen as

$$X_i \;=\; \langle x, \psi_i \rangle, \qquad i = 0, \ldots, N-1,$$

and the reconstruction formula (4.6) for $x = (x_0, \ldots, x_{N-1})^T$:

$$x \;=\; \sum_{i=0}^{N-1} X_i \psi_i \;=\; \sum_{i=0}^{N-1} \langle x, \psi_i \rangle \psi_i \;=\; \underbrace{\frac{1}{\sqrt{N}} \mathrm{DFT}_N}_{\Psi} \underbrace{\frac{1}{\sqrt{N}} \mathrm{DFT}_N^*}_{\Psi^*} x. \tag{4.9}$$

4.1.3 Lapped Orthogonal Transforms

In practice, the use of block transforms can produce artifacts known as "blocking effects" (since there is no overlap between the basis functions—processed blocks), and thus solutions were sought with longer basis functions. One such solution is the *Lapped Orthogonal Transform (LOT)*. The LOTs can be seen as a class of M-channel FBs implementing bases, originally developed for filters of length $L = 2M$ and later generalized to arbitrary integer multiples of M [82]. Compared to block transforms, the LOT keeps the same number of filters but doubles their length, which means that the basis functions of adjacent blocks overlap by half their size, thus removing the blocking effects. However, LOTs are not solely determined by their length, but by the specific form of their basis vectors as well.

In general, for a FB with filter length $L = 2M$, the time-domain matrix Ψ has a double diagonal, that is, in (4.2), only Ψ_0 and Ψ_1 exist:

$$\Psi \;=\; \begin{pmatrix} \ddots & \vdots & \vdots & \ddots \\ \cdots & \Psi_0 & 0 & \cdots \\ \cdots & \Psi_1 & \Psi_0 & \cdots \\ \cdots & 0 & \Psi_1 & \cdots \\ \ddots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{4.10}$$

Thus, (4.4) reduces to

$$\Psi_p(z) \;=\; \Psi_0 + z^{-1} \Psi_1, \tag{4.11}$$

where $\Psi_r, r = 0, 1$, are $M \times M$ matrices with $(\Psi_r)_{j,i} = \psi_{i,j}$ for $i = 0, \ldots, m-1$ and $j = Mr, \ldots Mr + M - 1$.

Since the LOT is a unitary transform, that is, $\Psi\Psi^* = \Psi^*\Psi = I$ the following must be satisfied:

$$\Psi_0 \Psi_0^* + \Psi_1 \Psi_1^* = \Psi_0^* \Psi_0 + \Psi_1^* \Psi_1 = I, \tag{4.12}$$

$$\Psi_0^* \Psi_1 = \Psi_1^* \Psi_0 = 0,\; \Psi_0 \Psi_1^* = \Psi_1 \Psi_0^* = 0. \tag{4.13}$$

Two main classes of LOTs exist distinguished by whether they use cosines or complex exponentials in their basis functions.

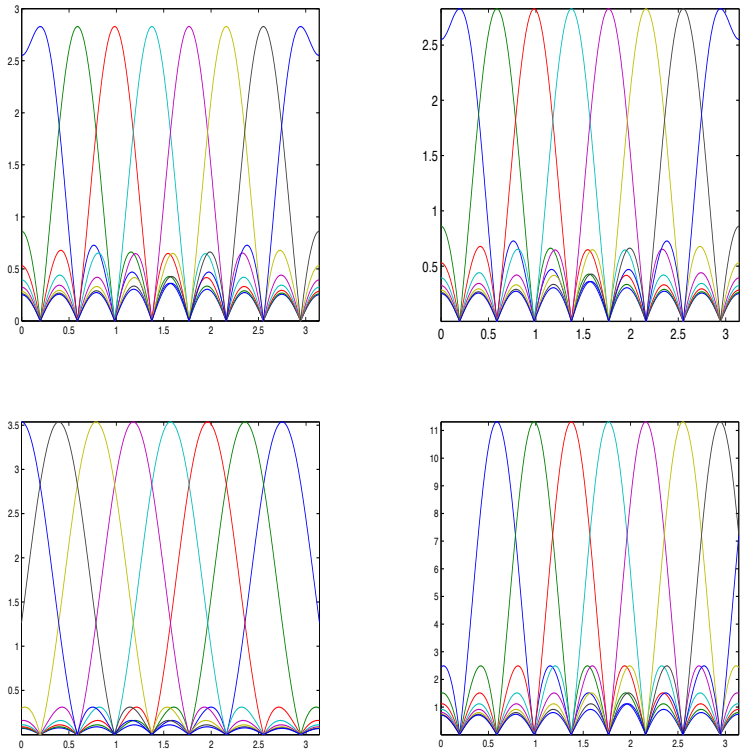


Figure 4.1: Lapped orthogonal transform families with $M = 8$ filters. (a) Princen-Johnson-Bradley, (b) Oddly modulated DCT, (c) Young-Kingsbury, (d) Malvar.

- Cosine Basis Functions

The Cassereau-Malvar LOT The LOT was introduced by Cassereau in [21]. The sole purpose of this transform at that time was image compression, so its basis functions were designed to maximize the transform coding gain. Therefore, the algorithm for the LOT was a recursive design algorithm solving the following optimization problem: $max(\Psi^\top R_{xx}\Psi)$ subject to (4.12) and (4.13), where R_{xx} is the covariance of the input signal x and is chosen to correspond to a first-order Markov model with $\rho = 0.095$. Note that in the optimization algorithm symmetries were forced on the LOT functions. Later on, Malvar [81] proposed a quasioptimal LOT to make the design more robust and allow the existence of a fast algorithm (the design in [21] uses nonlinear optimization steps), starting from the following matrix:

$$\Psi_{DCT} = \frac{1}{2}\begin{pmatrix} D_e - D_o & D_e - D_o \\ J(D_e - D_o) & -J(D_e - D_o) \end{pmatrix},$$

where J is the anti-diagonal matrix and D_e and D_o are the $M \times M/2$ matrices containing the even and odd DCT functions of length M, respectively. Then, to obtain Ψ, the author constructs a unitary matrix $\mathbf{Z}$ such that $\Psi = \Psi_{DCT}\mathbf{Z}$. In fact, the columns of $\mathbf{Z}$ are the eigenvectors of $R_{DCT} = \Psi_{DCT}^{\top} R_{xx} \Psi_{DCT}$.

The Princen-Johnson-Bradley LOT The Princen-Johnson-Bradley (PJB) LOT defined in [97] is an oddly-stacked time domain aliasing cancellation FB. Its basis functions are given by:

$$\psi_{i,k} = \frac{1}{\sqrt{M}} \cos\left(\frac{\pi(2i+1)(2k-M+1)}{4M}\right), \tag{4.14}$$

for $i = 0, \ldots, M-1$ and $k = 0, \ldots, 2M-1$. Thanks to the particular structure of the cosines:

$$\Psi_0 \Psi_0^* = \frac{1}{2}(I - J), \quad \Psi_1 \Psi_1^* = \frac{1}{2}(I + J), \tag{4.15}$$

where J is the anti-diagonal matrix. Fig 4.1 (a) shows the frequency response of the PJB LOT filters for $M = 8$.

With this construction, we will have, similarly to the DFT, fixed basis functions allowing no freedom in design. To allow for better designs, one can add a window that multiplies each filter resulting in a modulated FB over the frequency band. This modulated FB can be modeled as $\Delta\Psi$, where the window $\Delta = \mathrm{diag}\{\delta_k\}_{k=0}^{2M-1}$ is symmetric $\delta_k = \delta_{2M-1-k}, k = 0, \ldots 2M-1$. Now, the perfect reconstruction conditions in (4.12) become

$$\Delta\Psi_0\Psi_0^*\Delta + J\Delta J\Psi_1\Psi_1^* J\Delta J = I. \tag{4.16}$$

Substituting (4.15) into (4.16), we obtain

$$\frac{1}{2}(\Delta^2 + J\Delta^2 J) = I, \tag{4.17}$$

implying that the window has to satisfy $\delta_k^2 + \delta_{M-1-k}^2 = 2$, for $k = 0, \ldots, M-1$.

The Oddly-Modulated Discrete Cosine Transform LOT The oddly-modulated DCT basis functions are very similar to the PJB LOT ones and are defined as:

$$\psi_{i,k} = \frac{1}{\sqrt{M}} \cos\left(\frac{\pi(2i+1)(2k+1+M)}{4M}\right), \tag{4.18}$$

for $i = 0, \ldots, M-1$ and $k = 0, \ldots, 2M-1$. The frequency response of the filters of this family is depicted on Fig 4.1(b) for $M = 8$.

- Complex Exponentials

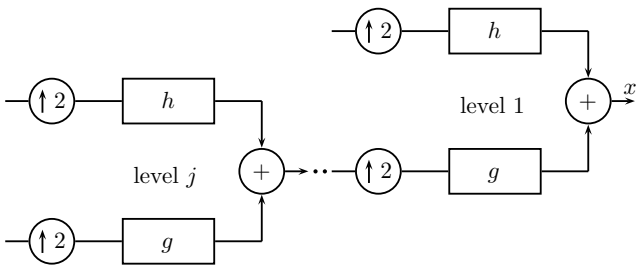


Figure 4.2: The synthesis part of the FB implementing the DWT with j levels. The analysis part is analogous (dual).

Young-Kingsbury LOTs In [131], Young and Kingsbury introduce the complex lapped transform for use in motion estimation applications, defined as:

$$\psi_{i,k} = \frac{1}{\sqrt{M}} \cos(\tfrac{k\pi}{2M}) \exp(\tfrac{-j(2i+1)k\pi}{2M}), \tag{4.19}$$

for $i = 0, \ldots, M-1$ and $k = -\left(M - \frac{1}{2}\right), \ldots, \left(M - \frac{1}{2}\right)$. The frequency response of the filters of this family is depicted on Fig 4.1(c) for $M = 8$.

Malvar Complex LOTs This modulated complex LOT is based on the PJB-LOT FB. Note that it is a complex basis but is redundant in the real space. Its basis functions are defined by cosine and sine modulation of the synthesis windows as follows [83]:

$$\psi_{i,k} = \frac{1}{2}(\psi^c_{i,k} + \alpha\psi^s_{i,k}), \qquad \psi^c_{i,k} = (\sqrt{2})\delta_k \psi^{(OMDCT)}_{i,k}, \tag{4.20}$$

$$\psi^s_{i,k} = \tfrac{\sqrt{2}}{\sqrt{M}}\delta_k \sin(\tfrac{\pi(2i+1)(2k+M+1)}{4M}), \tag{4.21}$$

where $\psi^{(OMDCT)}_{i,k}$ is as defined in (4.18), δ_k are the coefficients of the modulating window, and $i = 0, \ldots, M-1$, $k = 0, \ldots, 2M-1$. Fig. 4.1(d) shows eight frequency responses of the filters of this LOT family. The analysis filters are defined analogously. In [83], the author demonstrates that this LOT is well-suited for noise suppression and echo cancellation.

4.1.4 Discrete Wavelet Transform

The *Discrete Wavelet Transform (DWT)*, a famous MR tool, is a basis expansion and as such nonredundant (critically sampled). The *dyadic* DWT is built by iterating a two-channel FB with sampling factor $N = M = 2$ on the lowpass channel (Fig. 4.2 depicts the synthesis part). Assuming that the filter length $L = 2$, the two analysis filters act on 2 samples at a time and then, due to downsampling by 2,

the same filters act on the following 2 samples. In other words, there is no overlap. Hence in this case, the DWT is a block transform and the most prominent example is the *Haar transform* with synthesis filters

$$G(z) \;=\; \frac{1}{\sqrt{2}}(1+z^{-1}), \qquad H(z) \;=\; \frac{1}{\sqrt{2}}(1-z^{-1}).$$

If we consider the two-level DWT, then using the so-called the Noble identities [121] which allow us to exchange the order of filtering and sampling, we can collect all the filters and samplers along a path into a branch with a single filter and a single sampler. We can then write the filters of this equivalent filter bank as

$$\begin{aligned}
\psi_0(z) &= H(z) = \frac{1}{\sqrt{2}}(1-z^{-1}),\\
\psi_1(z) &= G(z)H(z^2) = \frac{1}{2}(1+z^{-1}-z^{-2}-z^{-3}),\\
\psi_2(z) &= G(z)G(z^2) = \frac{1}{2}(1+z^{-1}+z^{-2}+z^{-3}),
\end{aligned}$$

where the superscript here indicates the number of iterations.

A two-channel filter bank is orthonormal when the lowpass filter is orthogonal satisfying

$$G(z)G(z^{-1}) + G(-z)G(-z^{-1}) \;=\; 2,$$

and the highpass filter is build using

$$H(z) \;=\; -z^{-L+1}G(-z^{-1})$$

where the filter length L is even. The analysis filters are build from the synthesis ones by time-reversal. When the DWT is constructed from an orthonormal filer bank, it implements an orthonormal expansion with

$$x = \Psi X = \sum_{k\in\mathbb{Z}} X_k \psi_k$$

for $x \in \ell^2(\mathbb{Z})$ and $X_k = \langle x, \psi_k\rangle$. In this case, we can write Ψ in terms of the filters g, h as

$$\Psi = \begin{pmatrix}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\dots & g_0 & h_0 & 0 & 0 & 0 & 0 & \dots \\
\dots & g_1 & h_1 & 0 & 0 & 0 & 0 & \dots \\
\dots & g_2 & h_2 & g_0 & h_0 & 0 & 0 & \dots \\
\dots & g_3 & h_3 & g_1 & h_1 & 0 & 0 & \dots \\
\dots & g_4 & h_4 & g_2 & h_2 & g_0 & h_0 & \dots \\
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}$$

where g_{n-2k} and h_{n-2k} are the impulse responses of the synthesis filters g and h shifted by $2k$. Since the columns of Ψ are the basis functions, we have

$$\Psi = \{\psi_k\}_{k\in\mathbb{Z}} = \{\psi_{2k}, \psi_{2k+1}\}_{k\in\mathbb{Z}} = \{g_{\cdot-2k}, h_{\cdot-2k}\}_{k\in\mathbb{Z}},$$

namely, the even-indexed basis functions are the impulse responses of the synthesis lowpass filter and its even shifts, while the odd-indexed basis functions are the impulse responses of the synthesis highpass filter and its even shifts.

When the length of the filters is larger than the sampling factor, then the DWT is not a block transform anymore and a family of filters that is widely used are the *Daubechies filters* [36].

4.1.5 Wavelet Packets

One can also build arbitrary trees by, at each level, iterating on any subset of the branches of the FB. This is typically known as *wavelet packets* [28]. Depending on the length of the filters used, these may, or may not be block transforms or not. To analyze such tree-structured FBs, we typically compute the equivalent filter bank using the Noble identities [121]. The DWT is a particular case of wavelet packets, when only the lowpass branch is decomposed repeatedly to a certain level. For a fixed depth $\mathcal{J}$ of trees, there are as many as $2^{2^{\mathcal{J}}}$ possible wavelet packet trees. Many of these trees do not correspond to real-life signals, but one can build efficient search algorithms allowing a better match to signals at hand. Indeed, wavelet packets have the great advantage of being flexible and adaptive to the signals under consideration. This adaptivity property is enabled by the fact that these trees can be pruned according to some measure or cost function given whether it is better (or "cheaper") to keep a branch or not. For example, Coifman and Wickerhauser [29] present an efficient best basis search algorithm based on a divide-and-conquer strategy that uses additive cost functions. These are termed information cost functions and they measure the concentration or sparsity of the transform coefficients. An information cost function should be large when the coefficients are roughly the same size and small when all but a few coefficients are negligible. The best basis is the one that minimizes the information cost function over all bases in a library. Some examples of cost functions include the entropy, the logarithm of the energy and the norm in ℓ^p for $p < 2$ [129]. The idea of the best-basis algorithm consists of growing the full tree (up to a fixed depth), computing the information cost for all the nodes, then compare the costs of the parents nodes and their children nodes. Whenever a parent node has lower information cost than the children, it is kept, if not, the children nodes are kept. For more details on this algorithm, refer to [129, 29].

4.2 Fingerprint Recognition: Use of Nonredundant MR Bases

Due to the nature of fingerprint images, Hennings et al. [55] used an adaptive wavelet packet approach in combination with correlation filters to solve the recognition problem. The idea was to design a scheme that could adapt itself to the class at hand across multiple scales and resolutions where localized features might be found, clearly calling for the use of wavelet packets. Thus, instead of designing a single correlation filter for a pattern class, a correlation filter was designed for each leaf in the best wavelet packet tree found for that class. The design of the correlation filter was done in the training stage, where the filter was obtained to

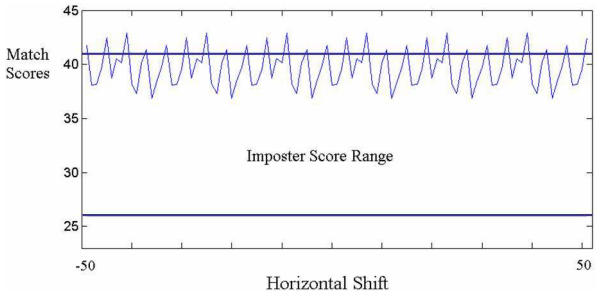


Figure 4.3: Periodic translation invariance of match scores in a fingerprint recognition system (from [55]).

match a few instances of a given class. Finally, if the image belongs to the pattern class of the filter, the correlation plane output contains a sharp peak; if not, no such peak exists. A measure of performance that measures the peak-to-correlation energy, called *match score*, was designed to discriminate between true and impostor classes. A significant improvement in all classes was achieved by using wavelet packets compared to other classical methods that use correlation filters. Indeed, an accuracy of 81.59% was achieved when using standard correlation filters classifiers as opposed to 98.32% with wavelet packets.

4.3 The Need for Redundant Multiresolution Techniques

Although correlation filters are translation invariant in the image intensity domain, they are not translation invariant in the wavelet domain, as the wavelet packets involve downsampling. To examine this effect, the authors in [55] took images from two classes (easy and difficult), translated them horizontally by t, where t ranges from 1 to 50 in each direction. At each pixel translation, they applied the wavelet correlation filters trained for one class in order to compute a peak-to-correlation energy match score. The resulting match score varies periodically with period $2^4 = 16$, as expected. For the easy class, there is still complete separation between authentic scores and the range of impostor scores, despite the translations. When this separation is not as wide, the impostor scores overlap with the match scores, thereby reducing the accuracy of the system (see Fig. 4.3). This clearly calls for the use of translation-invariant transforms. That is, to make the recognition system more robust under these distortions, redundant MR transforms are needed.

4.4 Redundant Multiresolution Techniques: Frames

We start with a brief account of redundant MR techniques called *frames (MR frames)*, first in finite dimensions and then follow up with how signals are represented using frames in infinite dimensions via FBs. An introductory account on frames was written by us [72, 73].

We define frames as follows: A family $\Phi = \{\varphi_i\}_{i \in I}$ in a Hilbert space $\mathbb{H}$ is called a *frame* if there exist two constants $0 < A \leq B < \infty$, such that for all x in

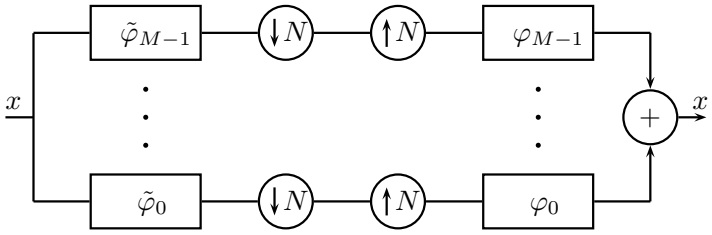


Figure 4.4: An FB implementation of a frame expansion: It is an M-channel FB with sampling by N.

$\mathbb{H}$,

$$A \, \|x\|^2 \;\leq\; \sum_{i \in I} |\langle x, \varphi_i \rangle|^2 \;\leq\; B \, \|x\|^2. \tag{4.22}$$

A, B are called *frame bounds*. The frame bounds are intimately related to the issues of stability. *Tight frames (TF)* are frames with equal frame bounds, that is, $A = B$. *Equal-norm frames* are those frames where all the elements have the same norm, $\|\varphi_i\| = \|\varphi_j\|$, for $i, j \in I$. *Unit-norm frames* are those frames where all the elements have norm 1, $\|\varphi_i\| = 1$, for $i \in I$. By combining this with the requirement of tightness, we can have *equal-norm tight frames*, as well as *unit-norm tight frames*. *A-tight frames* are tight frames with frame bound A. The special case of 1-tight frames are usually called *Parseval tight frames*.

In a finite-dimensional space ($\mathbb{R}^N$ or $\mathbb{C}^N$), a frame is defined as a set Φ of M frame vectors $\Phi = \{\varphi_0, \ldots, \varphi_{M-1}\}$ where M is larger than M. As for bases, we associate to the frame a matrix, also called Φ, that has the frame vectors as its columns:

$$\Phi \;=\; \begin{pmatrix} \varphi_{0,0} & \cdots & \varphi_{M-1,0} \\ \vdots & \ddots & \vdots \\ \varphi_{0,N-1} & \cdots & \varphi_{M-1,N-1} \end{pmatrix}.$$

Note that unlike for bases, Φ is now a rectangular matrix of size $N \times M$.

Similarly to bases, one can check that frames expand signals in $\mathbb{R}^N$ with $x = \Phi X = \Phi\tilde{\Phi}^* x$, where $\tilde{\Phi}$ represents the dual frame. Important operators in frame theory are the *frame operator* defined as $F = \Phi\Phi^*$ and the *Grammian* defined as $G = \Phi^*\Phi$. In matrix parlance, we have a tight frame when $\Phi = \tilde{\Phi}$, and the expansion becomes $\Phi\Phi^* = cI$ (c is a constant).

4.4.1 Filter-Bank View of Frames

In an M-channel FB with sampling factor N, if $M > N$, then we deal with an oversampled FB implementing a frame (see Fig. 4.4).

For a TF, $\tilde{\varphi}_i = \varphi_i$. The FB frame decomposition can be expressed as in (4.1) (substituting Φ for Ψ), where x is an infinite sequence belonging to $\ell^2(\mathbb{Z})$, X is

an infinite sequence of transform coefficients (inner products), and Φ is the frame expansion matrix.

Assuming again that the nonzero support of the filters (frame vectors) length is $L = kN$, we can write the frame operator Φ as in (4.2), with matrices $\Phi_r, r = 0, \ldots, k-1$, being rectangular of size $N \times M$.

We can rephrase the frame decomposition in the z-domain as well, where a FB implements a TF decomposition in $\ell^2(\mathbb{Z})$ *if and only if* its polyphase matrix $\Phi_p(z)$ is paraunitary [34]. For frames, the polyphase matrix $\Phi_p(z)$ is of size $N \times M$ and can be written as in (4.4) (substituting Φ for Ψ), where Φ_r are of size $N \times M$ as in (4.3).

4.4.2 Frame Properties

When designing a frame, particularly if we have a specific application in mind, it is useful to list potential requirements we might impose on our frame [72, 73].

- Tightness: This is a very common requirement and is typically imposed when we do need to reconstruct. Since tight frames (TFs) do not require inversion of matrices, they seem a natural choice. TFs are self dual and they preserve the norm.

- Equal norm: In the real world, the squared norm of a vector is usually associated with power. Thus, in situations where equal-power signals are desirable, equal norm is a must.

- Maximum robustness: We call a frame *maximally robust to erasures*, if every $N \times N$ submatrix of Φ is invertible. This requirement arose in using frames for robust transmission [50].

- Equiangularity: This is a geometrically intuitive requirement. We ask for angles between any two vectors to be the same. There are many more (tight) frames than those which are equiangular, so this leads to a very particular class of frames.

- Symmetry: Symmetries in a frame are typically connected to its geometric configuration. Harmonic and equiangular frames are good examples. See the work of Vale and Waldron [122] for details.

Table 4.1 summarizes all properties of the different classes of frames and writes them in terms of the frame bounds.

4.4.3 Seeding

In an ever-continuing search for new frame families, an appealing option is the process of obtaining TFs from ONBs in larger dimensions, known as the Naimark Theorem [3], or, *seeding* [98]. We give below a finite-dimensional instantiation of the theorem:

THEOREM 4.1 (NAIMARK [3], HAN & LARSON [52]). A set $\Phi = \{\varphi_i\}_{i\in I}$ in a Hilbert space $\mathbb{H}$ is a Parseval tight frame for $\mathbb{H}$ if and only if there is a larger Hilbert space $\mathbb{K}, \mathbb{H} \subset \mathbb{K}$, and an orthonormal basis $\{\psi_i\}_{i\in I}$ for $\mathbb{K}$ so that the orthogonal projection P of $\mathbb{K}$ onto $\mathbb{H}$ satisfies: $P\psi_i = \varphi_i$, for all $i \in I$.

We will use the term seeding when a frame is obtained from a basis and define it as follows:

DEFINITION 4.1. We say that a frame Φ is obtained by seeding from a basis Ψ by deleting a suitable set of columns of Ψ. We write $\Phi^* = \Psi[\mathbb{J}]$ where $\mathbb{J}$ is the index set of the retained columns.

All tight frames can be obtained this way. One of the most famous frame families, the Harmonic Tight Frames (HTFs) is the counterpart of the DFT, that is, HTFs are obtained by seeding the DFT. We can now reinterpret the Parseval tight frame identity $\Phi\Phi^* = I$: It says that the columns of Φ^* are orthonormal. In view of the above theorem, this makes a lot of sense as that frame was obtained by deleting columns from an ONB from a larger space.

In FB parlance, seeding is done on the polyphase matrix. Given $\Psi_p(z)$, the $M \times M$ polyphase matrix associated with a basis of size M, then $\Psi_p(z) = \Psi_0$, and

$$\Phi_p^*(z) = \Phi_0^* = \Psi_p[\mathbb{J}] \tag{4.23}$$

is the transpose of the frame polyphase matrix.

4.4.4 Invariance of Frame Properties

Another way of creating frames is to use the invariance of frame properties. Instead of starting from scratch, a reasonable way of trying to find new families is by constructing new ones from old ones by transformations. However, to do that, we must be sure that the transformation will preserve the properties our old family possesses. It is the aim of this brief discussion to enumerate when this is possible. The results for the nonpolynomial case are derived in [98]; here, we mimic those exactly and thus proofs are omitted. Assume all the matrix products below are compatible and $\Phi_p(z)$ is a frame. Then,

- $U_p(z)\Phi_p(z)V_p(z)$ is a frame for any matrices $U_p(z), V_p(z)$ of full rank (on the unit circle).
- If $\Phi_p(z)$ is tight (unit-norm tight), then $aU_p(z)\Phi_p(z)V_p(z)$ ($U_p(z)\Phi_p(z)V_p(z)$) is tight (unit-norm tight) for any paraunitary matrices $U_p(z), V_p(z)$ and $a \neq 0$.
- If $\Phi_p(z)$ is maximally robust, then $U_p(z)\Phi_p(z)D_p(z)$ is maximally robust for any full rank diagonal matrix $D_p(z)$ and any full rank matrix $U_p(z)$.
- If $\Phi_p(z)$ is unit-norm tight maximally robust, then $U_p(z)\Phi_p(z)D_p(z)$ is unit-norm tight maximally robust for any paraunitary diagonal matrix $D_p(z)$ and any paraunitary matrix $U_p(z)$.

4.4.5 Block Transforms

When $L = N$, that is, the length of the frame vectors equals the sampling factor, we obtain a block transform. One example of block transform with frames is HTFs, mentioned above, that we review briefly.

Harmonic Tight Frames HTFs are obtained by seeding from $\Psi = \mathrm{DFT}_M$ given in (4.7)-(4.9), by deleting the last $(M - N)$ columns:

$$\varphi_i \;=\; \sqrt{\frac{M}{N}}(W_M^0, W_M^i, \ldots, W_M^{i(N-1)}), \tag{4.24}$$

for $i = 0, \ldots, M-1$. Since obtained as an instance of the Naimark Theorem, this is thus a Parseval TF, that is, $\Phi\Phi^* = I$. The simplest example of an HTF is the Mercedes-Benz frame [74].

In [20], the authors define a more general version of the HTF, called general harmonic frames as follows:

$$\varphi_k = \left(c_1^k b_1, c_2^k b_2, \ldots, c_N^k b_N\right),$$

for $k = 0, \ldots, M-1$, with $|c| = 1$, $|b_i| = \frac{1}{\sqrt{M}}$ $(1 \le i \le N)$, and $\{c_i\}_{i=1}^N$ being distinct Mth roots of c. They also show that the HTFs are unique up to a permutation of the orthonormal basis and that every general harmonic frame is unitarily equivalent to a simple variation of an HTF. These frames have been generalized in an exhaustive work by Vale and Waldron [122], where the authors look at frames with symmetries. Some of these they term HTFs (their definition is more general than what is given in (4.24)), and are the result of the operation of a unitary U on a finite Abelian group G. When G is cyclic, the resulting frames are *cyclic*. In [20], the HTFs we showed above are with $U = I$ and generalized HTFs are with $U = D$ diagonal. These are cyclic in the parlance of [122]. An example of a cyclic frame are $(N + 1)$ vertices of a regular simplex in $\mathbb{R}^N$. There exist HTFs which are not cyclic.

4.4.6 Frame Families

As we said earlier, manipulating the parameters of a FB leads to different flavors of MR transforms. We focus here on four classes of frame families as we will use the first three in our classification system whereas the last one is closely related to the novel family of frames we developed in this work (see Chapter 8).

Algorithm à Trous The algorithm à trous is a fast implementation of the dyadic (continuous) wavelet transform. It was first introduced by Holschneider, Kronland-Martinet, Morlet, and Tchamitchian in 1989 [58]. The transform is implemented via a biorthogonal, nondownsampled FB, and is sometimes denoted as *Stationary Wavelet Transform (SWT)*. Its transform is completely redundant in contrast to a completely nonredundant scheme such as the DWT. An example for $j = 2$ levels is given in Fig. 4.5 (this is essentially the same as the 2-level DWT as in Fig. 4.2 with samplers removed).

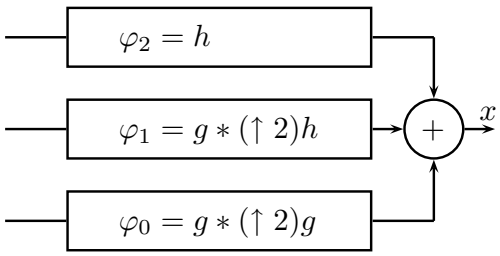


Figure 4.5: The synthesis part of the filter bank implementing the à trous algorithm. The analysis part is analogous. This is equivalent to Fig. 4.2 with sampling removed.

Let φ_0 and φ_1 be the filters used in this two-channel FB. At level $\mathcal{J}$ we will have equivalent upsampling by 2^j, which means that the filter moved across the upsampler will be upsampled by 2^j, inserting $(2^j - 1)$ zeros between every two samples and thus creating holes ("trou" means "hole" in French).

The bottom plot in Fig. 4.6 shows the sampling grid for the à trous algorithm. It is clear from the figure, that this scheme is completely redundant, as all the points exist. This is in contrast to a completely nonredundant scheme such as the DWT, given in the top plot of the figure. In fact, while the redundancy per level of this algorithm grows exponentially since $A_1 = 2, A_2 = 4, \ldots, A_j = 2^j, \ldots$, the total redundancy for j levels is linear, as $A = A_j 2^{-j} + \sum_{i=1}^{j} A_i 2^{-i} = (j + 1)$. Note that here we use a two-channel filter bank and that A_j is the frame bound when we use j levels. This growing redundancy is the price we pay for shift invariance as well as the simplicity of the algorithm. The 2D version of the algorithm is obtained by extending the 1D version in a separable manner.

The Dual-Tree Complex Wavelet Transform This transform was first introduced by Kingsbury in 1998 [69, 70, 71]. The basic idea behind it is to have two DWT trees working in parallel representing the real and complex parts of a complex transform. One tree represents the real part of the complex transform while the second tree represents the imaginary part. That is, when the dual-tree complex wavelet transform is applied to a real signal, the output of the first tree is the real part of the complex transform whereas the output of the second tree is its imaginary part. Each tree uses a different pair of lowpass and highpass filters designed to satisfy the perfect reconstruction condition (4.1).

Because the two DWT trees used in the dual-tree complex wavelet transform are fully downsampled, the redundancy of the transform is only 2 for the 1D case (it is 2^d for the d-dimensional case). Unlike the à trous algorithm, however, here the redundancy is independent of the number of levels used in the transform.

In 2D (or MD), the dual-tree complex wavelet transform possesses directional

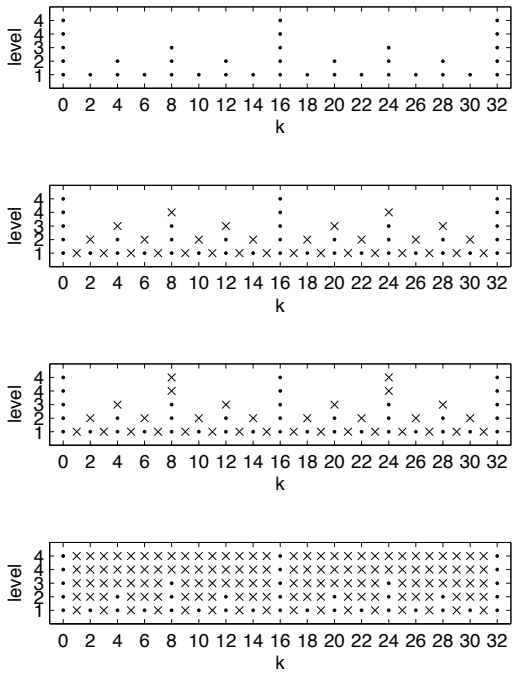


Figure 4.6: Sampling grids corresponding to time-frequency tilings of (top to bottom): DWT (nonredundant), double-density DWT, dual-tree complex wavelet transform, à trous family (completely redundant). Black dots correspond to the nonredundant (DWT-like) sampling grid. Crosses denote redundant points. Note that the last two ticks on the y-axis represent level 4 for the highpass and lowpass channels, respectively.

selectivity allowing us to capture edge or curve information, a property clearly absent from the usual separable DWT. In the real case, orientation selectivity is simply achieved by using two real separable 2D DWTs in parallel. Two pairs of filters are used to implement each DWT. These two transforms produce six subbands, three pairs of subbands from the same space-frequency region. By taking the sums and differences of each pair, one obtains the oriented wavelet transform. The near translation invariance and orientation selectivity properties of the dual-tree complex wavelet transform open up a window into a wide range of applications, among them denoising, motion estimation, image segmentation as well as building feature, texture and object detectors for images [113].

Double-Density Discrete Wavelet Transform Selesnick in [111] introduces the *double-density DWT*, which can approximately be implemented using a three-channel FB with sampling by two. The filters in the analysis bank are time-reversed versions of those in the synthesis bank. The redundancy of this FB tends towards 2 when iterated on the first channel. Actually, we have that $A_1 = \frac{3}{2}, A_2 = \frac{7}{4}, \ldots A_\infty = 2$,

where A_j is the redundancy at level $\mathcal{J}$. Like the dual-tree complex wavelet transform, the double-density DWT is nearly translation invariant when compared to the à trous construction. In [112], Selesnick introduces the combination of the double-density DWT and the dual-tree complex wavelet transform which he calls *double-density, dual-tree complex wavelet transform*. This transform can be seen as a dual-tree complex wavelet transform, with individual FBs being overcomplete ones (three channels, downsampling factor of two).

Gabor and Cosine-Modulated Frames The idea behind this class of frames, consisting of many families, dates back to Gabor [45] and the insight of constructing bases by modulation of a single prototype function. Gabor originally used complex modulation, and thus, all those families with complex modulation are termed *Gabor frames*. Other types of modulation are possible, such as cosine modulation, and again, all those families with cosine modulation are termed *cosine-modulated frames* (also often called Wilson bases). Both of these classes can be seen as general oversampled filter banks with m channels and sampling by n (see Fig. 4.4).

Gabor Frames. A Gabor frame is $\Phi = \{\varphi_i\}_{i=0}^{M-1}$, with

$$\varphi_{i,k} \;=\; W_M^{-ik}\varphi_{0,k}, \tag{4.25}$$

where index $i = 0, \ldots, M-1$ refers to the number of frame elements, $k \in \mathbb{Z}$ is the discrete-time index, W_M is the Mth root of unity and φ_0 is the prototype frame function. Comparing (4.25) with (4.24), we see that for filter length $L = N$ and $\varphi_{0,k} = 1, k = 0$ and 0 otherwise, the Gabor system is equivalent to a HTF frame. Thus, it is sometimes called the *oversampled DFT frame*.

For the critically-sampled case it is known that one cannot have Gabor bases with good time and frequency localization at the same time (this is similar in spirit to the Balian-Low theorem which holds for $\mathcal{L}^2(\mathbb{R})$ [37]); this prompted the development of oversampled (redundant) Gabor systems (frames). They are known under various names: *oversampled DFT FBs*, *complex-modulated FBs*, *short-time Fourier FBs* and *Gabor FBs* and have been studied in [33, 16, 15, 14, 42] (see also [118] and references within).

Cosine-Modulated Frames. This kind of modulation was used with great success within critically-sampled filter banks due to efficient implementation algorithms. Its oversampled version was introduced in [15], where the authors define the frame elements as:

$$\varphi_{i,k} \;=\; \sqrt{2}\cos\left(\frac{(i+1/2)\pi}{M} + \alpha_i\right)\varphi_{0,k}, \tag{4.26}$$

where index $i = 0, \ldots, M-1$ refers to the number of frame elements, $k \in \mathbb{Z}$ is the discrete-time index and φ_0 is the template frame function. Equation (4.26) defines the so-called *odd-stacked cosine modulated FBs*; even-stacked ones exist as well.

Cosine-modulated filter banks do not suffer from time-frequency localization problems, given by a general result stating that the generating window of an orthogonal cosine modulated FB can be obtained by constructing a tight complex filter bank with oversampling factor 2 while making sure the window function satisfies a certain symmetry property (for more details, see [15]).

4.5 Relevant Work on Multiresolution Classification

The idea of using multiresolution techniques for classification has been widely adopted, specifically in the texture research community, where there exists extensive literature on the subject. We mention below a few contributions. MR tools are essentially used as powerful feature extractors. Most of the approaches for MR classification may be divided in two categories: The first, which after representing the signals in a chosen MR representation, either uses the transform coefficients themselves (or a subset of them) as features or computes some local energies or statistical measures on these coefficients; The second chooses first which MR representation is the best in terms of some discriminative power metric and then uses the transform coefficients of this representation in the same fashion as the first approach.

Approach 1 We now give a few examples of the first approach. Lane and Fan [76] used both the standard wavelet and the wavelet packet representations to compute energy and entropy measures. These were used as texture features and input into a neural network classifier. Unser [120] used a similar idea with wavelet frames where he computed the energies of the subspaces of the discrete wavelet frame transform and used them as features to classify Brodatz textures using a Bayes classifier. Wavelet frames were also used in [124] for texture classification. Skretting and Husøy [115] proposed a frame texture classification method based on a deterministic texture model in which a small texture image block is modeled as a sparse linear combination of frame elements. For each texture class, a set of frame vectors are designed such that they give the sparsest representation of the class. In the test phase, the frame giving the best (sparsest) representation yields the class. In [100], the authors present a texture classification method based on curve fitting of wavelet domain singular values and probabilistic neural networks. Here, the feature vectors are based on singular values computed on local energies of wavelet packet subspaces. Probabilistic neural networks are a network formulation of probability density estimation and the classifier used in [100] is a weighted probabilistic neural network. Other MR-based texture classification methods include [68, 4, 5, 1, 51, 123, 22, 126, 91], and a complete list of articles can be found on [96].

Approach 2 For the second approach, the most noticeable contribution came from Saito and Coifman. In [106], they extended the best basis method in [29] for use in classification. They developed the so-called local discriminant basis algorithm that selects out of a dictionary of orthonormal bases (wavelet packets or lapped orthogonal transforms [82]) the "best" subspaces, namely the ones with the most discriminative power amongst classes. This metric is determined by computing the time-frequency energy distributions of each class. The subspaces are selected when they well separate these distributions according to a distance such as the Kullback-Leibler divergence [10]. Once the subspaces have been selected, to form the best basis, the transform coefficients are fed into a linear discriminant analysis classifier or a classification tree. In the testing phase, the corresponding coefficients are fed to the classifier to predict the class of the test signals. Later on, Saito et. al proposed

in [107] an improvement of their previous method by using estimates on empirical probability densities in the subspaces instead of energy distribution. In [60], we find a work of similar spirit, where the authors propose a best-basis type of algorithm that takes into account the existing correlations between the subspaces and bases the selection of these subspaces on a mutual information measure. Recently, the same authors presented in [61] a theoretical framework for signal classification with sparse representation (using frames) that achieves a sparse and robust representation of corrupted signals for effective classification. This is done by introducing in the objective function for sparse representation a fisher discrimination power term, a sparsity term and a reconstruction error term.

Outside of the texture classification literature, MR tools have been used on a wide range of classification tasks. In [110], the authors present a wavelet-based framework called TEMPLAR for recovering a pattern from a collection of noisy and misaligned observations. The method is iterative and combines the approximation capabilities of wavelets (a discrete wavelet transform) to minimum description length complexity-regularization to learn a low-dimensional template from the training data. In a pattern classification context, this method can be applied to produce a template of each class (pattern). Then, the resulting pattern models are used for likelihood-based classification. In [102], Richardson applied wavelets to the classification of mammograms and in [38], the authors look at a lung classification problem and use wavelet frames combined with grey-level histogram features along with a k-nearest neighbor classifier.

4.6 Towards Adaptive Multiresolution Classification

In the work we present in the next part of this thesis (Chapter 5), we do not follow either of the above approaches; rather, we perform MR decomposition and then classify within each subspace separately, and unlike the works above, we do not use subspace coefficients as features.

Indeed, in our work, we investigate five different biomedical and biometric applications (Chapter 2) that have classification as their underlying task. We develop an adaptive supervised classification algorithm based on MR techniques, aiming to extract discriminative features within space-frequency localized MR subspaces. These are obtained by MR decomposition; that is, rather than add MR features to existing features, we instead choose to compute these features in the MR-decomposed subspaces themselves. Thus, our system has an upfront MR decomposition block which is followed by feature computation and classification in each of the MR subspaces, which, in turn, are then combined through an adaptive weighting process. We present the details of this algorithm in the next chapter and its performance on the applications we consider in this work in Chapter 6.

Frame	Constraints	Properties
General	$\{\varphi_i\}_{i\in I}$	$A\|x\|^2 \leq \sum_{i\in I} \|\langle\varphi_i, x\rangle\|^2 \leq B\|x\|^2$
	is a Riesz basis for $\mathbb{H}$	$AI \leq F \leq BI$
		$\mathrm{tr}(F) = \sum_{k=1}^{N} \lambda_k = \mathrm{tr}(G) = \sum_{i=1}^{M} \|\varphi_i\|^2$
ENF	$\|\varphi_i\| = \|\varphi_k\| = a$	$A\|x\|^2 \leq \sum_{i\in I} \|\langle\varphi_i, x\rangle\|^2 \leq B\|x\|^2$
	for all i and k	$AI \leq F \leq BI$
		$\mathrm{tr}(F) = \sum_{k=1}^{N} \lambda_k = \mathrm{tr}(G) = \sum_{i=1}^{M} \|\varphi_i\|^2 = Ma^2$
TF	$A = B$	$\sum_{i\in I} \|\langle\varphi_i, x\rangle\|^2 = A\|x\|^2$
		$F = AI$
		$\mathrm{tr}(F) = \sum_{k=1}^{N} \lambda_k = NA = \mathrm{tr}(G) = \sum_{i=1}^{M} \|\varphi_i\|^2$
PTF	$A = B = 1$	$\sum_{i\in I} \|\langle\varphi_i, x\rangle\|^2 = \|x\|^2$
		$F = I$
		$\mathrm{tr}(F) = \sum_{k=1}^{N} \lambda_k = N = \mathrm{tr}(G) = \sum_{i=1}^{M} \|\varphi_i\|^2$
ENTF	$A = B$	$\sum_{i\in I} \|\langle\varphi_i, x\rangle\|^2 = A\|x\|^2$
	$\|\varphi_i\| = \|\varphi_k\| = a$	$F = AI$
	for all i and k	$\mathrm{tr}(F) = \sum_{k=1}^{N} \lambda_k = NA = \mathrm{tr}(G) = \sum_{i=1}^{M} \|\varphi_i\|^2 = Ma^2$
		$A = (m/n)a^2$
UNTF	$A = B$	$\sum_{i\in I} \|\langle\varphi_i, x\rangle\|^2 = A\|x\|^2$
	$\|\varphi_i\| = 1$	$F = AI$
	for all i	$\mathrm{tr}(F) = \sum_{k=1}^{N} \lambda_k = NA = \mathrm{tr}(G) = \sum_{i=1}^{M} \|\varphi_i\|^2 = M$
		$A = M/N$
ENPTF	$A = B = 1$	$\sum_{i\in I} \|\langle\varphi_i, x\rangle\|^2 = \|x\|^2$
	$\|\varphi_i\| = \|\varphi_k\| = a$	$F = I$
	for all i and k	$\mathrm{tr}(F) = \sum_{j=1}^{N} \lambda_k = N = \mathrm{tr}(G) = \sum_{i=1}^{M} \|\varphi_i\|^2 = Ma^2$
		$a = \sqrt{N/M}$
UNPTF	$A = B = 1$	$\sum_{i\in I} \|\langle\varphi_i, x\rangle\|^2 = \|x\|^2$
$\Leftrightarrow$	$\|\varphi_i\| = 1$	$F = I$
ONB	for all i	$\mathrm{tr}(F) = \sum_{k=1}^{N} \lambda_k = N = \mathrm{tr}(G) = \sum_{i=1}^{M} \|\varphi_i\|^2 = M$
		$N = M$

Table 4.1: Summary of properties for various classes of frames. All trace identities are given for $\mathbb{H} = \mathbb{R}^N, \mathbb{C}^N$. ENF = Equal-norm frame, TF = tight frame, PTF = Parseval tight frame, ENTF = Equal-norm tight frame, UNTF = Unit-norm tight frame, ENPTF = Equal-norm Parseval tight frame, UNPTF = Unit-norm Parseval tight frame, ONB = Orthonormal basis

Part III

Algorithm and Applications

Chapter 5

Multiresolution Classification Algorithm

Contents

In Chapter 2, we saw that the classification problem is ubiquitous in biomedical imaging, and that MR techniques might make classification more accurate. The results obtained in [55] seem to indicate that adaptive MR techniques, frames in particular, might be needed.

Having motivated the use of adaptive MR in classification as well as the need for redundant MR transforms, we now test that hypothesis. In this chapter, we describe the adaptive MR classification algorithm we developed and detail each step involved. In the next chapter, we proceed to present the performance of this algorithm in the five application domains we consider.

We now describe the adaptive MR classification algorithm we developed, based on our previous discussion in Chapter 4 on why MR is needed. While we have developed the current algorithm by learning from each application as we went along, we decided to first present all algorithmic accomplishments and then discuss results in various application domains in Chapter 6. Our results using this algorithm in various application fields are described in [116, 86, 84, 117, 23, 66, 24].

5.1 Main Idea

As argued in Chapter 4, we would like to extract discriminative features within space-frequency localized subspaces. These are obtained by MR decomposition; that is, instead of adding MR features as in [62], we compute features in the MR-decomposed subspaces.

Our initial idea was to use wavelet packets since they adapt themselves to the signal at hand, and just as in the fingerprint case, prove that adaptivity significantly

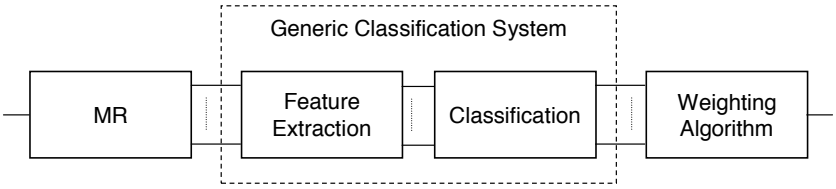


Figure 5.1: Our proposed adaptive MR classification system.

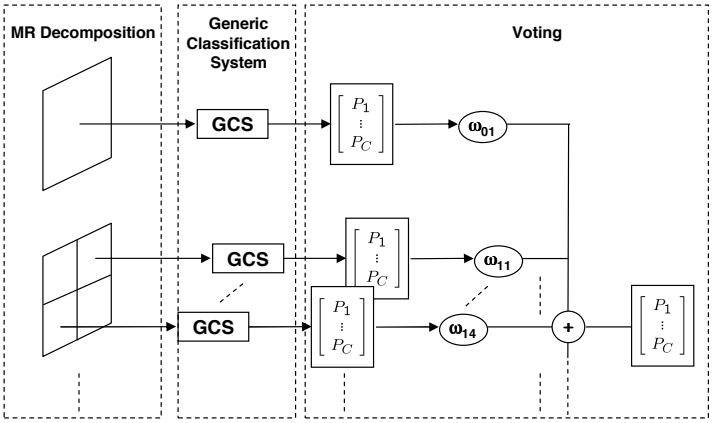


Figure 5.2: Detailed view of our proposed adaptive MR classification system.

improves the recognition system. So, ideally, we would characterize each class by the best wavelet packet tree that represents it. However, this is possible only if a suitable cost function can be found. Given that we have no natural cost function available, we decided to mimic a wavelet-packet like system by adding a weighting procedure at the end of our system, allowing us to weigh the decisions of each subband in a fully grown tree. This way, a very low weight emulates a pruned branch in the tree. Thus, we propose a system with an MR decomposition block in front (see Fig. 5.1), followed by feature computation and classification in each of the subspaces, which are then combined through a weighting process, providing adaptivity.

5.2 Multiresolution Block

In Sections 4.1,4.4, we have seen that MR transforms are many and the adaptivity of MR transforms manifests itself in many guises, including a number of popular transforms:

1. Growing a full tree to $\mathcal{J}$ levels with specific filters of the same length as the downsampling factor yields the DFT of size $2^{\mathcal{J}}$.

2. Growing a full tree to $\mathcal{J}$ levels but allowing the filters to be longer, leads to the short-time Fourier transform, or, the Gabor transform.

3. Growing the tree only on the lowpass branch to $\mathcal{J}$ levels leads to the $\mathcal{J}$-level DWT.

4. Growing an arbitrary tree leads to wavelet packets.

5. Splitting the signal into more than two channels, allowing filters in the above transforms to be orthogonal and/or linear phase, allowing for true multidimensional filters and/or samplers, etc., leads to even more degrees of freedom.

In our classification system, any MR transform can be used. In particular, amongst the MR bases, we used the DWT, DFT, discrete cosine transform and others, while amongst the MR frames, we used the double-density DWT, dual-tree complex wavelet transform and the stationary wavelet transform (that implements the algorithm à trous). (Note that here, we use all the subbands of the decomposition tree, not only the leaves. Thus, it might be abuse of language to call a transform a DWT.) For example, for 2 levels, we have a total of $S = 21$ subbands (original image + 4 subbands at the first level + 16 subbands at the second level).

5.3 Feature Extraction and Classifier

We start with the feature sets used in [62]: Haralick texture features (Haralick set T_1, 13 features), morphological (16 features) and Zernike moments (49 features). Unlike in [62], we do not use wavelet/Gabor features because the MR advantage given by these will be achieved by our MR decomposition. Therefore, our total number of features is 78, as opposed to 174 in [62].

Instead of combining all features into a single feature vector, we allow each feature set its own feature vector per subband. For example, for two levels of decomposition, this effectively brings the number of subbands to $3 \cdot S = 63$ when using all three feature sets. Note that although we have decreased the number of features significantly, we have also increased the number of classifiers, because we now have one classifier per subband. Evaluating this computational trade-off is a task for future work.

For the classifier block, we originally started by using a maximum likelihood classifier. This classifier proved that MR subspaces do indeed contain discriminatory information that improves the classification accuracy of the system. However, this type of classifier uses strong assumptions on the data, namely, a probabilistic modeling of the feature vectors. In practice, this model can be quite far from reality and might be misleading. Therefore, we chose to use a different model for the classifier based on learning and adaptivity: neural networks. All results presented in Chapter 6 are based on the use of NN as the core classifier [40].

5.3.1 New Texture Feature Set

As we will show later on, the Haralick texture features seem to possess the most discriminative power, so we looked more closely into these. We changed the way

that Haralick combines the initial four sets of features defined in (3.2). We note that P_H and P_V are fundamentally different from P_{LD} and P_{HD} because adjacent neighboring pixels are spatially closer than diagonal neighboring pixels. Therefore, instead of averaging the features from all four sets, we create our first set of 13 features by averaging $f_{(H,i)}$ and $f_{(V,i)}$, and a second set of 13 features by averaging $f_{(LD,i)}$ and $f_{(RD,i)}$. Thus, we end up with two sets of 13 features, which are concatenated into a new feature set, denoted T_3, of 26 features:

$$f_i^{(T_3)} = \frac{f_{H,i} + f_{V,i}}{2}, \qquad f_{i+13}^{(T_3)} = \frac{f_{LD,i} + f_{RD,i}}{2}, \qquad i = 1, \cdots, 13.$$

5.3.2 K-means and Gaussian Modeling

In our initial efforts to design a classification system for the recognition of protein subcellular location patterns presented in Section 2.1, we developed a maximum likelihood type of classifier. This fits within our generic framework shown in Fig 5.1. The hypothesis behind this classifier is that the feature vectors from each class form clusters in the feature space. Note that there are S feature spaces, one for each MR subspace. For each feature space, we model each cluster with a Gaussian distribution. Then, we use a maximum likelihood rule to assign class labels. These labels are "'local" to each of the feature spaces (or subbands) involved. Therefore, to make a final or global decision, we add a weighting procedure that acts as a mediator to make everyone agree on the final class label (see Section 5.4 for a detailed description of this process). More specifically, after computing Haralick texture features (feature set T_1), we use K-means clustering algorithm in each of the S feature spaces. This allows us to form at most K clusters for each class. We then model each cluster by a multi-variate Gaussian distribution using the training set. Note that the mean associated with each distribution is the mean or center of each cluster. As a result, in each feature space, every feature vector of every class is now represented by a single probability vector. The ith element of this vector is the probability that the feature vector in question belongs to class i. This probabilities are used to train the weighting procedure (we used the open procedure described in Section 5.4) to output a final weight vector for this system. During the testing phase, the Gaussian models obtained in the training phase are used to compute the probability vectors of a test image. These vectors are then weighed with the weighting vector (from the training phase). Finally, the class label with the maximum likelihood is assigned to the image.

We used this system on the protein subcellular location data set presented in Section 2.1, and demonstrated that, by adding an MR block in front, we were able to raise the classification accuracy by roughly 10% (from 71.8% to 82.2%) as compared to the system with no MR. We concluded that selecting features in MR subspaces allows us to custom-build discriminative feature sets. However, although the MR block substantially increased classification accuracy, the accuracy of the overall system was still not high enough, and thus, in our subsequent work, we reexamined the classification and weighting steps of the system. We present the ensued work in the next two sections.

5.3.3 Neural Networks

We decided to use a two-layer NN classifier. The first layer contains a node for each of the input features, each node using the Tan-Sigmoid transfer function. The second layer contains a node for each output and uses a linear transfer function (no hidden layers are used). We then train the NN using a one-hot design, that is, since each output from the second layer corresponds to a class, when training, each training image will have an output of 1 for the class of which it is a member and a 0 for all other classes. To maximize the use of our data, our training process of the NN block uses five-fold cross validation.

5.4 Weighting Procedure

Fig. 5.2 shows a detailed graphical representation of our MR classification system, including the process of combining all of the subband decisions into one. We use weights for each subband to adjust the importance that a particular subband has on the overall decision made by the classification system. If the weights are chosen such that the no-decomposition weight is equal to 1, and all other weights are 0, we will achieve the same output vector as we would have without using the adaptive MR system. Therefore, we know that there exists a weight combination that will do at least as well as the generic classifier (when no MR is involved) in the training phase. Our goal is to decide how to find the weight vector that achieves the highest overall classification accuracy on a given data set. We developed three versions of the weighting algorithm: open-form, per-dataset closed-form and per class closed-form. The per-dataset algorithm assigns one weight vector for the entire data set, whereas the per-class one assigns a weight vector for each class in the data set. The latter goes back to our original idea of having a wavelet packet tree characterizing each class, only in this case, we do not necessarily obtain a tree.

The difference between the open- and closed-form algorithms is that in the open-form version (see Algorithms 1 and 2), we optimize classification accuracy on the training set as opposed to the closed-form where we look for the least-squares solution.

The NN block outputs a series of decision vectors for each subband of each training image. Each decision vector $d_s^{(r)}$ contains C numbers (where C denotes number of classes) that correspond to the "local" decisions made by the subband s for a specific image r.

The classifier is evaluated using nested cross validations (we chose to use five-fold cross validation in the NN block and ten-fold during the weighting process, but one can use different numbers). One problem with this technique is that the initial ordering of the images determines which images are grouped together for training and testing in each fold of the cross validation. A different original ordering of the images would result in different groupings, which would be equivalent to presenting different data sets to the classifier, and would thus result in a different overall result. We solve this problem by running multiple trials, each with a random initial ordering of the images.

5.4.1 Open-Form Algorithm

If using the open-form algorithm, we initialize all the weights (see Algorithm 1), and a global decision vector is computed using a weighted sum of the local decisions. An initial class label will be given to an image using this global decision vector. If that class label is correct, we go to the next image. If it is incorrect, we look at the local decisions of each subband and adjust the weights of each subband s as follows:

$$w_s^{iter} = \begin{cases} w_s^{iter-1} \cdot (1+\epsilon) & \text{if subband } s \text{ is correct,} \\ w_s^{iter-1} \cdot (1-\epsilon) & \text{otherwise,} \end{cases}$$

where *iter* is the iteration number and ϵ is a small positive constant. This can be viewed as a reward/punishment method where the subbands taking the correct decisions will have their weights increased, and those taking wrong decisions will have their weights decreased. We continue cycling through the images until there is no increase in classification accuracy on the training set for a given number of iterations.

5.4.2 Per-Dataset Closed-Form Algorithm

The closed-form solution does not use an iterative algorithm; rather, it finds the weight vector by solving a minimization problem in the least-square sense.

Assume we have R training images. For each training image $r = 1, \ldots, R$, the vector $d_s^{(r)} = (d_{s,c}^{(r)})^T$ for $c = 1, \ldots, C$, is the $C \times 1$ decision vector at the output of each subband classifier s, where $d_{s,c}^{(r)}$ indicates the confidence of subband s that the training image r belongs to class c. For each training image r, the weighting block takes as input the subband (local) decision vectors $d_s^{(r)}$ and combines them into a single output decision vector as follows:

$$\sum_{s=1}^{S} w_s d_s^{(r)}. \tag{5.1}$$

We can rewrite the above by, for each training image r, forming a matrix $D^{(r)}$ of size $C \times S$, where each element $D_{c,s}^{(r)}$ is the value at position c of the decision vector $d_s^{(r)}$ of subband classifier s. We can now compute:

$$D^{(r)} w,$$

where $w = (w_1, \ldots, w_S)^T$ is of size $S \times 1$. Thus, we want to find a weight vector w common to all training images $r = 1, \ldots, R$. A possible solution for w is the one that minimizes the error in the least-square sense:

$$w_{win} \;=\; \arg\min_{w} \sum_{r=1}^{R} \|d^{(r)} - D^{(r)} w\|^2, \tag{5.2}$$

where $d^{(r)}$ is the desired target decision vector of size $C \times 1$, with a 1 in the position of the true class, and 0 elsewhere.

Algorithm 1 Classification Training Phase

Input: d_s (local decision vector).
Output: w (weight vector).
TrainingPhase(d_s)
  initialize $w_{1,s}$ to classification accuracy of subband s
  initialize $w_{2,s}$ to emphasize the decisions of the 0th subband
  initialize $w_{3,s}$ to positive random entries
  for all weight vectors, $i = 1$ to 3 **do**
    normalize w_i, initialize counter, $cnt = 0$
    compute classification accuracy with w_i and store in p_i
    while $cnt <$ maxEpochs **do**
      increment counter, $cnt = cnt + 1$
      for all images, $r = 1$ to R **do**
        set g_{dec} = the image is classified correctly with w_i
        if g_{dec} is false **then**
          for all subbands, $s = 0$ to $S - 1$ **do**
            set $l_{dec}^{(s)}$ = subband classified correctly
            if $l_{dec}^{(s)}$ is true **then**
              $w_{i,s} = w_{i,s} \cdot (1 + \epsilon)$
            else
              $w_{i,s} = w_{i,s} \cdot (1 - \epsilon)$
            end if
          end for
        end if
      end for
      compute classification accuracy with w_i and store in p_i^{new}
      if $p_i^{new} > p_i$ **then**
        set $p_i = p_i^{new}$, save w_i as w_i^{best}, reset counter, $cnt = 0$
      end if
    end while
  end for
  set w to w_i^{best} with the greatest p_i
  return w

We need to rewrite the above in a direct error-minimization form. We thus define a target output vector d of size $CR \times 1$, as a vector which concatenates all the target decision vectors $d^{(r)}$ as follows:

$$d = \left(\underbrace{d_1^{(1)}, d_2^{(1)}, \ldots, d_C^{(1)}}_{\text{training image 1}}, \ldots, \underbrace{d_1^{(R)}, \ldots, d_C^{(R)}}_{\text{training image } R} \right)^T ,$$

and a $CR \times S$ matrix D consisting of the all the decision matrices $D^{(r)}$ of all the

Algorithm 2 Testing Phase

Input: d_s (local decision vector), w(weight vector).
Output: g (global decision vector), ACC (classification accuracy).
TestingPhase(d_s, w)
  set $g = \sum_{s=1}^{S} w_s \cdot d_s$
  set ACC equal to the classification accuracy of g
  return g and ACC

training data stacked on top of each other:

$$D = \begin{pmatrix} D_{1,1}^{(1)} & \dots & D_{1,S}^{(1)} \\ \vdots & \ddots & \vdots \\ D_{C,1}^{(1)} & \dots & D_{C,S}^{(1)} \\ \vdots & \ddots & \vdots \\ D_{1,1}^{(R)} & \dots & D_{1,S}^{(R)} \\ \vdots & \ddots & \vdots \\ D_{C,1}^{(R)} & \dots & D_{C,S}^{(R)} \end{pmatrix}.$$

We can now rewrite (5.2) as:

$$w_{win} = \arg\min_{w} \|d - Dw\|, \tag{5.3}$$

which possesses a closed-form solution and can be computed efficiently.

Then, for a testing image t, we compute its decision vector $\delta^{(t)} = (\delta_1^{(t)}, \delta_2^{(t)}, \dots, \delta_C^{(t)})$ as follows:

$$\delta^{(t)} = \sum_{s=1}^{S} w_{win,s} d_s^{(t)},$$

where $d_s^{(t)}$ are the local decision vectors for t. The classification decision is then made as

$$c_{win} = \arg\max_{c} \delta_c,$$

that is, the winning class corresponds to the index of the highest coefficient in δ.

5.4.3 Per-Class Closed-Form Algorithm

To make the system truly adaptive, it is reasonable to assume that different classes require different weight vectors. Thus, we propose a system where, instead of a single weight vector w for the whole training data set, each class c has its own weight vector w_c. As opposed to (5.1), the entries in the output decision vector are now computed as:

$$D^{(r)} w_c, \qquad c = 1, \dots, C. \tag{5.4}$$

Now, the weights can be grouped together to form an $S \times C$ matrix W so that each column represents a class-specific weight vector. Equation (5.4) can be rewritten as:

$$\text{diag}\left(D^{(r)}W\right). \tag{5.5}$$

Recall that $D^{(r)}$ is of size $C \times S$ and thus d is of size $C \times C$ (compare this to (5.1)). To learn these weights, we again use the training set and look for a solution that minimizes the squared error:

$$W_{win} \;=\; \arg\min_{W} \sum_{r=1}^{R} \|d^{(r)} - \text{diag}\left(D^{(r)}W^{(r)}\right)\|^2. \tag{5.6}$$

To obtain an expression analogous to (5.3) and be able to apply standard methods, we have to define v as the vector concatenating all class-specific weight vectors:

$$v \;=\; (W_{1,1}, W_{1,2} \ldots W_{1,C}, \ldots, W_{S,1}, \ldots, W_{S,C})^{\top}. \tag{5.7}$$

We now define D as the following block matrix, where $c_k^{(l)}$, is the vector $(D_{c,1}^{(r)}, D_{c,2}^{(r)}, \ldots, D_{C,S}^{(r)})$

$$D = \begin{pmatrix} d_1^{(1)} & 0 & 0 & \ldots & 0 \\ 0 & d_2^{(1)} & 0 & \ldots & 0 \\ 0 & 0 & d_3^{(1)} & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & d_C^{(1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_1^{(R)} & 0 & 0 & \ldots & 0 \\ 0 & d_2^{(R)} & 0 & \ldots & 0 \\ 0 & 0 & d_3^{(R)} & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & d_C^{(R)} \end{pmatrix}, \tag{5.8}$$

for $r = 1, \ldots, R$. We can now write a minimization problem equivalent to the one in (5.6), and which we can solve using standard techniques:

$$v_{win} \;=\; \arg\min_{v} \|d - Dv\|. \tag{5.9}$$

5.4.4 Decomposition Tree Pruning

Our long-term goal in developing an adaptive MR classification system was to find a wavelet packet-like decomposition, where each class would induce a different MR subtree. While the authors have done just that in [55], we needed a cost function which is specific to the data set used. Our goal is thus have a more generic system and to achieve a wavelet packet-like system but without the need for a cost function.

We come close to this goal here, where we identify the set of discriminative subbands for each class (not necessarily a subtree).

Once the weight vectors are computed (using any version of the weighting algorithm), we use the values of the weights to regulate the MR decomposition. In particular, subbands which are given a low weight by the weighting procedure can be pruned away as long as the remaining subbands are still sufficient to classify the image correctly. This way, the pruned subbands and their associated features need not be computed, resulting in computational savings. We propose to keep the high-weight subbands, so that at least a certain ratio κ, defined as the fraction of the sum of kept weights over the sum of all the weights, of subbands are kept.

This pruning can be done over a single weight vector and is thus suitable for both the previous model with a weight vector per entire data set as well as for the new model with a weight vector per-class (5.4). The process is formalized as Algorithm 3.

Algorithm 3 Pruning the Decomposition Tree

Input: The vector of weights w, fraction of kept weights/subbands κ $(0 < \kappa \leq 1)$
Output: Set of subbands $\mathcal{S}$

$\mathcal{S} \leftarrow \{\}$
while $\left(\sum_{i \in \mathcal{S}} |w_i|\right) < \kappa \sum_{i=1}^{S} |w_i|$ **do**
  $s \leftarrow \arg\max_{s \notin \mathcal{S}} w_s$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{s\}$
end while
return $\mathcal{S}$

Chapter 6

Biomedical Applications

Contents

In this chapter, we discuss the performance of our classification method and use different instantiations of the MR classification algorithm depending on the data set at hand. In each case, we will first describe the data set, then the particular instantiation of the algorithm and finally present classification results.

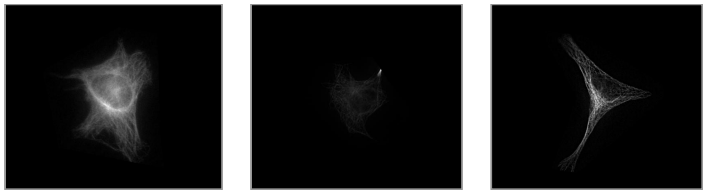
6.1 Determination of Protein Subcellular Location Patterns

This is the application domain discussed in Section 2.1. The goal here is to recognize proteins based on their subcellular location patterns.

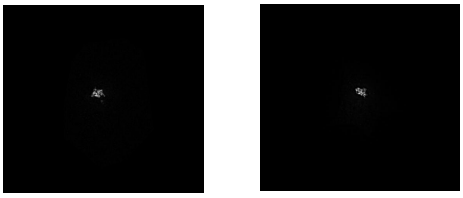
The details of our results in this area can be found in [116, 86, 84, 117, 23].

6.1.1 Data Set

To evaluate our MR approach, we use the 2D HeLa set depicting protein subcellular location described previously [13]. The proteins in the data set were labeled using immunofluorescence, and thus, we know the ground truth, that is, which protein was labeled in each cell and subsequently imaged. This is useful for algorithm development as we can test the accuracy of classification schemes.



(a)



(b)

Figure 6.1: (a) Intra-class variation: The three images show the spatial distribution of tubulin within a cell. (b) Inter-class similarity: The first image shows the spatial distribution of giantin and the second image shows the spatial distribution of gpp130. Both are Golgi proteins. (Images courtesy of Dr. R. F. Murphy, CMU [87].)

The challenge in this data set is that images from the same class may look different while those from different classes may look very similar (see Fig. 6.1).

This data set is publicly available [87] and contains approximately 90 single-cell images of size 512×512, in each of $C = 10$ classes. The 10 classes of subcellular location patterns were obtained by labeling an endoplasmic reticulum protein, two Golgi proteins (giantin and gpp130), a lysosomal protein, a mitochondrial protein, a nucleolar protein, two cytoskeletal proteins (actin and tubulin), an endosomal protein, and DNA. The best previously described overall classification accuracy on this data set is 91.5% [62].

6.1.2 Algorithm

To test our system, we used the following for each block: For the MR block, we use the DWT for MR bases and the stationary wavelet transform that implements the à trous algorithm for MR frames. The feature extraction block uses Haralick texture feature sets T_1, T_2, T_3, morphological features and Zernike moments. The classifier is a neural networks (NN) classifier and for the weighting procedure, we use both open form and per-dataset closed-form versions.

6.1.3 Results

The results are given in Table 6.1, while Fig. 6.2 depicts the following trends (note that No MR denotes the version of the algorithm where no MR transform is used):

1. For all feature combinations, MR bases significantly outperforms no MR, thus demonstrating that classifying in MR subspaces indeed improves classification

System	T	Weight.	Classification accuracy [%]						
			M	T	Z	T, M	M, Z	T, Z	All
NMR	T_1	NW	66.12	85.49	51.20	85.76	72.48	85.06	85.04
	T_2	NW	66.12	85.76	51.20	86.64	72.48	85.78	86.24
	T_3	NW	66.12	**87.46**	51.20	87.38	72.48	87.12	86.86
MRB	T_3	OF	81.62	91.82	65.42	92.04	83.38	91.66	92.36
	T_3	CF	81.48	**92.32**	65.84	92.62	83.58	92.34	92.54
MRF	T_3	OF	84.92	94.72	65.82	94.64	86.80	94.74	94.52
	T_3	CF	85.16	**95.26**	65.24	**95.40**	85.88	95.26	**95.38**

Table 6.1: Classification accuracy for 2D HeLa images depicting protein subcellular location patterns. NMR = no MR, MRB = MR bases, MRF = MR frames, T = texture features, M= morphological features, Z = Zernike moment features, T_1, T_2, T_3 = Haralick texture feature sets T_1, T_2, T_3, NW = no weighting, OF = open-form weighting, CF = per-dataset closed-form weighting.

accuracy.

2. MR frames outperforms MR bases (the only set showing no improvement is the Zernike feature set alone) and gives the best classification accuracy of 95.40%.

3. While a slightly higher classification accuracy is obtained by using all three feature sets as well as both Haralick texture and morphological sets, the larger number of features and additional complexity of using morphological and Zernike features do not justify the slight improvement in accuracy (Haralick texture features T_3 alone achieve 95.26% with MR frames). This "flat" trend (see Fig. 6.2) is good news as we can use a significantly reduced feature set and still obtain a fairly high classification accuracy.

4. For the two versions of the weighting algorithm, open and closed forms, the closed-form algorithm slightly outperforms the open-form one for all feature combinations except for the morphological feature set alone (fourth and fifth rows of Table 6.1). In particular, for Haralick texture features T_3, the accuracy rose from 91.82% to 92.32% in the MR bases case, and from 94.72% to 95.26% in the MR frames case.

6.2 Detection of Developmental Stages in *Drosophila* Embryos

This is the application domain we discussed in Section 2.2. In this classification problem, the aim is to distinguish between three developmental stages of the ventral furrow in fruit fly embryos. The stages consist of the initial stage (no development yet), open stage (during development) and closed or final (development is complete).

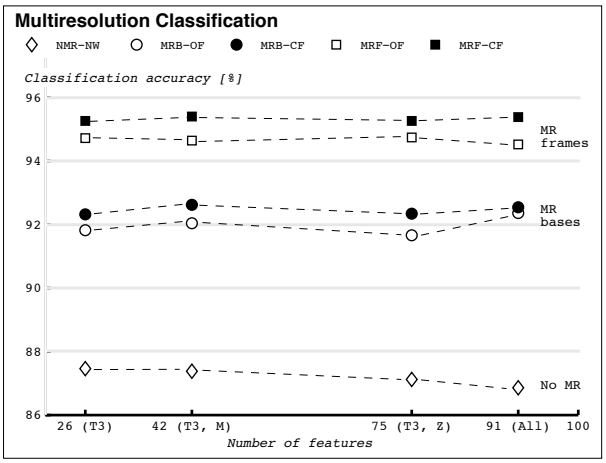


Figure 6.2: Pictorial representation of classification accuracy for 2D HeLa images depicting protein subcellular location patterns.

An automated classification algorithm for this data set can integrated in a high-throughput system allowing for an efficient and accurate identification and screening of large amounts of data.

The details of our results in this area can be found in [66].

6.2.1 Data Set

The data set consists of 60 time-lapse, fluorescence microscopy z-stacks (3D volumes in time) of developmental stages of *Drosophila* embryos. The stacks are acquired roughly every 10 minutes. The number of slices per stack varies; it is 5 slices for normal sets and 7 slices for delayed/abnormal. The number of time points is typically 15 for normal/abnormal and around 30 for delayed. All the slices have been tagged by a human expert so we have reliable ground truth.

6.2.2 Algorithm

For this data set, our task can be divided into two parts: First determine the developmental stage, then associate the time point to be able to tag the development as normal, delayed or abnormal.

Classifier We use the following instantiation of the algorithm: For the MR Block we use the DWT for MR bases, and stationary wavelet transform (à trous algorithm) for MR frames. In the feature extraction phase, we use Haralick texture feature sets T_1, T_3 as well as a combination of Haralick texture features T_3 and morphological features. The classifier is a neural networks (NN) classifier and the weighting algorithm uses both the open form and the per-dataset closed-form.

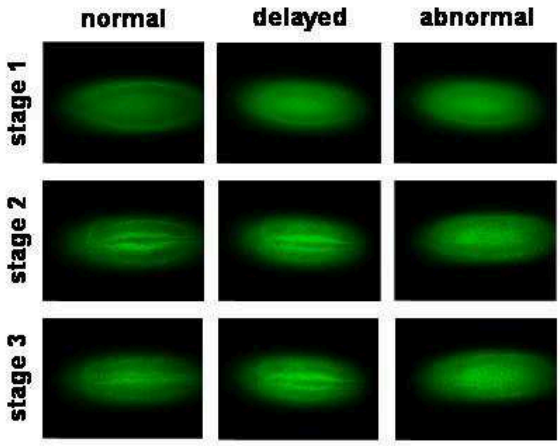


Figure 6.3: Representative examples of each stage. Top: Stage 1, no ventral furrow, for normal (t=30min), delayed (t=60min) and abnormal (t=20min) embryos. Middle: Stage 2, ventral furrow opening, for normal (t=60min), delayed (t=110min) and abnormal (t=72min) embryos. Bottom: Stage 3, ventral furrow closed, for normal (t=75min), delayed (t=140min) and abnormal (t=82min) embryos. (Images courtesy of J. S. Minden, CMU [85].)

Tagging Chart					
	(1,2,3)	(1,1,2)	(1,2,2)	(1,1,1)	(1,3,3)
Tag	Normal	Delayed	Delayed	Abnormal	Abnormal

Table 6.2: Tagging chart for the detection of developmental stages in *Drosophila* embryos. All combinations starting with 2 or 3 will be assumed to be a classifier mistake. Those combinations should be converted to (1,x,y) where x and y are the original stage determination. Any combination starting with 1 and not in the above chart is assumed to be abnormal.

Screening For each time-lapse series, we consider slices at three time points; the first is during the time when Stage 1 is expected to occur, the second is during the time Stage 2 is expected to occur, and likewise for the third time point (these times are known). We then determine normal/delayed/abnormal tags by comparing the expected stages with what the classifier outputs for each set of time points. For example, if the three time points are classified as (1,2,3) (numbers refer to stages), then this is a normal image series. If the classifier labels the images as (1,1,2), then this is a delayed image series. If the classification is (1,1,1), then this is abnormal because it means development did not occur at all. For each combination, we assign a normal/delayed/abnormal tag. Our current assignment is given in Table 6.2. Of

Classification Accuracy [%]				
2D	Weight	T_1	T_3	T_3, M
NMR	NW	82.94	88.39	78.33
MRB	OF	88.11	91.06	89.89
	CF	90.94	92.22	92.78
MRF	OF	83.44	89.95	90.51
	CF	84.83	91.06	**93.17**
3D	Majority rule on 2D			**98.35**

Table 6.3: Classification accuracy for 2D slices of *Drosophila* embryos. We use these in majority voting classification for 3D stacks yielding the accuracy of 98.35%. NMR = no MR, MRB = MR bases, MRF = MR frames, M= morphological features, T_1, T_3 = Haralick texture feature sets T_1, T_3, NW = no weighting, OF = open-form weighting, CF = per-dataset closed-form weighting.

course, it is possible that the sequence (1,1,1) is the correct classification in the first case and incorrect in the last two, leading to an incorrect tag. We will assume that any combination starting with 2 or 3, that is, (2,x,y) or (3,x,y), is a classifier mistake and will convert it into (1,x,y). The combinations starting with 1 not shown in Table 6.2 are assumed to lead to abnormal tags.

6.2.3 Results

Table 6.3 shows the accuracies obtained for the *Drosophila* data set. We can say the following:

1. In all cases, MR does significantly better than no MR.
2. MR frames is better than MR bases when using Haralick feature set T_3 combined with morphological features, with the best classification accuracy of 93.35%.
3. The closed-form version of the weighting process gives better results than the open-form.
4. Access to 3D stacks allows to classify three slices out of each stack and then use a majority rule to make a decision. The classification results for adjacent slices were 92.38% and 91.64% using (T_3, M) features and closed-form weighting. Using the majority rule process, the classification accuracy reaches 98.35%. (Note that this rule assumes independence of the slices in a 3D stack.)

We can use the same process in time as the one for volumes to improve the screening, but since we do not have enough slices in time and need to acquire time-lapse series with better time resolution, this is left for future work. Note that using the majority rule assumes the slices in a 3D stack to be independent. We have not verified this assumption and will leave it for future work.

6.3 Classification of Histological Stem-Cell Teratomas

This is the application domain discussed in Section 2.3 and the details of our results are presented in [26].

For this application, we use our MR classification algorithm to design a system to recognize tissue types within teratomas derived from ES cells. For example, we are given a histological image of an embryonic tissue and would like to recognize that the image depicts skin or muscular tissue. We will show that MR classification again outperforms the no MR one. We test various MR decomposition modules bases (MR bases) frames (MR frames). The complex nature of the images in this work makes it unlikely that the texture features alone could produce high accuracies (in mid to upper 90s as in previous the biomedical application). Therefore, we develop a novel feature set for specific use in recognizing tissues in H&E stained images.

6.3.1 Data Set

Human ES cells (H7 line) and putative nonhuman primate ES cells (derived as outlined by Navara [89]) were introduced into the testes of immune deficient SCID mice by modified efferent duct injection [90]. Cells were injected using an Eppendorf Femtojet pressure injector into the interstitial space of the testis. Tumors typically developed between two and four months after injection. Tumors were removed and processed by routine histological methods. Multiple serial sections were examined for evidence of tissue derived from the three germ layers.

To train our classifier, we need to populate this data set with a fair amount of single-class images. We start with H&E images that depict multiple tissues (classes) contained in the teratomas. These are hand segmented by an expert to separate the classes. Then, the segmentation masks are used to generate single-class images. Because the set of multi-class images available to us is small (only 23 images), we decided to take advantage of their large size (1600×1200), and use a window to extract single-class images of size 200×200. We thus obtain 45 images per class. We use six classes for this experiment: mesenchyme (embryonic connective tissue), skin, myenteric plexus, bone, necrotic (dying or dead tissue), and striated muscle. The images have been taken at 100x magnification (see Fig. 2.2) and have been labeled by a pathologist, thus we have access to ground truth. Note that the test images were never seen by the system during the training phase.

6.3.2 Algorithm

The classification system described in the previous chapter works with gray-scale intensity images. The images obtained through histological techniques have the advantage of being highly detailed and showing distinctive features of the tissues at different resolutions. One of these important features is color. Therefore, it is important that the classification system takes advantage of all the information available from the histological images and exploits a feature as important as color. In particular, when using H&E staining, the nuclei turn a marked purplish/blue color from hematoxylin while the cytoplasm becomes varying shades of red due to exposure to eosin. As a result, we developed new texture features for this data set based on the nuclei present in all images.

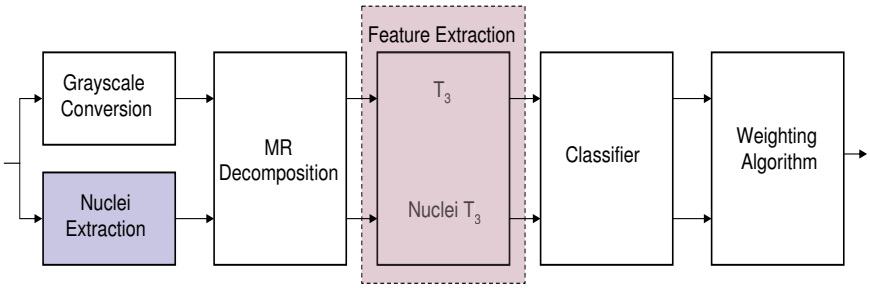
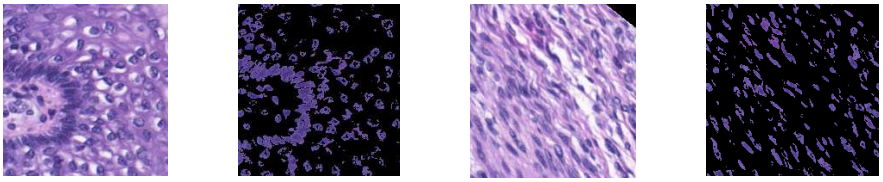


Figure 6.4: Overview of the proposed H&E-stained tissue recognition system. The input is an H&E-stained image of one of the six tissue classes given in Fig. 2.2. The multiresolution (MR) nature of the system is accomplished through the MR decomposition block, after which all the processing is done in MR subspaces. We use Haralick T_3 features [23] and propose new nuclear texture features (see Fig. 6.6). The classifier is a simple neural network one. We use two versions of the weighting algorithm (open form and closed form). The output is the tissue class label.



(a) (b) (c) (d)

Figure 6.5: Examples of tissue and nuclear-only images. (a) Skin, (b) corresponding extracted image of skin nuclei, (c) striated muscle, (d) corresponding extracted image of striated muscle nuclei. (Images (a) and (c) courtesy of Dr. J. A. Ozolek and Dr. C. A. Castro, university of Pittsburgh medical center [92].)

New Nuclear Texture Feature Set (NT_3) We observed that the cell nuclei have a distinctive distribution and texture depending on the tissue type. For example, nuclei in striated muscle image are mostly elongated and have the same orientation (Fig. 6.5(d)), whereas the nuclei in skin images are irregular with a jagged pattern (Fig. 6.5(b)). This prompted us to extract nuclei images from the original histological images and then compute texture features (NT_3) on these, to incorporate them in our classification system. Nuclear features in conjunction with other morphological features have also been used to classify breast cancer tumors in [94].

Fig. 6.6 depicts a block diagram of the nuclei extraction method; We first convert the original images from RGB to the perceptually uniform $L^*a^*b^*$ color space. The L^* channel is the luminance channel and a^* and b^* are the chrominance (color) channels that indicate where the color falls along the red-green and blue-yellow axes, respectively. Given that the images depict mainly three colors: white, blue and pink, we then use the K-means clustering algorithm on the a^* and b^*

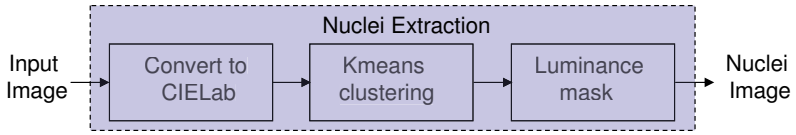


Figure 6.6: Nuclear image extraction.

channels to derive three clusters corresponding to those three colors. We observe experimentally that the centroid with the largest difference between its red and blue channel pixel values corresponds to the blue cluster. Finally, since the nuclei take on the dark shades of blue, we use an adaptive threshold on the luminance channel to mask the lighter shades of blue from the blue cluster image, thus obtaining the nuclei images [128].

Experimental Setup We use the following in our algorithm: The MR block uses the DWT for MR bases and the stationary wavelet transform (à trous algorithm) for MR frames. In the feature extraction step, we use the new nuclear texture features, Haralick texture features T_3 and a combination of both sets at the same time. Neural networks is the classifier and open form and per-dataset closed-form are used as weighting procedures.

Classification Accuracy [%]				
	Weight	NT_3	T_3	T_3, NT_3
NMR	NW	54.44	65.74	71.74
MRB	OF	63.22	70.02	77.29
	CF	64.10	71.38	78.20
MRF	OF	71.51	82.37	86.56
	CF	73.05	84.40	**87.72**

Table 6.4: Classification accuracy for tissue types in teratomas derived from ES cells. Along each row, feature sets are arranged by increased accuracy (with (T_3, NT_3) being the best). Along each column, MR blocks as well as weighting algorithms are arranged by increased accuracy as well; the MR frames gives the best results. NMR = no MR, MRB = MR bases, MRF = MR frames, T_3 = Haralick texture features, NT_3 = new nuclear Haralick texture features, NW = no weighting, OF = open-form weighting, CF = per-dataset closed-form weighting.

6.3.3 Results

The results are given in Table 6.4. We note the following trends (the first three are consistent with the trends observed in all previous applications):

1. For all feature combinations, MR transforms (both MR bases and MR frames)

significantly outperform no MR, thus showing that classifying in MR subspaces indeed improves the classification accuracy.

2. MR frames considerably outperforms MR bases and give the best classification accuracy of 87.72%.

3. The closed-form version of the algorithm outperforms the open-form one.

4. Incorporating nuclear texture features significantly improves the accuracy using any MR block and any weighting algorithm. In particular, classification accuracy increases from 73.05% using nuclear texture features NT_3 only, to 84.40% using Haralick T_3 features only, and to 87.72% using the combined feature set T_3, NT_3.

For tissue type identification, we hope to improve the performance of the classification system by adding morphological features, as the shape of the cell and nucleus, which as noted demonstrates distinctive variations across tissue types. We will also incorporated 3D features as well as obtain more images for the training set so the classifier can see a larger variation during training. Our larger goal is to build an automated toolbox for extraction, recognition and quantification of the varied tissue types present in teratomas derived from ES cells and other pathological specimens using only routine hematoxylin and eosin stained tissue sections.

6.4 Classification of Otitis Media Stages

This is the application described in Section 2.4. It concerns infections of the middle ear and our goal is to distinguish between three possible stages: normal when no infection is present, OME when the infection is mild and AOM when the infection is the most severe.

6.4.1 Data Set

The images in this data set were acquired with a spectroscopic otoscope. A standard otoscope was connected to a spectrometer by an optical fiber wire. The otoscope collects the light reflected from the eardrum and directs it to spectrometer. The reflected light was then sampled and transferred to a computer [109]. The data set used in this work was supplied by Dr. Hoberman at the University of Pittsburgh Medical Center. The images were labeled by an expert, providing us with the ground truth. Due to very high variability in the images within and across classes dues to different lighting conditions and other elements of noise (see Fig. 2.3), we decided to hand-segment the images so as to keep only relevant portions of the images consisting of the tympanic membranes.

We use three classes for this experiment: normal, otitis media with effusion (OME) and acute otitis media (AOM). Each class contains 50 images of size 640×480 (see Fig. 2.3).

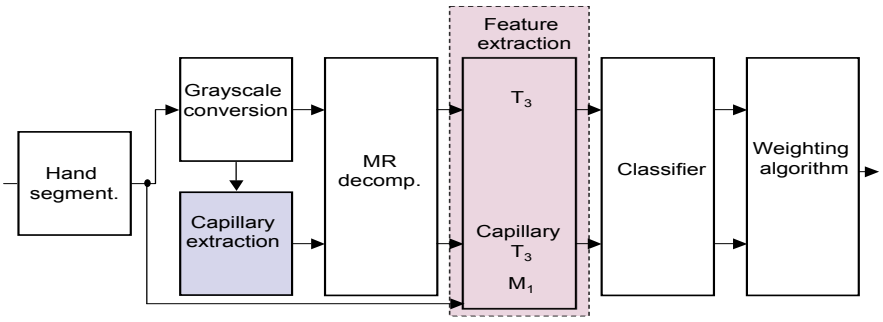


Figure 6.7: MR Classification system for otitis media data set.

	Otitis Media Stage		
Finding	Normal	OME	AOM
Color	Grey, pink	White, amber, grey, blue	White, pale yellow, marked redness
Position	Neutral, retracted	Neutral, retracted	Distinct fullness, bulging
Translucency	Translucent	Opacified, semi-opaque	Opacified

Table 6.5: Otoscopic findings associated with stages of otitis media. These were gathered by the University of Pittsburgh Medical Center [114] and are some of the observations used by physicians to diagnose otitis media.

6.4.2 Algorithm

Similarly to the previous application, this data set contains color images with specific characteristics. Here again, we need to take advantage of this additional information and create new features tailored to the otitis media images. As usual, the goal is to obtain discriminative features that allow us to distinguish between the three classes. As discussed in Section 2.4, this is an arduous task since even experts very often misdiagnose infections in the middle ear. In particular, it is hard to differentiate between OME and AOM.

Table 6.5 identifies important findings that physicians consider while examining and diagnosing middle ear effusion. We tried to inspire ourselves from these features. Unfortunately, most of them, such as translucency, are very hard to translate in terms of analytical formulas. Instead, we use these features indirectly. We combined the experts' observations on patients with our own observations of the images and extract discriminative features, that we used in addition to Haralick texture features (T_3). These novel features can be grouped into two main categories: Capillary Haralick texture features (CT_3) and morphological otitis media features (M_1).

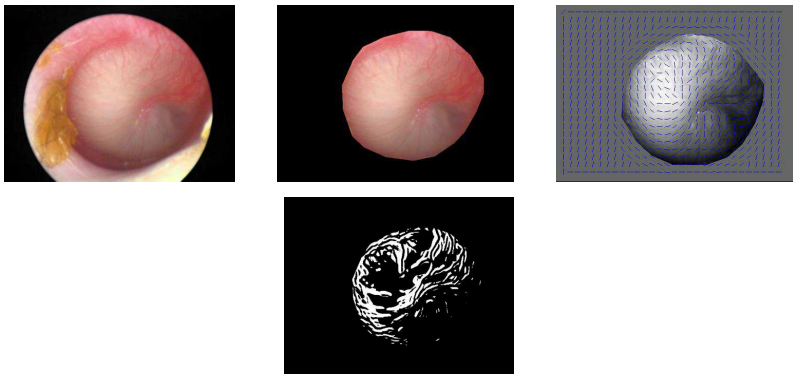


Figure 6.8: Example of otitis media and capillary-only images. (a) Original otitis media image (from the AOM class, courtesy of Dr. Hoberman, university of Pittsburgh medical center [56]), (b) hand-segmented image, (c) capillary orientation image, (d) capillary-only image.

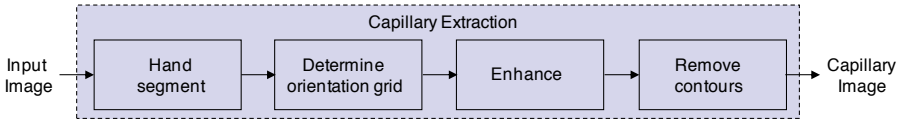


Figure 6.9: Capillary image extraction.

New Capillary Texture Feature Set (CT_3) Capillaries are small blood vessels that are present in the tympanic membrane of the ear. When the ear is normal, none to little capillaries are visible. However, when the ear becomes more and more infected and starts to bulge, as observed in the OME and AOM conditions, the capillaries of the tympanic membrane become more and more visible. Therefore, the textures of capillary images were deemed important features to discriminate between the classes. We extract the capillary images using a ridge detection algorithm inspired by the work in [59] and developed by Kovesi [75]. The ridge detection algorithm takes as its input the hand-segmented, grayscale version of the original RGB image and first identifies ridge-like regions in the image. Then, the orientation of the capillaries is determined by comparing each pixel to its surroundings in the gradient images. Regions with reliable capillary orientation are found in areas that have a significant gradient difference. Here, reliability is a value between 0 and 1 that measures the orientation detected. The orientation measure is deemed good if the reliability is above 0.5. Next, oriented filters are used to enhance the ridge pattern. Finally, the image is converted to a black and white mask and the outer contour is removed. This final step is performed because, although it appears as a ridge in the image, the outer contour is in fact the edge of the tympanic membrane and does not represent any capillaries. Fig 6.9 shows a block diagram for the ridge detection

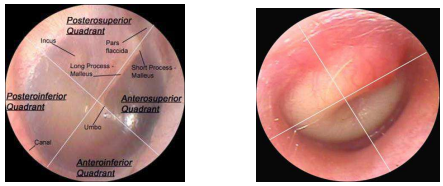


Figure 6.10: Positioning of the tympanic membrane (courtesy of [114]). (a) Neutral position of the short process in the normal or OME case, (b) obscured short process when bulging occurs (AOM).

algorithm. We compute Haralick texture features T_3 on the binary capillary images and obtain the new feature set CT_3. Fig 6.8 shows an example of an AOM image, its hand-segmented version, its capillary orientation image and the capillary-only final black and white image.

New Morphological Feature Set (M_1) The positioning of the tympanic membrane (TM) distinguishes AOM from other otitis media classes and can be translated into a contour-type of feature. Middle ears afflicted with AOM exhibit a distinctive bulging, fullness of the TM, whereas in OME and normal middle ears, the TM positioning is either neutral or retracted. Bulging in otitis media images is detected by an obscurity in the short, lateral process of the malleus depicted in Fig. 6.10. Moreover, a greater level of obscurity of the short process produces an increasingly elliptical shape of the TM [114]. Thus, we quantify the level of obscurity in the short process by computing the convex hull area, the eccentricity, and the ratio of the minor to major axes of the TM region. The convex hull area, the area of the smallest convex polygon that contains the TM region, was computed to extract the relative shape and size of this region. The eccentricity is computed as the ratio of the distance between the foci, the center point of the ellipse and its major axis length. The ratio of the minor to major axis of the TM is the scalar quotient of the length in pixels of the minor to the major axes of the TM.

In addition to these features, we compute color based features. The images were first preprocessed to remove the black pixels of the background in the hand-segmented RGB images. After some investigative work, we found that the mean of the red channel in the RGB images can be used as a discriminative feature. Indeed, these mean values yielded the best separation amongst the three classes. The images of the AOM class contained on average higher number of red valued pixels. This was to be expected given the observation that AOM cases exhibit a distinctive marked redness (see Table 6.5).

Experimental Setup We used the following instantiation of our algorithm: The MR block has the DWT for MR bases and the stationary wavelet transform (à trous algorithm) for MR frames. The feature extraction block uses capillary texture features CT_3, Haralick T_3 and new morphological features, as well as a combination of all these feature sets. The classifier has neural networks and the open-form

weighting algorithm is used.

6.4.3 Results

Classification Accuracy [%]				
	Weight	CT_3	CT_3, M_1, T_3	T_3
NMR	NW	44.40	54.67	55.67
MRB	OF	44.40	64.73	65.93
MRF	OF	45.73	71.93	**73.13**

Table 6.6: Classification accuracy for otitis media. MR frames gives the best results with Haralick texture features T_3. NMR = no MR, MRB = MR bases, MRF = MR frames, T_3 = Haralick texture features, CT_3 = new capillary Haralick texture features, M_1 = new morphological features, NW = no weighting, OF = open-form weighting, CF = per-dataset closed-form weighting.

The classification accuracies are reported in Table 6.6. We observe the following:

1. MR substantially outperforms no MR in all cases.

2. MR frames is always better than MR bases, with the best classification accuracy of 73.13% achieved with Haralick texture features T_3.

3. Surprisingly, adding capillary texture features CT_3 and new morphological features to Haralick T_3 decreases slightly the classification accuracy from 73.13% to 71.93%.

In this application, the MR classification system does not perform as in the previous ones. Moreover, the addition of novel features unexpectedly leads to lower accuracies than the simple use of Haralick texture features T_3. We hypothesize that this might be due to two aspects of the feature extraction step: The first is that the ridge extraction algorithm did not perform well on this type of images. The second, is that the set of new features along with T_3 were in "competition" and instead of helping the classifier make correct decisions, it confused it. If this is the case, a feature selection procedure should solve the problem.

The confusion matrices for no MR, MR bases and MR frames are presented in Table 6.7. We observe that in every case, more than 50% of the images were assigned the correct class label. As expected, we note that AOM and OME classes have the largest percentages of incorrect classifications. In addition, these two classes hold most of the confusion where one class is mistaken to be the other and vice versa. This result supports the fact that the features of AOM and OME are very similar and can often be confused.

	Classification Accuracy [%]		
	Normal	OME	AOM
NMR			
Normal	65.80	22.60	11.60
OME	28.60	42.60	28.60
AOM	15.00	29.40	55.60
MR bases			
Normal	81.80	11.60	6.60
OME	17.20	52.40	30.40
AOM	12.00	28.00	60.00
MR frames			
Normal	93.00	3.80	3.20
OME	6.80	59.80	33.40
AOM	6.00	31.00	63.00

Table 6.7: Confusion matrices for otitis media classification when using all features (new capillary features CT_3, new morphological features M_1, Haralick texture features T_3). Note that in each case, the classification accuracy can be computed as the average of the diagonal element of each matrix. NMR = no MR, MRB = MR bases, MRF = MR frames.

6.5 Application in Other Domains: Fingerprint Recognition

This is the application domain discussed in Section 2.5. Given an image of a fingerprint, the goal here is to recognize the individual (class) to whom the fingerprint belongs. The details of our results can be found in [24].

6.5.1 Data Set

To test our system we used images from a subset of the NIST 24 fingerprint database [127]. The data set contains 10 classes with 50 512×512 images each (45 images are used to train the system). The images were acquired while individuals were rolling their thumbs, inducing different plastic distortions making the data set realistic and challenging (see Fig. 2.4).

6.5.2 Algorithm

To classify this data set, we use different MR transforms as well as various weighting procedures. In the MR block, for MR bases we use the DWT of size 2×2, and the following 4×4 transforms: the discrete Fourier transform, the discrete cosine transform, the discrete Hartley transform [9]), the Walsh-Hadamard transform [64]) and the discrete triangle transform [99] which is nonseparable. We also use two random unitary transforms, the first one has an all ones row, whereas the second is completely random. For MR frames we use the double-density DWT, the dual-tree complex wavelet transform (both of these were presented in Section 4.4.6) and the stationary wavelet transform that implements the algorithm à trous. We use

Haralick texture feature set T_3 and neural networks in the generic classification system. For the weighting process, we use per-dataset closed-form and per-class closed-form. When using the pruning procedure, we set the value for κ at 0.8 as initial observations showed that this value achieved a good balance between pruning away the decomposition tree while keeping the accuracy high.

6.5.3 Results

	Pruned		Not pruned	
	Per-Class	CF	Per-Class	CF
NMR	96.22	96.22	96.22	96.22
MR bases				
DWT	98.86	98.82	98.58	98.68
DFT	98.26	98.18	98.42	98.46
DCT	95.08	94.46	98.10	98.02
DHT	95.48	95.06	98.00	97.78
WHT	95.02	94.34	98.12	98.08
DTT	98.02	97.92	98.30	98.28
RU1	97.12	97.00	99.00	98.98
RU2	94.90	94.84	98.12	98.18
MR frames				
DD-DWT	98.96	99.10	98.70	99.12
DT-CWT	99.06	98.52	99.14	98.80
SWT	99.36	99.38	99.42	**99.50**

Table 6.8: Classification accuracies for fingerprint images obtained with different MR transforms, Haralick texture features T_3, using two weighting algorithms and a pruning procedure. For MR bases, we have the following transforms: discrete Hartley transform (DHT), Walsh-Hadamard transform (WHT), discrete triangle transform (DTT), random unitary transforms RU1 and RU2. For MR frames, we used the double-density DWT (DD-DWT), dual-tree complex wavelet transform (DT-CWT), Algorithm à trous (SWT). CF = per-dataset weighting procedure.

All the results are shown in Table 6.8. By observing the results, we can draw the following conclusions:

1. MR does better than no MR.
2. The redundant transforms (MR frames) do better than the unitary ones (MR bases) and the SWT achieves the best classification accuracy of 99.50%.
3. The choice of the transform amongst MR bases does not seem to be crucial. One might as well use a random unitary transform and still achieve similar performances.
4. As expected, pruning does not improve the accuracy of the system, but it does make it more efficient.

5. In general, the class-adaptive method seems to do better than the data set adaptive one.

6. Considering the two main MR decompositions DWT and SWT, using $\kappa = 0.8$ in the pruning procedure removed almost half of the subbands, enabling significant computational savings in computation with a small impact on the classification accuracy.

For future work, we intend to use a much smaller training set of images to train our system, use a much larger data set as well as optimize κ for each transform.

6.6 Towards a Theory of Frame Multiresolution Classification

We investigated in this work five different biomedical and biometrics applications that had classification as their underlying task. We have developed an accurate and efficient adaptive supervised classification algorithm based on multiresolution (MR) techniques, aiming to extract discriminative features within space-frequency localized MR subspaces. These are obtained by MR decomposition; that is, rather than add MR features to existing features, we instead chose to compute these features in the MR-decomposed subspaces themselves. Thus, our system has an upfront MR decomposition block which is followed by feature computation and classification in each of the MR subspaces, which, in turn, are then combined through an adaptive weighting process.

In the table below, we summarize the classification accuracies obtained for each application:

Project	**Accuracy [%]**	**Reference**
Recognizing proteins using subcellular location patterns	95.4	[23]
Determination of developmental stages in *Drosophila* embryos	93.17 (2D) 98.35 (3D)	[66]
Tissue recognition in stem cell teratomas	87.72	[26]
Diagnosis of otitis media	73.13	Initial work
Fingerprint recognition	99.5	[24]

In all the applications we examined, and despite the variation in the nature of the data sets (cellular, tissue nature, various imaging modalities, etc), we observed two persistent trends:

- Using the MR block in front of the generic classifier is always better and produces a higher accuracy than not having an MR transform.

- MR frames always outperform MR bases and lead to the best classification results.

The first trend was expected as it does subscribe to our philosophy: "MR subspaces contain useful information for classification". However, the second trend

was not a result we had anticipated. This directed us to ask the following two fundamental questions:

1. Why do MR frames perform better than MR bases in a classification context?
2. Can we design new frame families that are application-specific?

Our answers to the above questions are the topic of the next part of this thesis. In the next chapter, we simplify the first question to a more tractable set up and provide a rigorous and formal approach to explaining why frames outperform bases in classification. Specifically, we consider only one class of signals which is generalized to be a convex set. In addition, we suppress the feature extraction step and consider the transform coefficients of a signal as features themselves. Finally, we simplify the classifier to a classification scheme akin to a majority voting. This basic but fundamental setting helps us understand the role of frames and redundancy in classification and is a general mathematical framework that allows rigorous analysis of frame-based classification. In Chapter 8, we design new frame families that generalize lapped orthogonal transforms to frames. As we will see, these new families are flexible, tight and efficient by construction and have other desirable properties as well.

Part IV

Theory of Frame Multiresolution Classification

Chapter 7

Frame Classification

Contents

As the five biomedical and biometrics applications we investigated in this thesis have very different images and are often obtained through different imaging modalities, it came as a surprise that all five shared the same trend: The system using MR frames invariably had a higher accuracy than the one using MR bases. This led us to ask the following fundamental question: Why do frames always perform better than bases when it comes to classification?

In this chapter, we detail our first steps towards answering this question. Our approach involves a frame theory-based scheme for the classification of convex sets, and a measure-theoretic framework for evaluating its performance. We look at a single-class classification problem (sometimes also called recognition problem) in $\mathbb{R}^N$ where the class $\mathcal{C}$ to be recognized is a compact convex subset of $\mathbb{R}^N$. We assume we have the complete knowledge of the class $\mathcal{C}$; we do not address the process by which we may reach this level of knowledge by learning the support and distribution of the points inside of $\mathcal{C}$ from a finite number of training samples. Rather, we focus on the problem of how best to determine whether a given point $x \in \mathbb{R}^N$ lies in $\mathcal{C}$ when one only knows the point's transform coefficients $\{\langle x, \tilde{\varphi}_m^* \rangle\}_{m=1}^M$ with respect to some frame $\{\varphi_m\}_{m=1}^M$ for $\mathbb{R}^N$. We also examine this problem in the presence of noise.

This chapter is organized as follows: In Section 7.1, we first prove that for a compact convex class $\mathcal{C}$, the set of points $\mathcal{E}_p(\mathcal{C})$ at which classification errors due to noise will occur, may be approximated by $\mathcal{E}_p(\hat{\mathcal{C}})$, provided the *approximating set* $\hat{\mathcal{C}}$ is

sufficiently close to $\mathcal{C}$. We thus rephrase our problem as the search for "convenient" approximating sets, $\hat{\mathcal{C}}$. As $\mathcal{C}$ is convex, convex polytopes are natural candidates for the approximating set $\hat{\mathcal{C}}$. In Section 7.2, we show how convex polytopes are a special case of sets that naturally arise in the context of frame theory, and term these sets *frame sets*. In Section 7.3, we show how frame sets may be used to classify elements of $\mathcal{C}$, and study the errors associated with such a classification scheme. We then conclude in Section 7.4 by deriving upper bounds on the size of the set of points at which these errors are most likely to occur.

7.1 Classification of Convex Sets in the Presence of Noise

Let $\mathcal{C} \subset \mathbb{R}^N$ be a compact convex set that represents our class of signals; in the absence of noise, the ideal classifier function $\chi_{\mathcal{C}}$ perfectly determines whether a given point is in $\mathcal{C}$ or not. However, even an ideal classifier may be affected by the presence of noise, resulting in classification errors. In this work, we shall consider a probabilistic, radially symmetric noise model. Specifically, let P be a probability measure on $\mathbb{R}^N$ which is absolutely continuous with respect to Lebesgue measure, and take $p \in L^1(\mathbb{R}^N)$ to be its density function, that is,

$$P(\mathcal{S}) := \int_{\mathbb{R}^N} \chi_{\mathcal{S}}(\eta) p(\eta)\, \mathrm{d}\eta,$$

for any Lebesgue measurable set $\mathcal{S} \subseteq \mathbb{R}^N$, where $p(\eta) \geq 0$ for all $\eta \in \mathbb{R}^N$ and $P(\mathbb{R}^N) = 1$. For a given point $x \in \mathbb{R}^N$, the quantity:

$$P(x + \eta \in \mathcal{S}) := \int_{\mathbb{R}^N} \chi_{\mathcal{S}}(x + \eta) p(\eta)\, \mathrm{d}\eta \tag{7.1}$$

represents the probability that the additive noise η perturbs x into being an element of $\mathcal{S}$.

As we have seen in Section 3.3.2, there are two types of classification errors: Type-I errors (false positives) in which a point is not an element of $\mathcal{C}$ but is classified as one, and Type-II errors (false negatives) in which a point is an element of $\mathcal{C}$ but is classified as not. Though either type of error may occur at any point, here we are especially concerned with those points x for which the addition of the noise η results in x being misclassified more than half of the time. To be precise, the set of points for which Type-I errors occur more frequently than not is:

$$\mathcal{E}_{\mathrm{I}} = \{x \notin \mathcal{C} : P(x + \eta \in \mathcal{C}) > \tfrac{1}{2}\}, \tag{7.2}$$

while the corresponding set for Type-II errors is:

$$\mathcal{E}_{\mathrm{II}} = \{x \in \mathcal{C} : P(x + \eta \notin \mathcal{C}) > \tfrac{1}{2}\}. \tag{7.3}$$

For a given class $\mathcal{C}$ and noise distribution p, we define the corresponding total classification error $\mathrm{Er}(\mathcal{C}, p)$ to be the sum of the Lebesgue measures of $\mathcal{E}_{\mathrm{I}}$ and $\mathcal{E}_{\mathrm{II}}$; these sets are indeed measurable as (7.1) is a continuous function of x.

7.1.1 Classification Error of Convex Sets in the Presence of Additive Radially Symmetric Noise

We now show that $\mathcal{E}_{\mathrm{I}}$ is empty whenever $\mathcal{C}$ is convex and p is radially symmetric, that is, when $p(U\eta) = p(\eta)$ for all orthogonal matrices U. Before stating this result, we note that radial symmetry implies $p(-\eta) = p(\eta)$ and so:

$$P(x+\eta \in \mathcal{C}) = \int_{\mathbb{R}^N} \chi_{\mathcal{C}}(x+\eta)p(\eta)\,\mathrm{d}\eta = \int_{\mathbb{R}^N} \chi_{\mathcal{C}}(x-\eta)p(\eta)\,\mathrm{d}\eta = (\chi_{\mathcal{C}} * p)(x),$$

where $\chi_{\mathcal{C}} * p$ is the convolution of $\chi_{\mathcal{C}}$ and p. Since $(\chi_{\mathcal{C}} * p)(x) \in [0,1]$ for all $x \in \mathbb{R}^N$, we may write:

$$\begin{aligned}\{x \in \mathbb{R}^N : P(x+\eta \notin \mathcal{C}) > \tfrac{1}{2}\} &= \{x \in \mathbb{R}^N : P(x+\eta \in \mathcal{C}) \leq \tfrac{1}{2}\} \\ &= (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}],\end{aligned}$$

that is, the preimage of $[0, \frac{1}{2}]$ under the convolution $\chi_{\mathcal{C}} * p$, while

$$\{x \in \mathbb{R}^N : P(x+\eta \in \mathcal{C}) > \tfrac{1}{2}\} = (\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2}, 1].$$

Thus, $\mathcal{E}_{\mathrm{I}}$ may be written as $\mathcal{C}^{\mathrm{c}} \cap (\chi_{\mathcal{C}} * p)^{-1}(\frac{1}{2}, 1]$ and $\mathcal{E}_{\mathrm{II}}$ may be written as $\mathcal{C} \cap (\chi_{\mathcal{C}} * p)^{-1}[0, \frac{1}{2}]$.

LEMMA 7.1. If $\mathcal{C}$ is convex and p is a radially symmetric probability density function over $\mathbb{R}^N$, then for any $x \notin \mathcal{C}$,

$$P(x+\eta \in \mathcal{C}) = \int_{\mathbb{R}^N} \chi_{\mathcal{C}}(x+\eta)p(\eta)\,\mathrm{d}\eta = (\chi_{\mathcal{C}} * p)(x) \leq \tfrac{1}{2}.$$

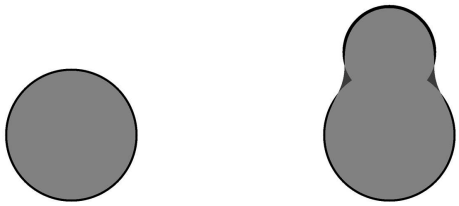
Equivalently, $(\chi_{\mathcal{C}} * p)^{-1}(\frac{1}{2}, 1] \subseteq \mathcal{C}$.

Proof. Since $\mathcal{C}$ is convex and $x \notin \mathcal{C}$, there exists $y \in \mathbb{R}^N$ such that $\langle z - x, y\rangle > 0$ for all $z \in \mathcal{C}$ [8], that is, such that $\mathcal{C}$ lies in a half-space whose boundary contains x. Thus,

$$\begin{aligned}&P(x+\eta \in \mathcal{C}) \\ &\quad= \int_{\mathbb{R}^N} \chi_{\mathcal{C}}(x+\eta)p(\eta)\,\mathrm{d}\eta \\ &\quad= \int_{\{\eta:\langle \eta, y\rangle > 0\}} \chi_{\mathcal{C}}(x+\eta)p(\eta)\,\mathrm{d}\eta + \int_{\{\eta:\langle \eta, y\rangle \leq 0\}} \chi_{\mathcal{C}}(x+\eta)p(\eta)\,\mathrm{d}\eta.\end{aligned}$$

If $\langle \eta, y\rangle \leq 0$, then $\langle (x+\eta) - x, y\rangle \leq 0$ and so $x+\eta \notin \mathcal{C}$, that is, $\chi_{\mathcal{C}}(x+\eta) = 0$. Thus,

$$P(x+\eta \in \mathcal{C}) = \int_{\{\eta:\langle \eta, y\rangle > 0\}} \chi_{\mathcal{C}}(x+\eta)p(\eta)\,\mathrm{d}\eta \leq \int_{\{\eta:\langle \eta, y\rangle > 0\}} p(\eta)\,\mathrm{d}\eta. \tag{7.4}$$



(a) (b)

Figure 7.1: Type-I and Type-II errors in a convex and a nonconvex class for a Gaussian noise model p of mean zero and standard deviation $\sigma = 0.25$. The classes are indicated in medium gray, Type-I errors in dark gray, and Type-II errors in black. (a) Error set $\mathcal{E}_p(\mathcal{C})$ is a ring inscribed in $\mathcal{C}$ when $\mathcal{C}$ is the disk centered at $(0,0)$, of radius 1. Here, only Type-II errors exist (black). (b) Nonconvex class $\mathcal{S}$, and the associated Type-I (dark gray) and Type-II (black) errors.

However, as $\{\eta : \langle \eta, y\rangle = 0\}$ has measure zero and the radial symmetry of p implies $p(-\eta) = p(\eta)$ for all $\eta \in \mathbb{R}^N$, then:

$$\begin{aligned} 1 &= \int_{\mathbb{R}^N} p(\eta)\,\mathrm{d}\eta \\ &= \int_{\{\eta:\langle \eta,y\rangle>0\}} p(\eta)\,\mathrm{d}\eta + \int_{\{\eta:\langle \eta,y\rangle<0\}} p(\eta)\,\mathrm{d}\eta \\ &= 2\int_{\{\eta:\langle \eta,y\rangle>0\}} p(\eta)\,\mathrm{d}\eta, \end{aligned}$$

which, when combined with (7.4), yields the result. □ Note that when $\mathcal{C}$ is convex

and p is radially symmetric, Lemma 7.1 implies that $\mathcal{E}_{\mathrm{I}}$ is empty. That is, in this case, Type-I errors, though still possible, do not occur more than half of the time at any point x. Therefore, in this case, we define the error set of a compact convex set $\mathcal{C}$ with respect to a radially symmetric noise model p as just the set of points at which Type-II will occur more often than not:

$$\mathcal{E}_p(\mathcal{C}) := \{x \in \mathcal{C} : P(x + \eta \notin \mathcal{C}) > \tfrac{1}{2}\}. \tag{7.5}$$

Here, the total classification error is simply $\mathrm{Er}(\mathcal{C}, p) = m(\mathcal{E}_p(\mathcal{C}))$, where m is the Lebesgue measure.

Example 7.1. To visualize the error set of a convex versus a nonconvex set $\mathcal{C}$, we consider the class $\mathcal{C}$ to be the disk in $\mathbb{R}^2$ centered at $(0,0)$ and of radius 1 (convex set, Fig. 7.1(a)). In Fig. 7.1(b) a nonconvex example is given. The figure illustrates Type-I and Type-II errors for a Gaussian noise model p of mean zero and standard

deviation $\sigma = 0.25$. For the convex set $\mathcal{C}$, there are only Type-II errors (black), whereas for the nonconvex set both types of errors exist (Type-I is in dark gray). ∎

7.1.2 Classification of Convex Sets via Approximating Sets

For a given convex class $\mathcal{C}$, a useful, explicit expression of its error set (7.5) may be difficult to obtain. In the sections below, we show that this problem becomes easier when the class in question is a convex polytope, dubbed a frame set. The question therefore arises: if the actual class $\mathcal{C}$ is well approximated by some geometrically nice *approximating set* $\hat{\mathcal{C}}$, is it true that the error set of $\mathcal{C}$ is well approximated by the error set of $\hat{\mathcal{C}}$?

We now show this is indeed true, provided the distance between $\mathcal{C}$ and $\hat{\mathcal{C}}$ is taken to be:

$$d(\mathcal{C}, \hat{\mathcal{C}}) := m(\mathcal{C} \cap \hat{\mathcal{C}}^{\mathrm{c}}) + m(\hat{\mathcal{C}} \cap \mathcal{C}^{\mathrm{c}}), \tag{7.6}$$

and provided that $\mathcal{C}$ is sufficiently regular. We say that a compact convex set $\mathcal{C}$ is *regular* with respect to p if:

$$\begin{aligned} 0 &= m\{x \in \mathbb{R}^N : P(x+\eta) = \tfrac{1}{2}\} \\ &= m\Big\{x \in \mathbb{R}^N : \int_{\mathbb{R}^N} \chi_{\mathcal{C}}(x+\eta) p(\eta)\, \mathrm{d}\eta = \tfrac{1}{2}\Big\}. \end{aligned} \tag{7.7}$$

We note that for typical p and $\mathcal{C}$, the set being measured in (7.7) is a level surface of the convolution $\chi_{\mathcal{C}} * p$, which one usually expects to have measure zero. Nevertheless, the explicit need for this additional assumption on p and $\mathcal{C}$ will become apparent in the proof of the following result; the study of sufficient conditions on p and $\mathcal{C}$ so as to guarantee regularity is left as future work.

THEOREM 7.2. For any fixed radially symmetric noise model p, and any compact convex set $\mathcal{C}$ which is regular as in (7.7) with respect to p, we have:

$$\lim_{\hat{\mathcal{C}} \to \mathcal{C}} d((\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}], (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}]) = 0, \tag{7.8}$$

where $\hat{\mathcal{C}}$ may be any compact convex set. As a consequence, we also have that the error set function $\mathcal{E}_p(\mathcal{C})$ is continuous, that is, $\lim_{\hat{\mathcal{C}} \to \mathcal{C}} \mathcal{E}_p(\hat{\mathcal{C}}) = \mathcal{E}_p(\mathcal{C})$.

Proof. We first prove (7.8), that is, taking any regular compact convex set $\mathcal{C}$ and any $\varepsilon > 0$, we shall show there exists $\delta > 0$ such that

$$\begin{aligned} & d((\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}], (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}]) && (7.9) \\ &\quad = m((\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}] \cap (\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2}, 1]) && (7.10) \\ &\qquad + m((\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}] \cap (\chi_{\hat{\mathcal{C}}} * p)^{-1}(\tfrac{1}{2}, 1]) && (7.11) \\ &\quad < \varepsilon, \end{aligned}$$

whenever $\hat{\mathcal{C}}$ is any compact convex set such that $d(\hat{\mathcal{C}}, \mathcal{C}) < \delta$. To estimate the size of (7.10), note that since $|p(\eta)| \leq 1$ for all $\eta \in \mathbb{R}^N$,

$$\begin{aligned}
|(\chi_{\hat{\mathcal{C}}} * p)(x) - (\chi_{\mathcal{C}} * p)(x)| &= \Big| \int_{\mathbb{R}^N} \big[\chi_{\hat{\mathcal{C}}}(x+\eta) - \chi_{\mathcal{C}}(x+\eta)\big] p(\eta)\, \mathrm{d}\eta \Big| \\
&\leq \int_{\mathbb{R}^N} \big|\chi_{\hat{\mathcal{C}}}(x+\eta) - \chi_{\mathcal{C}}(x+\eta)\big|\, \mathrm{d}\eta \\
&= \int_{\mathbb{R}^N} \chi_{(\hat{\mathcal{C}} \cap \mathcal{C}^{\mathrm{c}}) \cup (\mathcal{C} \cap \hat{\mathcal{C}}^{\mathrm{c}})}(x+\eta)\, \mathrm{d}\eta \\
&= m((\hat{\mathcal{C}} \cap \mathcal{C}^{\mathrm{c}}) \cup (\mathcal{C} \cap \hat{\mathcal{C}}^{\mathrm{c}})) \\
&= d(\hat{\mathcal{C}}, \mathcal{C}),
\end{aligned}$$

and thus, for any fixed positive integer k, any compact convex set $\hat{\mathcal{C}}$ such that $d(\hat{\mathcal{C}}, \mathcal{C}) < \frac{1}{k}$, and any $x \in (\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \frac{1}{2}] \cap (\chi_{\mathcal{C}} * p)^{-1}(\frac{1}{2}, 1]$,

$$\begin{aligned}
\tfrac{1}{2} &< (\chi_{\mathcal{C}} * p)(x) \\
&\leq |(\chi_{\mathcal{C}} * p)(x) - (\chi_{\hat{\mathcal{C}}} * p)(x)| + |(\chi_{\hat{\mathcal{C}}} * p)(x)| \\
&< \tfrac{1}{k} + \tfrac{1}{2}.
\end{aligned}$$

Thus, for any such k and $\hat{\mathcal{C}}$, we have:

$$(\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}] \cap (\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2}, 1] \subseteq (\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2}, \tfrac{1}{2} + \tfrac{1}{k}). \tag{7.12}$$

Applying a similar reasoning to the set in (7.11), we note that if $x \in (\chi_{\mathcal{C}} * p)^{-1}[0, \frac{1}{2}] \cap (\chi_{\hat{\mathcal{C}}} * p)^{-1}(\frac{1}{2}, 1]$, then

$$\begin{aligned}
\tfrac{1}{2} &\geq (\chi_{\mathcal{C}} * p)(x) \\
&\geq (\chi_{\hat{\mathcal{C}}} * p)(x) - |(\chi_{\hat{\mathcal{C}}} * p)(x) - (\chi_{\mathcal{C}} * p)(x)| \\
&> \tfrac{1}{2} - \tfrac{1}{k},
\end{aligned}$$

and so for any such k and $\hat{\mathcal{C}}$, we also have:

$$(\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}] \cap (\chi_{\hat{\mathcal{C}}} * p)^{-1}(\tfrac{1}{2}, 1] \subseteq (\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2}]. \tag{7.13}$$

Summing the measures of (7.12) and (7.13), we therefore obtain an upper bound on the left hand side of (7.10):

$$\begin{aligned}
&d((\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}], (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}]) \\
&\quad \leq m((\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2}, \tfrac{1}{2} + \tfrac{1}{k})) + m((\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2}]) \\
&\quad = m((\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2} + \tfrac{1}{k})).
\end{aligned} \tag{7.14}$$

We now claim that (7.14) converges to zero as k grows large. This fact will follow from the continuity of the Lebesgue measure [104], provided we first show that the sets in (7.14) are of finite measure for sufficiently large k. In particular, for $k = 3$, since $p \in L^1(\mathbb{R}^N)$, then

$$1 = \int_{\mathbb{R}^N} p(\eta)\, \mathrm{d}\eta = \lim_{r \to \infty} \int_{B(0,r)} p(\eta)\, \mathrm{d}\eta,$$

where $B(0,r) = \{\eta \in \mathbb{R}^N : \|\eta\| < r\}$. Thus, there exists r_0 such that

$$\int_{B(0,r_0)} p(\eta)\,\mathrm{d}\eta \geq \tfrac{5}{6}.$$

Next, note that for any $x \in \mathbb{R}^N$ such that $\inf_{y\in\mathcal{C}} \|y - x\| \geq r_0$, any $\eta \in B(0,r_0)$ has the property that $x + \eta \notin \mathcal{C}$, since otherwise, we have

$$r_0 \leq \inf_{y\in\mathcal{C}} \|y - x\| \leq \|(x+\eta) - x\| = \|\eta\| < r_0,$$

a contradiction. Thus, if x satisfies $\inf_{y\in\mathcal{C}} \|y - x\| \geq r_0$, then $\chi_{\mathcal{C}}(x+\eta) = 0$ for all $\eta \in B(0,r_0)$ and so:

$$\begin{aligned}
(\chi_{\mathcal{C}} * p)(x) &= \int_{\mathbb{R}^N} \chi_{\mathcal{C}}(x+\eta)p(\eta)\,\mathrm{d}\eta \\
&= \int_{B(0,r_0)^{\mathrm{c}}} \chi_{\mathcal{C}}(x+\eta)p(\eta)\,\mathrm{d}\eta \\
&\leq \int_{B(0,r_0)^{\mathrm{c}}} p(\eta)\,\mathrm{d}\eta \\
&= 1 - \int_{B(0,r_0)} p(\eta)\,\mathrm{d}\eta \\
&\leq 1 - \tfrac{5}{6} \\
&= \tfrac{1}{6}.
\end{aligned}$$

In particular, if $x \in (\chi_{\mathcal{C}} * p)^{-1}(\frac{1}{2} - \frac{1}{3}, \frac{1}{2} + \frac{1}{3})$, then $(\chi_{\mathcal{C}} * p)(x) > \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$, and so such an x does not satisfy $\inf_{y\in\mathcal{C}} \|y - x\| \geq r_0$. In other words:

$$(\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{3}, \tfrac{1}{2} + \tfrac{1}{3}) \subseteq \{x \in \mathbb{R}^N : \exists\, y \in \mathcal{C} \text{ s.t. } \|y - x\| < r_0\}. \tag{7.15}$$

As $\mathcal{C}$ is compact, then the sets in (7.15) are bounded, implying

$$m((\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2} + \tfrac{1}{k})) < \infty$$

for $k = 3$. Next, noting that the sets in (7.14) are nested, that is

$$(\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k+1}, \tfrac{1}{2} + \tfrac{1}{k+1}) \subseteq (\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2} + \tfrac{1}{k}),$$

the continuity of the Lebesgue measure [104] then gives:

$$\begin{aligned}
\lim_{k\to\infty} m((\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2} + \tfrac{1}{k})) &= m\Big(\bigcap_{k=3}^{\infty} (\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2} + \tfrac{1}{k})\Big) \\
&= m\Big((\chi_{\mathcal{C}} * p)^{-1} \bigcap_{k=3}^{\infty} (\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2} + \tfrac{1}{k})\Big) \\
&= m((\chi_{\mathcal{C}} * p)^{-1}\{\tfrac{1}{2}\}) \\
&= 0,
\end{aligned} \tag{7.16}$$

where the final conclusion follows from the assumption that $\mathcal{C}$ is regular. In particular, (7.14) and (7.16) together imply that we may pick $\delta = \frac{1}{k}$ small enough so that:

$$d((\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}], (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}]) \leq m((\chi_{\mathcal{C}} * p)^{-1}(\tfrac{1}{2} - \tfrac{1}{k}, \tfrac{1}{2} + \tfrac{1}{k})) < \varepsilon,$$

for every compact convex set $\hat{\mathcal{C}}$ such that $d(\hat{\mathcal{C}}, C) < \delta$, thus proving our first conclusion (7.8).

For the second conclusion, namely that $\lim_{\hat{\mathcal{C}} \to \mathcal{C}} \mathcal{E}_p(\hat{\mathcal{C}}) = \mathcal{E}_p(\mathcal{C})$, note that

$$d(\mathcal{E}_p(\hat{\mathcal{C}}), \mathcal{E}_p(\mathcal{C})) = m(\mathcal{E}_p(\hat{\mathcal{C}}) \cap \mathcal{E}_p(\mathcal{C})^{\mathrm{c}}) + m(\mathcal{E}_p(\mathcal{C}) \cap \mathcal{E}_p(\hat{\mathcal{C}})^{\mathrm{c}}). \tag{7.17}$$

Since

$$\mathcal{E}_p(\mathcal{C}) = \{x \in \mathcal{C} : P(x + \eta \notin \mathcal{C}) > \tfrac{1}{2}\} = \mathcal{C} \cap (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}],$$

the first term of (7.17) may be rewritten as:

$$\begin{aligned}
&m(\mathcal{E}_p(\hat{\mathcal{C}}) \cap \mathcal{E}_p(\mathcal{C})^{\mathrm{c}}) \\
&= m(\hat{\mathcal{C}} \cap (\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}] \cap (\mathcal{C} \cap (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}])^{\mathrm{c}}) \\
&= m(\hat{\mathcal{C}} \cap (\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}] \cap (\mathcal{C}^{\mathrm{c}} \cup (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}]^{\mathrm{c}})) \\
&\leq m(\hat{\mathcal{C}} \cap (\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}] \cap \mathcal{C}^{\mathrm{c}}) + m(\hat{\mathcal{C}} \cap (\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}] \cap (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}]^{\mathrm{c}}) \\
&\leq m(\hat{\mathcal{C}} \cap \mathcal{C}^{\mathrm{c}}) + m((\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}] \cap (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}]^{\mathrm{c}}).
\end{aligned} \tag{7.18}$$

Adding (7.18) to the inequality obtained by interchanging $\hat{\mathcal{C}}$ and $\mathcal{C}$ in (7.18) gives:

$$d(\mathcal{E}_p(\hat{\mathcal{C}}), \mathcal{E}_p(\mathcal{C})) \leq d(\hat{\mathcal{C}}, \mathcal{C}) + d((\chi_{\hat{\mathcal{C}}} * p)^{-1}[0, \tfrac{1}{2}], (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}]). \tag{7.19}$$

As the first half of (7.19) tends to zero by definition as $\hat{\mathcal{C}}$ approaches $\mathcal{C}$, and the second half of (7.19) tends to zero by our first conclusion (7.8), we have $\lim_{\hat{\mathcal{C}} \to \mathcal{C}} d(\mathcal{E}_p(\hat{\mathcal{C}}), \mathcal{E}_p(\mathcal{C})) = 0$. □

7.2 Frame Sets

In the previous section, we showed that for a convex class $\mathcal{C}$, the points at which classification errors due to noise will occur more often than not, namely $\mathcal{E}_p(\mathcal{C})$, may be approximated by $\mathcal{E}_p(\hat{\mathcal{C}})$, provided $\hat{\mathcal{C}}$ is sufficiently close to $\mathcal{C}$. Our aim is thus to find convenient $\hat{\mathcal{C}}$ such that the performance of a classifier can be analyzed using $\hat{\mathcal{C}}$ instead of an arbitrary $\mathcal{C}$. As $\mathcal{C}$ is convex, convex polytopes are natural candidates for the approximating set $\hat{\mathcal{C}}$. In this section, we show how convex polytopes are a special case of sets that naturally arise in the context of frame theory. We term these sets *frame sets*. In the next section, we shall show how these frame sets may be used to classify elements of $\mathcal{C}$, and study the errors associated with such a classification scheme.

As we have seen in Chapter 4, a sequence of vectors $\{\varphi_m\}_{m=1}^M$ in $\mathbb{R}^N$ is a *frame* for $\mathbb{R}^N$ if they span $\mathbb{R}^N$. Equivalently, $\{\varphi_m\}_{m=1}^M$ is a frame if its analysis operator $\tilde{\Phi}^* : \mathbb{R}^N \to \mathbb{R}^M$, $(\tilde{\Phi}^* x)(m) = \langle x, \tilde{\varphi}_m^* \rangle$ is injective; $\tilde{\Phi}^*$ may be regarded as an $M \times N$ matrix whose mth row is $\tilde{\varphi}_m^*$.

DEFINITION 7.1. Given a frame $\{\tilde{\varphi}_m^*\}_{m=1}^M$ for $\mathbb{R}^N$ and some subset $\Omega \subseteq \mathbb{R}^M$, the corresponding frame set is:

$$(\tilde{\Phi}^*)^{-1}(\Omega) := \{x \in \mathbb{R}^N : \tilde{\Phi}^* x \in \Omega\}.$$

7.2.1 Properties of Frames Sets

Being defined in terms of preimages of sets under the action of a function, frame sets immediately inherit many of convenient set relations often encountered in topology, such as:

$$\begin{aligned}(\tilde{\Phi}^*)^{-1}(\Omega_1 \cup \Omega_2) &= (\tilde{\Phi}^*)^{-1}(\Omega_1) \cup (\tilde{\Phi}^*)^{-1}(\Omega_2),\\ (\tilde{\Phi}^*)^{-1}(\Omega_1 \cap \Omega_2) &= (\tilde{\Phi}^*)^{-1}(\Omega_1) \cap (\tilde{\Phi}^*)^{-1}(\Omega_2),\\ (\tilde{\Phi}^*)^{-1}(\Omega^{\mathrm{c}}) &= ((\tilde{\Phi}^*)^{-1}(\Omega))^{\mathrm{c}}.\end{aligned}$$

More can be said since the functions which generate these preimages are linear:

PROPOSITION 7.3. The translation of a frame set is a frame set of a translation. In particular, for any $x_0 \in \mathbb{R}^N$:

$$x_0 + (\tilde{\Phi}^*)^{-1}(\Omega) = (\tilde{\Phi}^*)^{-1}(\tilde{\Phi}^* x_0 + \Omega). \tag{7.20}$$

Also, if $\{\theta_n\}_{n=1}^N$ is a frame for $\mathbb{R}^P$ with analysis operator $\tilde{\Theta}^*$, then:

$$(\tilde{\Phi}^* \tilde{\Theta}^*)^{-1}(\Omega) = (\tilde{\Theta}^*)^{-1}\left((\tilde{\Phi}^*)^{-1}(\Omega)\right). \tag{7.21}$$

Proof. To prove (7.20), let x_0 be an element of $\mathbb{R}^N$. Then,

$$\begin{aligned}x \in (\tilde{\Phi}^*)^{-1}(\tilde{\Phi}^* x_0 + \Omega) &\Leftrightarrow \tilde{\Phi}^* x \in \tilde{\Phi}^* x_0 + \Omega\\ &\Leftrightarrow \tilde{\Phi}^*(x - x_0) \in \Omega\\ &\Leftrightarrow x - x_0 \in (\tilde{\Phi}^*)^{-1}(\Omega)\\ &\Leftrightarrow x \in x_0 + (\tilde{\Phi}^*)^{-1}(\Omega).\end{aligned}$$

To prove (7.21), note:

$$\begin{aligned}(\tilde{\Phi}^* \tilde{\Theta}^*)^{-1}(\Omega) &= \{x \in \mathbb{R}^P : \tilde{\Phi}^* \tilde{\Theta}^* x \in \Omega\}\\ &= \{x \in \mathbb{R}^P : \tilde{\Theta}^* x \in (\tilde{\Phi}^*)^{-1}(\Omega)\}\\ &= \{x \in \mathbb{R}^P : x \in (\tilde{\Theta}^*)^{-1}\left((\tilde{\Phi}^*)^{-1}(\Omega)\right)\}\\ &= (\tilde{\Theta}^*)^{-1}\left((\tilde{\Phi}^*)^{-1}(\Omega)\right).\end{aligned}$$

□

Proposition 7.3 has interesting consequences. In particular, given Ω, if $M = N$ and $\tilde{\Phi}^*$ is a rotation, (7.21) implies that a frame set of a rotated analysis operator is another frame set of the rotated subset Ω. If $\tilde{\Phi}^*$ were a dilation, we would obtain a similar result.

In general, to compute a frame set explicitly, one needs to find a left-inverse of the analysis operator, that is, an operator $\Phi : \mathbb{R}^M \to \mathbb{R}^N$ such that $\Phi\tilde{\Phi}^* = \mathrm{I}$. To be precise, we have:

PROPOSITION 7.4. If $\Phi\tilde{\Phi}^* = \mathrm{I}$, then:

$$\tilde{\Phi}^*(\tilde{\Phi}^*)^{-1}(\Omega) = \Omega \cap \tilde{\Phi}^*\mathbb{R}^N, \tag{7.22}$$

$$(\tilde{\Phi}^*)^{-1}(\Omega) = \Phi(\Omega \cap \tilde{\Phi}^*\mathbb{R}^N). \tag{7.23}$$

Moreover, if $\tilde{\Phi}^*$ is a Parseval tight frame, that is, $(\tilde{\Phi}^*)^{\mathrm{T}}\tilde{\Phi}^* = \mathrm{I}$, then $(\tilde{\Phi}^*)^{-1}(\Omega)$ is isometric to $\Omega \cap \tilde{\Phi}^*\mathbb{R}^N$.

Proof. For (7.22), note that if $y \in \tilde{\Phi}^*(\tilde{\Phi}^*)^{-1}(\Omega)$, then $y = \tilde{\Phi}^*x$ with $x \in (\tilde{\Phi}^*)^{-1}(\Omega) \subseteq \mathbb{R}^N$, that is, $y \in \Omega$ and $y \in \tilde{\Phi}^*\mathbb{R}^N$. Thus, $\tilde{\Phi}^*(\tilde{\Phi}^*)^{-1}(\Omega) \subseteq \Omega \cap \tilde{\Phi}^*\mathbb{R}^N$. Meanwhile, if $y \in \Omega \cap \tilde{\Phi}^*\mathbb{R}^N$, then $y = \tilde{\Phi}^*x$ for some $x \in \mathbb{R}^N$; since $y \in \Omega$ then $x \in (\tilde{\Phi}^*)^{-1}(\Omega)$, implying $y \in \tilde{\Phi}^*(\tilde{\Phi}^*)^{-1}(\Omega)$. Thus, $\tilde{\Phi}^*(\tilde{\Phi}^*)^{-1}(\Omega) \supseteq \Omega \cap \tilde{\Phi}^*\mathbb{R}^N$.

Next, (7.23) is obtained by taking Φ of (7.22), that is:

$$(\tilde{\Phi}^*)^{-1}(\Omega) = \Phi\tilde{\Phi}^*(\tilde{\Phi}^*)^{-1}(\Omega) = \Phi(\Omega \cap \tilde{\Phi}^*\mathbb{R}^N).$$

Note that the injectivity of $\tilde{\Phi}^*$ along with (7.22) gives that $(\tilde{\Phi}^*)^{-1}(\Omega)$ is isomorphic to $\Omega \cap \tilde{\Phi}^*\mathbb{R}^N$. Moreover, when $\tilde{\Phi}^*$ is Parseval, then for any $x \in \mathbb{R}^N$, the fact that $\|\tilde{\Phi}^*x\|^2 = \langle(\tilde{\Phi}^*)^{\mathrm{T}}\tilde{\Phi}^*x, x\rangle = \langle x, x\rangle = \|x\|^2$ implies the two sets are isometric. □ We

note that neither (7.22) nor (7.23) claims that $(\tilde{\Phi}^*)^{-1}(\Omega)$ is equal to $\Phi(\Omega)$; indeed the second set is larger than the first, in general. In fact, when $\tilde{\Phi}^*$ is Parseval and Φ is chosen to be $(\tilde{\Phi}^*)^{\mathrm{T}}$, the set $\Phi(\Omega)$ is isometric to the orthogonal projection of Ω onto $\tilde{\Phi}^*\mathbb{R}^N$, whereas $(\tilde{\Phi}^*)^{-1}(\Omega)$ is isometric to the intersection of Ω and $\tilde{\Phi}^*\mathbb{R}^N$.

As needed in the next section, the following result shows that a frame set will inherit many of the characteristics of the set Ω which generates it.

PROPOSITION 7.5. For any frame analysis operator $\tilde{\Phi}^*$ and subset Ω of $\mathbb{R}^N$,

a. if Ω is convex, then $(\tilde{\Phi}^*)^{-1}(\Omega)$ is convex,

b. if Ω is closed, then $(\tilde{\Phi}^*)^{-1}(\Omega)$ is closed,

c. if Ω is bounded, then $(\tilde{\Phi}^*)^{-1}(\Omega)$ is bounded.

Proof. If Ω is convex, then for any $x_1, x_2 \in (\tilde{\Phi}^*)^{-1}(\Omega)$,

$$\tilde{\Phi}^*(\lambda x_1 + (1-\lambda)x_2) = \lambda\tilde{\Phi}^*x_1 + (1-\lambda)\tilde{\Phi}^*x_2 \in \Omega,$$

and so $\lambda x_1 + (1-\lambda)x_2 \in (\tilde{\Phi}^*)^{-1}(\Omega)$. Meanwhile, $\tilde{\Phi}^*$, being a linear operator over a finite-dimensional domain, is continuous and therefore Ω being closed implies $(\tilde{\Phi}^*)^{-1}(\Omega)$ is closed. Now assume that Ω is bounded, that is, that there exists a $\rho \geq 0$ such that $\|y\| \leq \rho$ for all $y \in \Omega$. Let $\{\tilde{\varphi}_m^*\}_{m=1}^M$ be the frame corresponding to $\tilde{\Phi}^*$, and let Φ satisfy $\Phi\tilde{\Phi}^* = \mathrm{I}$ and have operator norm $\|\Phi\|$. Then, for all $x \in (\tilde{\Phi}^*)^{-1}(\Omega)$, we have $\tilde{\Phi}^* x \in \Omega$ and so:

$$\|x\| = \|\Phi\tilde{\Phi}^* x\| \leq \|\Phi\|\|\tilde{\Phi}^* x\| \leq \|\Phi\|\rho.$$

□

As a corollary to the previous result, note that if $\tilde{\Phi}^*$ is a frame analysis operator and Ω is compact and convex, the previous result implies that $(\tilde{\Phi}^*)^{-1}(\Omega)$ is also compact and convex.

7.2.2 Convex Polytope Frame Sets

In the special case where $\Omega = \prod_{m=1}^M [a_m, b_m]$, that is a parallelepiped rectangle, then $(\tilde{\Phi}^*)^{-1}(\Omega)$ is a convex polytope:

$$\begin{aligned}(\tilde{\Phi}^*)^{-1}(\Omega) &= \{x \in \mathbb{R}^N : \tilde{\Phi}^* x \in \Omega\} \\ &= \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m],\ m = 1, \ldots, M\} \\ &= \bigcap_{m=1}^M \{x \in \mathbb{R}^N : a_m \leq \langle x, \tilde{\varphi}_m^* \rangle \leq b_m\}. \end{aligned} \tag{7.24}$$

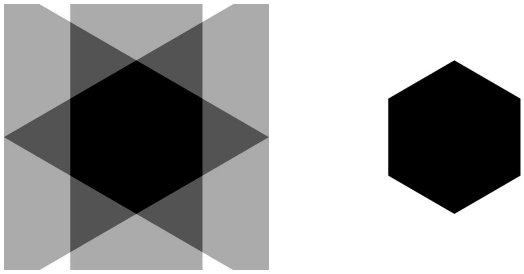
In the following section, we will show how sets of this form may be used to approximate an arbitrary compact convex set $\mathcal{C}$, and propose a classification scheme for $\mathcal{C}$ in terms of $(\tilde{\Phi}^*)^{-1}(\Omega)$. To facilitate this process, we consider the following *decision function*:

Definition 7.2. For any frame $\{\varphi_m\}_{m=1}^M$ of $\mathbb{R}^N$ and any $\Omega = \prod_{m=1}^M [a_m, b_m]$, the associated decision function is $D_{\tilde{\Phi}^*,\Omega} : \mathbb{R}^N \to [0,1]$,

$$D_{\tilde{\Phi}^*,\Omega}(x) := \frac{1}{M} \sum_{m=1}^M \chi_{[a_m,b_m]}(\langle x, \tilde{\varphi}_m^* \rangle).$$

Note that the frame set is equal to the set of points where its decision function is 1, that is, $(\tilde{\Phi}^*)^{-1}(\Omega) = D^{-1}_{\tilde{\Phi}^*,\Omega}(\{1\})$. Moreover, $D_{\tilde{\Phi}^*,\Omega}(x) = \frac{m}{M}$ precisely when x belongs to exactly m hyperbands of the form
$\{x \in \mathbb{R}^N :\ a_m \leq \langle x, \tilde{\varphi}_m^* \rangle \leq b_m\}$.

Example 7.2. Let us choose the set of frame vectors $\tilde{\varphi}_m^* = [\cos\frac{(m-1)\pi}{M}\ \sin\frac{(m-1)\pi}{M}]^{\mathrm{T}}$ in the plane, let $\tilde{\Phi}^*$ be their analysis frame operator and $\Omega = [-1,1]^M$. Then, decision function $D_{\tilde{\Phi}^*,\Omega}$ is depicted in Fig. 7.2(a), where shades of gray correspond to a value between 0 (black) and 1 (white), and represents, how many inequalities in (7.24) are satisfied. Fig. 7.2(b) shows the corresponding frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$ as the region where the decision function is equal to 1. ■



(a) (b)

Figure 7.2: A frame set example for $M = 3$ and $\Omega = [-1, 1]^2$. Each shade of gray corresponds to a value between 0 (black) and 1 (white), and represents, how many inequalities in (7.24) are satisfied. (a) Decision function $D_{\tilde{\Phi}^*,\Omega}$. (b) Frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$.

In the context of classification, frame sets of the form (7.24) provide the additional advantage of easing the decision making process. To be precise, membership in the class $(\tilde{\Phi}^*)^{-1}(\Omega)$ may be decided by independently determining whether each $\langle x, \tilde{\varphi}_m^* \rangle$ belongs to $[a_m, b_m]$, a classification method which is elaborated upon below. In a more general case, for instance, when Ω is a ball, deciding class membership requires the use of multiple frame coefficients at once. For these reasons, the remainder of this work is dedicated to the special case $\Omega = \prod_{m=1}^{M} [a_m, b_m]$.

7.3 Classification of Convex Sets with Frame Sets

In Section 7.1, we considered a classification problem for a convex set $\mathcal{C}$, namely, deciding whether a given $x \in \mathbb{R}^N$ is an element of $\mathcal{C}$, in the presence of noise. However, issues arise in applying this analysis to a real-world problem. For example, one is seldom given the classification set $\mathcal{C}$ explicitly; in reality, it is often approximated via training, that is, using a small set of actual examples of signals which are known to lie in that class. Moreover, even in the ideal case when one has a perfect understanding of $\mathcal{C}$, determining whether a given point lies in the convex set $\mathcal{C}$ may require an arbitrary large amount of computation, as the set may only be expressed as an infinite intersection of half-spaces. This second issue may be partially resolved by approximating $\mathcal{C}$ by convex polytopes, which, as seen in Section 7.2, are frame sets.

In particular, we propose the following classification scheme, in which an arbitrary convex class $\mathcal{C}$ is approximated by a frame set $\hat{\mathcal{C}} = (\tilde{\Phi}^*)^{-1}(\Omega)$. To be precise, given $x \in \mathbb{R}^N$, we shall decide whether $x \in \mathcal{C}$ by instead deciding whether $x \in \hat{\mathcal{C}} = (\tilde{\Phi}^*)^{-1}(\Omega)$, that is, whether $\tilde{\Phi}^* x \in \Omega$. In other words, letting $\{\tilde{\varphi}_m^*\}_{m=1}^M$ be the frame vectors of $\tilde{\Phi}^*$ and $\Omega = \prod_{m=1}^{M} [a_m, b_m]$, we say $x \in \mathcal{C}$ precisely when $a_m \leq \langle x, \tilde{\varphi}_m^* \rangle \leq b_m$ for all $m = 1, \ldots, M$.

7.3.1 Classification with Frame Sets when No Noise is Present

We note that even in the noiseless case, such an approximation will inevitably lead to classification errors whenever $x \in \mathcal{C} \cap (\tilde{\Phi}^*)^{-1}(\Omega)^{\mathrm{c}}$ or $x \in (\tilde{\Phi}^*)^{-1}(\Omega) \cap \mathcal{C}^{\mathrm{c}}$. The frequency of such errors may only be reduced by choosing an alternative $\tilde{\Phi}^*$ and Ω so that the resulting new frame set is closer to $\mathcal{C}$. In the following result, we write an arbitrary compact convex set as the limit of a sequence of frame sets, and show that these errors vanish asymptotically.

THEOREM 7.6. Let $u \in \mathbb{R}^d$, $u \neq 0$ be fixed, and let $\{\tilde{\varphi}_m^*\}_{m=1}^{\infty}$ be any countable dense set in the hemisphere $\{x \in \mathbb{R}^N : \|x\| = 1, \langle x, u\rangle \geq 0\}$. Then for any compact convex set $\mathcal{C}$, there exists a corresponding sequence of intervals $\{[a_m, b_m]\}_{m=1}^{\infty}$ such that

$$\mathcal{C} = \lim_{M\to\infty} \bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^*\rangle \in [a_m, b_m]\}, \tag{7.25}$$

in which the limit is taken with respect to the distance (7.6). In particular, letting $\tilde{\Phi}_M^*$ be the frame analysis operator of $\{\varphi_m\}_{m=1}^M$ and letting $\Omega_M = \prod_{m=1}^M [a_m, b_m]$, we have:

$$\mathcal{C} = \lim_{M\to\infty} (\tilde{\Phi}^*)_M^{-1}(\Omega_M).$$

Before we prove this theorem, let us give an illustrative example of this result.

EXAMPLE 7.3. We go back to our initial example in the plane, and use the disk as the compact convex class $\mathcal{C}$. Let $\tilde{\varphi}_m^* = [\cos\frac{(m-1)\pi}{M}\ \sin\frac{(m-1)\pi}{M}]^{\mathrm{T}}$ be the sequence of frame vectors and $\Omega_M = [-1, 1]^M$. Fig. 7.3 shows the set of misclassified points when we approximate the disk $\mathcal{C}$ with the frame set $(\tilde{\Phi}^*)_M^{-1}(\Omega_M)$ (black), that is, the set of points where the disk and the frame set differ. Note that our choice of Ω_M implies that the frame sets will be approaching the disk from the outside, namely, $\mathcal{C}$ is inscribed in $(\tilde{\Phi}^*)_M^{-1}(\Omega_M)$. However, for a given M, we can choose $\Omega_M = [-a, a]^M$ for $a > 0$ and compute the approximation error $d(C, (\tilde{\Phi}^*)_M^{-1}(\Omega_M))$ as a function of a that can be optimized. Indeed, we have:

$$d(C, (\tilde{\Phi}^*)_M^{-1}([-a, a]^M)) = (2M-1)\pi - 2M\big(2\sin^{-1}(a) + 2a\sqrt{1-a^2} - a^2\tan(\tfrac{\pi}{M})\big),$$

which leads to the expression of the minimal approximation error for the optimal parameter a_0:

$$d(C, (\tilde{\Phi}^*)_M^{-1}([-a_0, a_0]^M)) = (2M-1)\pi - 4M\sin^{-1}(a_0), \tag{7.26}$$

where $a_0 = 2(4 + tan^2(\frac{\pi}{M}))^{-\frac{1}{2}}$. Moreover, we can prove that the error in (7.26) goes to zero as M grows large, illustrating our result in Theorem 7.6. ∎

Proof. We first note that such countable dense sets of vectors indeed exist. In particular, for any positive integer k, the compactness of the hypersphere implies that there exists a finite number of points $\{(\tilde{\varphi}_l^*)^{(k)}\}_{l=1}^{L_k}$ such that for any x with $\|x\| = 1$, we have $\|x - (\tilde{\varphi}_l^*)^{(k)}\| < \frac{1}{k}$ for at least one index $l = 1, \ldots, L_k$. The

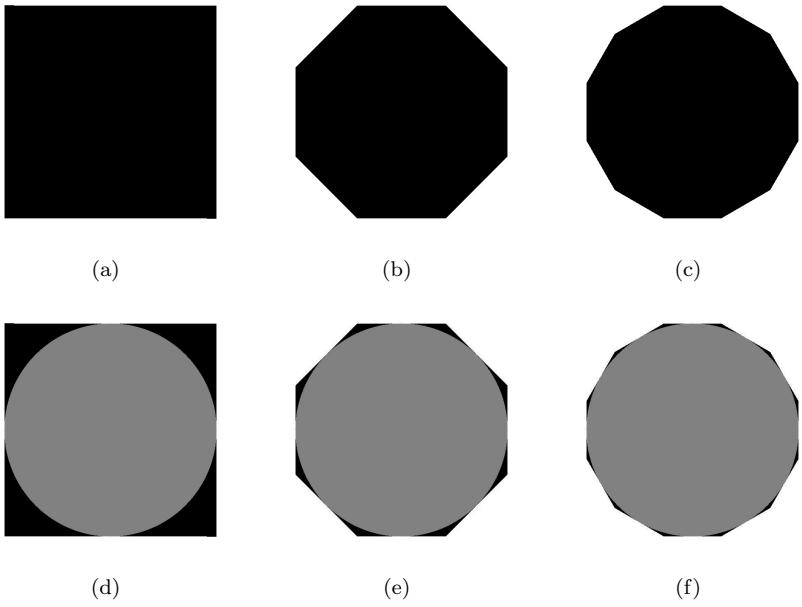


Figure 7.3: Example of frame sets $\hat{\mathcal{C}}$ and corresponding approximation error sets for three values of M, when the class $\mathcal{C}$ is the unit disk. The approximation errors are due to approximating $\mathcal{C}$ by the frame sets $(\tilde{\Phi}^*)^{-1}(\Omega)$ where $\tilde{\Phi}^*$ and Ω are as in Example 7.3. First row: Frame sets (a) $M = 2$, (b) $M = 4$ (c) $M = 6$. Second row: Corresponding approximation error sets in black and convex class $\mathcal{C}$ in a medium shade of gray (d) $M = 2$, (e) $M = 4$, (f) $M = 6$.

concatenation $\{\tilde{\varphi}_m^*\}_{m=1}^{\infty}$ of the sequences $\{(\tilde{\varphi}_l^*)^{(k)}\}_{l=1}^{L_k}$ over all $k \geq 1$ is then a countable set which is dense in the whole sphere, and as such, is dense in any hemisphere.

For any m, let

$$a_m = \min_{x \in \mathcal{C}} \langle x, \tilde{\varphi}_m^* \rangle, \quad b_m = \max_{x \in \mathcal{C}} \langle x, \tilde{\varphi}_m^* \rangle.$$

Then, for any $x \in \mathcal{C}$, we immediately have that $\langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]$ for all m, that is,

$$\mathcal{C} \subseteq \bigcap_{m=1}^{\infty} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\}, \tag{7.27}$$

To prove equality in (7.27), note that if $x \notin \mathcal{C}$, then the fact that $\mathcal{C}$ is a compact convex set implies there exists $y \in \mathbb{R}^N$, $\|y\| = 1$, such that $\langle z - x, y \rangle > 0$ for all $z \in \mathcal{C}$. Letting $\alpha = \min_{z \in \mathcal{C}} \langle z - x, y \rangle$, we have $\alpha > 0$. Since $\{\tilde{\varphi}_m^*\}_{m=1}^{\infty}$ is dense in a

hemisphere of $\{x \in \mathbb{R}^N : \|x\| = 1\}$, there exists m_0 such that:

$$\min\big\{\|y - \tilde{\varphi}^*_{m_0}\|, \|y + \tilde{\varphi}^*_{m_0}\|\big\} \leq \frac{\alpha}{2\max\limits_{z\in\mathcal{C}}\|z - x\|}. \tag{7.28}$$

In particular, if $\|y - \tilde{\varphi}^*_{m_0}\| \leq \dfrac{\alpha}{2\max\limits_{z\in\mathcal{C}}\|z - x\|}$, then for any $z \in \mathcal{C}$,

$$\begin{aligned}
\langle z - x, \tilde{\varphi}^*_{m_0}\rangle &= \langle z - x, y\rangle - \langle z - x, y - \tilde{\varphi}^*_{m_0}\rangle \\
&\geq \alpha - \|z - x\|\,\|y - \tilde{\varphi}^*_{m_0}\| \\
&\geq \alpha - \max_{\tilde{z}\in\mathcal{C}}\|\tilde{z} - x\|\frac{\alpha}{2\max\limits_{\tilde{z}\in\mathcal{C}}\|\tilde{z} - x\|} \\
&= \tfrac{\alpha}{2},
\end{aligned}$$

that is, $\langle z, \tilde{\varphi}^*_{m_0}\rangle \geq \frac{\alpha}{2} + \langle x, \tilde{\varphi}^*_{m_0}\rangle$ for all $z \in \mathcal{C}$. Thus, in this case we have:

$$a_{m_0} = \min_{z\in\mathcal{C}}\langle z, \tilde{\varphi}^*_{m_0}\rangle \geq \tfrac{\alpha}{2} + \langle x, \tilde{\varphi}^*_{m_0}\rangle > \langle x, \tilde{\varphi}^*_{m_0}\rangle,$$

implying $x \notin \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}^*_{m_0}\rangle \in [a_{m_0}, b_{m_0}]\}$.

Meanwhile, in the case where $\|y + f_{m_0}\| \leq \dfrac{\alpha}{2\max\limits_{z\in\mathcal{C}}\|z - x\|}$, then for any $z \in \mathcal{C}$,

$$\begin{aligned}
\langle z - x, \tilde{\varphi}^*_{m_0}\rangle &= \langle z - x, y + \tilde{\varphi}^*_{m_0}\rangle - \langle z - x, y\rangle \\
&\leq \|z - x\|\,\|y - \tilde{\varphi}^*_{m_0}\| - \alpha \\
&\leq \max_{\tilde{z}\in\mathcal{C}}\|\tilde{z} - x\|\frac{\alpha}{2\max\limits_{\tilde{z}\in\mathcal{C}}\|\tilde{z} - x\|} - \alpha \\
&= -\tfrac{\alpha}{2},
\end{aligned}$$

that is, $\langle z, \tilde{\varphi}^*_{m_0}\rangle \leq \langle x, \tilde{\varphi}^*_{m_0}\rangle - \frac{\alpha}{2}$ for all $z \in \mathcal{C}$. Thus, in this case we have:

$$b_{m_0} = \max_{z\in\mathcal{C}}\langle z, \tilde{\varphi}^*_{m_0}\rangle \leq \langle x, \tilde{\varphi}^*_{m_0}\rangle - \tfrac{\alpha}{2} < \langle x, \tilde{\varphi}^*_{m_0}\rangle,$$

which again implies $x \notin \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}^*_{m_0}\rangle \in [a_{m_0}, b_{m_0}]\}$. To summarize, if $x \notin \mathcal{C}$, then (7.28) holds, which, regardless of whether $\|y - \tilde{\varphi}^*_{m_0}\|$ or $\|y + \tilde{\varphi}^*_{m_0}\|$ is the smaller quantity, implies that:

$$x \notin \bigcap_{m=1}^{\infty}\{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}^*_m\rangle \in [a_m, b_m]\},$$

Thus, $\mathcal{C} = \bigcap\limits_{m=1}^{\infty}\{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}^*_m\rangle \in [a_m, b_m]\}$.

To prove (7.25), note that since $\mathcal{C} \subseteq \bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\}$, then:

$$\begin{aligned} & d\Big(\mathcal{C}, \bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\}\Big) \\ & \quad = m(\emptyset) + m\Big(\bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\} \cap \mathcal{C}^{\mathrm{c}}\Big) \\ & \quad = m\Big(\bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\} \cap \mathcal{C}^{\mathrm{c}}\Big). \end{aligned} \tag{7.29}$$

We claim that the sets in (7.29) have finite measure when M is sufficiently large. Indeed, as $\{\tilde{\varphi}_m^*\}_{m=1}^{\infty}$ is dense in the hemisphere, then there exists M_0 such that $\{\tilde{\varphi}_m^*\}_{m=1}^{M_0}$ spans $\mathbb{R}^N$, that is, is a frame for $\mathbb{R}^N$. Letting $\Omega_{M_0} = \prod_{m=1}^{M_0} [a_m, b_m]$, the third statement of Proposition 7.5 gives that since Ω is bounded, then

$$(\tilde{\Phi}^*)_{M_0}^{-1}(\Omega_{M_0}) = \bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\}$$

is bounded. In particular, the measures in (7.29) are finite whenever $M \geq M_0$. Moreover, as the sets in (7.29) are nested, that is,

$$\bigcap_{m=1}^{M+1} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\} \cap \mathcal{C}^{\mathrm{c}} \subseteq \bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\} \cap \mathcal{C}^{\mathrm{c}},$$

the continuity of the Lebesgue measure implies:

$$\begin{aligned} & \lim_{M \to \infty} d\Big(\mathcal{C}, \bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\}\Big) \\ & \quad = \lim_{M \to \infty} m\Big(\bigcap_{m=1}^{M} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\} \cap \mathcal{C}^{\mathrm{c}}\Big) \\ & \quad = m\Big(\bigcap_{m=1}^{\infty} \{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\} \cap \mathcal{C}^{\mathrm{c}}\Big) \\ & \quad = m(\mathcal{C} \cap \mathcal{C}^{\mathrm{c}}) \\ & \quad = m(\emptyset) \\ & \quad = 0. \end{aligned}$$

□

7.3.2 Classification with Frame Sets in the Presence of Noise

The above result shows that in the noiseless case, frame sets may be used to approximate any compact convex set to within any degree of accuracy. In the noisy case,

additional classification errors will occur. Errors of this type are unavoidable, and limit the performance of any classification scheme. As it is impossible to eliminate such errors, we instead focus on characterizing those points which are most in danger to be misclassified due to noise. Indeed, as discussed in Section 7.1, the error set $\mathcal{E}_p(\mathcal{C})$ for a given convex set $\mathcal{C}$ and noise model p, given in (7.5), is the set of all points which are misclassified more than half of the time. We now refine this error set concept further, so as to include both errors due to noise and those that arise in approximating $\mathcal{C}$ by a frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$. Specifically, we let $\mathcal{E}_p(\mathcal{C}, (\tilde{\Phi}^*)^{-1}(\Omega))$ be the set of all $x \in \mathbb{R}^N$ for which either Type-I or II errors arise more than half of the time while attempting to determine membership in class $\mathcal{C}$ using its approximation $(\tilde{\Phi}^*)^{-1}(\Omega)$:

$$\begin{aligned}\mathcal{E}_p(\mathcal{C}, (\tilde{\Phi}^*)^{-1}(\Omega)) := &\{x \in C : P(x + \eta \notin (\tilde{\Phi}^*)^{-1}(\Omega)) > \tfrac{1}{2}\} \\ &\cup \{x \notin C : P(x + \eta \in (\tilde{\Phi}^*)^{-1}(\Omega)) > \tfrac{1}{2}\}. \end{aligned} \tag{7.30}$$

That is, $\mathcal{E}_p(\mathcal{C}, (\tilde{\Phi}^*)^{-1}(\Omega))$ is the set of those points where we expect to err when using our decision rule, namely deciding $x \in \mathcal{C}$ whenever $\tilde{\Phi}^* x \in \Omega$. Note that when $\mathcal{C} = (\tilde{\Phi}^*)^{-1}(\Omega)$, this definition generalizes (7.5), namely $\mathcal{E}_p(\mathcal{C}, \mathcal{C}) = \mathcal{E}_p(\mathcal{C})$. More importantly, the next result shows that $\mathcal{E}_p(\mathcal{C}, (\tilde{\Phi}^*)^{-1}(\Omega))$ will asymptotically approximate the intrinsic error in $\mathcal{C}$ due to noise, namely $\mathcal{E}_p(\mathcal{C})$, as the frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$ is taken ever closer to $\mathcal{C}$.

THEOREM 7.7. Let p be a radially symmetric noise model, let $\mathcal{C}$ be a compact convex set which is regular (7.7), and let $\{(\tilde{\Phi}_k^*)^{-1}(\Omega_k)\}_{k=1}^{\infty}$ be any sequence of frame sets where $(\tilde{\Phi}^*)_k$ is a $M_k \times N$ frame analysis operator, $\Omega_k = \prod_{m=1}^{M_k}[a_{m_k}, b_{m_k}]$, and

$$\mathcal{C} = \lim_{k\to\infty} (\tilde{\Phi}_k^*)^{-1}(\Omega_k).$$

Then, the error sets of $(\tilde{\Phi}_k^*)^{-1}(\Omega_k)$ converge to the error set of $\mathcal{C}$:

$$\mathcal{E}_p(\mathcal{C}) = \lim_{k\to\infty} \mathcal{E}_p((\tilde{\Phi}_k^*)^{-1}(\Omega_k)), \tag{7.31}$$

and furthermore the decision rule "decide $x \in \mathcal{C}$ if $(\tilde{\Phi}^*)x \in \Omega$" asymptotically attains the best classification accuracy possible in the presence of noise:

$$\mathcal{E}_p(\mathcal{C}) = \lim_{k\to\infty} \mathcal{E}_p(\mathcal{C}, (\tilde{\Phi}_k^*)^{-1}(\Omega_k)). \tag{7.32}$$

Proof. To prove (7.31), let $\{(\tilde{\Phi}_k^*)^{-1}(\Omega_k)\}_{k=1}^{\infty}$ be any sequence of frame sets such that $\mathcal{C} = \lim_{k\to\infty}(\tilde{\Phi}_k^*)^{-1}(\Omega_k)$. Note that since $\Omega_k = \prod_{m=1}^{M_k}[a_{m_k}, b_{m_k}]$ is compact and convex, Proposition 7.5 gives that each set $(\tilde{\Phi}_k^*)^{-1}(\Omega_k)$ is compact and convex. Thus, for any radially symmetric noise model p, the continuity of $\mathcal{E}_p(\mathcal{C})$, as given in Theorem 7.2, immediately implies our first result (7.31):

$$\lim_{k\to\infty} \mathcal{E}_p((\tilde{\Phi}_k^*)^{-1}(\Omega_k)) = \lim_{\hat{\mathcal{C}}\to\mathcal{C}} \mathcal{E}_p(\hat{\mathcal{C}}) = \mathcal{E}_p(\mathcal{C}).$$

Next, to show (7.32), we, in a manner similar to the proof of Theorem 7.2, may write the error set of $\mathcal{C}$, abbreviated to $\mathcal{E}_1$, as:

$$\mathcal{E}_1 := \mathcal{E}_p(\mathcal{C}) = \{x \in \mathcal{C} : P(x + \eta \notin C) > \tfrac{1}{2}\} = \mathcal{C} \cap (\chi_{\mathcal{C}} * p)^{-1}[0, \tfrac{1}{2}].$$

Similarly, we write $\mathcal{E}_p(\mathcal{C}, (\tilde{\Phi}_k^*)^{-1}(\Omega_k)) = \mathcal{E}_2 \cup \mathcal{E}_3$, where:

$$\mathcal{E}_2 := \{x \in \mathcal{C} : P(x + \eta \notin (\tilde{\Phi}_k^*)^{-1}(\Omega_k)) > \tfrac{1}{2}\} = \mathcal{C} \cap (\chi_{(\tilde{\Phi}_k^*)^{-1}(\Omega_k)} * p)^{-1}[0, \tfrac{1}{2}],$$
$$\mathcal{E}_3 := \{x \notin \mathcal{C} : P(x + \eta \in (\tilde{\Phi}_k^*)^{-1}(\Omega_k)) > \tfrac{1}{2}\} = \mathcal{C}^{\mathrm{c}} \cap (\chi_{(\tilde{\Phi}_k^*)^{-1}(\Omega_k)} * p)^{-1}(\tfrac{1}{2}, 1].$$

Thus, under this notation:

$$\begin{aligned} d(\mathcal{E}_p(\mathcal{C}), \mathcal{E}_p(\mathcal{C}, (\tilde{\Phi}_k^*)^{-1}(\Omega_k))) &= d(\mathcal{E}_1, \mathcal{E}_2 \cup \mathcal{E}_3) \\ &= m(\mathcal{E}_1 \cap (\mathcal{E}_2 \cup \mathcal{E}_3)^{\mathrm{c}}) + m((\mathcal{E}_2 \cup \mathcal{E}_3) \cap \mathcal{E}_1^{\mathrm{c}}) \\ &= m(\mathcal{E}_1 \cap \mathcal{E}_2^{\mathrm{c}} \cap \mathcal{E}_3^{\mathrm{c}}) + m((\mathcal{E}_2 \cap \mathcal{E}_1^{\mathrm{c}}) \cup (\mathcal{E}_3 \cap \mathcal{E}_1^{\mathrm{c}})) \\ &\leq m(\mathcal{E}_1 \cap \mathcal{E}_2^{\mathrm{c}}) + m(\mathcal{E}_2 \cap \mathcal{E}_1^{\mathrm{c}}) + m(\mathcal{E}_3 \cap \mathcal{E}_1^{\mathrm{c}}) \\ &= d(\mathcal{E}_1, \mathcal{E}_2) + m(\mathcal{E}_3). \end{aligned} \tag{7.33}$$

Thus, it suffices to show that both $d(\mathcal{E}_1, \mathcal{E}_2)$ and $m(\mathcal{E}_3)$ tend to zero as k grows large. To prove the former, note that letting $\mathcal{S}_1 = (\chi_{\mathcal{C}} * p)^{-1}[0, \frac{1}{2}]$ and $\mathcal{S}_2 = (\chi_{(\tilde{\Phi}_k^*)^{-1}(\Omega_k)} * p)^{-1}[0, \frac{1}{2}]$,

$$\begin{aligned} d(\mathcal{E}_1, \mathcal{E}_2) &= d(\mathcal{C} \cap \mathcal{S}_1, \mathcal{C} \cap \mathcal{S}_2) \\ &= m(\mathcal{C} \cap \mathcal{S}_1 \cap (\mathcal{C} \cap \mathcal{S}_2)^{\mathrm{c}}) + m(\mathcal{C} \cap \mathcal{S}_2 \cap (\mathcal{C} \cap \mathcal{S}_1)^{\mathrm{c}}) \\ &= m(\mathcal{C} \cap \mathcal{S}_1 \cap (\mathcal{C}^{\mathrm{c}} \cup \mathcal{S}_2^{\mathrm{c}})) + m(\mathcal{C} \cap \mathcal{S}_2 \cap (\mathcal{C}^{\mathrm{c}} \cup \mathcal{S}_1^{\mathrm{c}})) \\ &= m(\mathcal{C} \cap \mathcal{S}_1 \cap \mathcal{S}_2^{\mathrm{c}}) + m(\mathcal{C} \cap \mathcal{S}_2 \cap \mathcal{S}_1^{\mathrm{c}}) \\ &\leq m(\mathcal{S}_1 \cap \mathcal{S}_2^{\mathrm{c}}) + m(\mathcal{S}_2 \cap \mathcal{S}_1^{\mathrm{c}}) \\ &= d(\mathcal{S}_1, \mathcal{S}_2), \end{aligned}$$

which tends to zero by (7.8) of Theorem 7.2. Meanwhile, by applying Lemma 7.1 to $(\tilde{\Phi}_k^*)^{-1}(\Omega_k)$, we see that $(\chi_{(\tilde{\Phi}_k^*)^{-1}(\Omega_k)} * p)^{-1}(\frac{1}{2}, 1] \subseteq (\tilde{\Phi}_k^*)^{-1}(\Omega_k)$, and so $m(\mathcal{E}_3)$ will also tend to zero:

$$\begin{aligned} m(E_3) &= m(\mathcal{C}^{\mathrm{c}} \cap (\chi_{(\tilde{\Phi}_k^*)^{-1}(\Omega_k)} * p)^{-1}(\tfrac{1}{2}, 1]) \\ &\leq m(\mathcal{C}^{\mathrm{c}} \cap (\tilde{\Phi}_k^*)^{-1}(\Omega_k)) \\ &\leq d((\tilde{\Phi}_k^*)^{-1}(\Omega_k), \mathcal{C}). \end{aligned}$$

□

Fig. 7.4 illustrates the result in (7.31) for a given white Gaussian noise with mean zero and standard deviation $\sigma = 0.25$. The error sets of the frame sets for different values of M ($M = 2, M = 4$ and $M = 6$) converge toward the error set of the disk $\mathcal{C}$.

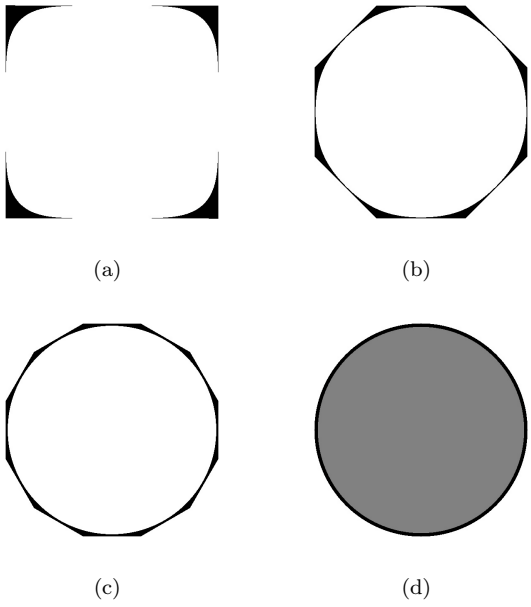


Figure 7.4: Error sets $\mathcal{E}_p((\tilde{\Phi}^*)^{-1}(\Omega))$ for $M = 2$ (a), $M = 4$ (b), $M = 6$ (c) and $\mathcal{E}_p(\mathcal{C})$ (d), where p is a Gaussian noise model of mean zero and standard deviation $\sigma = 0.25$. We see that as M increases, $\mathcal{E}_p((\tilde{\Phi}^*)^{-1}(\Omega))$ better approximates the error set of the data set $\mathcal{C}$.

7.4 Estimating the Classification Error of Frame Sets

In the previous section we showed how an arbitrary compact convex set $\mathcal{C}$ may be approximated to within an arbitrary precision by a frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$, specifically those of the form $\Omega = \prod_{m=1}^{M}[a_m, b_m]$. We further showed that when $\mathcal{C}$ is regular, the set of points we expect to misclassify due to noise, namely $\mathcal{E}_p(\mathcal{C})$, is well-approximated by the analogous error set of a frame set approximation of $\mathcal{C}$, namely a set $\mathcal{E}_p((\tilde{\Phi}^*)^{-1}(\Omega))$ where $(\tilde{\Phi}^*)^{-1}(\Omega)$ is close to $\mathcal{C}$ in measure. That is, the properties of $\mathcal{E}_p(\mathcal{C})$ may be understood by studying the less complicated sets $\mathcal{E}_p((\tilde{\Phi}^*)^{-1}(\Omega))$, which are the subject of this section.

For any frame $\{\tilde{\varphi}_m^*\}_{m=1}^{M}$ of $\mathbb{R}^N$, any $\Omega = \prod_{m=1}^{M}[a_m, b_m]$, any noise model p, and any fixed $x \in \mathbb{R}^N$, let $D_{\tilde{\Phi}^*,\Omega}$ be the decision function defined in (7.2). Then, the expected value of this function evaluated at the noisy point $x + \eta$ is:

$$\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta)) := \frac{1}{M}\sum_{m=1}^{M}\int_{\mathbb{R}^N} \chi_{[a_m,b_m]}(\langle x+\eta, \tilde{\varphi}_m^*\rangle)p(\eta)\,\mathrm{d}\eta. \tag{7.34}$$

When p is radially symmetric, and in particular, when p is Gaussian, the next result shows that this expected value may be computed more explicitly.

THEOREM 7.8. For any frame $\{\tilde{\varphi}_m^*\}_{m=1}^M$ of $\mathbb{R}^N$, any $\Omega = \prod_{m=1}^M [a_m, b_m]$, and any $x \in \mathbb{R}^N$:

a. If p is radially symmetric, then writing $\eta = (\eta_1, \ldots, \eta_N)$, we have:

$$\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta)) = \frac{1}{M}\sum_{m=1}^{M} \int_{\mathbb{R}^N} \chi_{[a_m,b_m]-\langle x,\tilde{\varphi}_m^*\rangle}(\|\tilde{\varphi}_m^*\|\eta_1) p(\eta)\,\mathrm{d}\eta. \tag{7.35}$$

b. If $\eta \sim \mathcal{N}(0, \sigma^2\mathrm{I})$, that is, if $p(\eta) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp(-\frac{\|\eta\|^2}{2\sigma^2})$, then:

$$\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta)) = \frac{1}{2M}\sum_{m=1}^{M} \Big[\mathrm{erf}\big(\tfrac{b_m-\langle x,\tilde{\varphi}_m^*\rangle}{\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma}\big) - \mathrm{erf}\big(\tfrac{a_m-\langle x,\tilde{\varphi}_m^*\rangle}{\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma}\big)\Big], \tag{7.36}$$

where $\mathrm{erf}(x) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^x \exp(-t^2)\,\mathrm{d}t$.

Proof. To prove (7.35), for any $m = 1, \ldots, M$ let U_m be an orthogonal matrix such that $U_m\tilde{\varphi}_m^* = \|\tilde{\varphi}_m^*\| e_1$, where e_1 is the first vector in the standard basis for $\mathbb{R}^N$. Thus, making the change of variables $\eta = U_m^{\mathrm{T}} u$ in the mth summand of (7.34) gives (7.35):

$$\begin{aligned}
&\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta)) \\
&= \frac{1}{M}\sum_{m=1}^{M} \int_{\mathbb{R}^N} \chi_{[a_m,b_m]}(\langle x, \tilde{\varphi}_m^*\rangle + \langle \eta, \tilde{\varphi}_m^*\rangle) p(\eta)\,\mathrm{d}\eta \\
&= \frac{1}{M}\sum_{m=1}^{M} \int_{\mathbb{R}^N} \chi_{[a_m,b_m]-\langle x,\tilde{\varphi}_m^*\rangle}(\langle U_m^{\mathrm{T}} u, \tilde{\varphi}_m^*\rangle) p(U_m^{\mathrm{T}} u) |\det(U_m^{\mathrm{T}})|\,\mathrm{d}u \\
&= \frac{1}{M}\sum_{m=1}^{M} \int_{\mathbb{R}^N} \chi_{[a_m,b_m]-\langle x,\tilde{\varphi}_m^*\rangle}(\langle u, U_m\tilde{\varphi}_m^*\rangle) p(u)\,\mathrm{d}u \\
&= \frac{1}{M}\sum_{m=1}^{M} \int_{\mathbb{R}^N} \chi_{[a_m,b_m]-\langle x,\tilde{\varphi}_m^*\rangle}(\|\tilde{\varphi}_m^*\| u_1) p(u)\,\mathrm{d}u.
\end{aligned}$$

Next, as the Gaussian distribution $\eta \sim \mathcal{N}(0, \sigma^2\mathrm{I})$ is radially symmetric, we may

apply (7.35) in this special case, and obtain:

$$\begin{aligned}
&\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+w)) \\
&= \frac{1}{M}\sum_{m=1}^{M}\int_{\mathbb{R}^N} \chi_{[a_m,b_m]-\langle x,\tilde{\varphi}_m^*\rangle}(\|\tilde{\varphi}_m^*\|u_1)(2\pi\sigma^2)^{-\frac{N}{2}}\exp(-\tfrac{\|u\|^2}{2\sigma^2})\,\mathrm{d}u \\
&= \frac{1}{M}\sum_{m=1}^{M}\Big\{(2\pi\sigma^2)^{-\frac{1}{2}}\int_{\mathbb{R}} \chi_{[a_m,b_m]-\langle x,\tilde{\varphi}_m^*\rangle}(\|\tilde{\varphi}_m^*\|u_1)\exp(-\tfrac{u_1^2}{2\sigma^2})\,\mathrm{d}u_1 \\
&\qquad \times \prod_{n=2}^{N}\Big[(2\pi\sigma^2)^{-\frac{1}{2}}\int_{\mathbb{R}}\exp(-\tfrac{u_n^2}{2\sigma^2})\,\mathrm{d}u_n\Big]\Big\} \\
&= \frac{1}{M\sigma\sqrt{2\pi}}\sum_{m=1}^{M}\int_{\mathbb{R}} \chi_{[a_m,b_m]-\langle x,\tilde{\varphi}_m^*\rangle}(\|\tilde{\varphi}_m^*\|u_1)\exp(-\tfrac{u_1^2}{2\sigma^2})\,\mathrm{d}u_1.
\end{aligned}$$

Making the change of variables $t = \frac{u_1}{\sqrt{2}\sigma}$, we continue simplifying the above expression to obtain (7.36):

$$\begin{aligned}
\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+w)) &= \frac{1}{M\sqrt{\pi}}\sum_{m=1}^{M}\int_{\mathbb{R}} \chi_{[a_m,b_m]-\langle x,\tilde{\varphi}_m^*\rangle}(\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma t)\exp(-t^2)\,\mathrm{d}t \\
&= \frac{1}{M\sqrt{\pi}}\sum_{m=1}^{M}\int_{(a_m-\langle x,\tilde{\varphi}_m^*\rangle)/(\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma)}^{(b_m-\langle x,\tilde{\varphi}_m^*\rangle)/(\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma)} \exp(-t^2)\,\mathrm{d}t \\
&= \frac{1}{2M}\sum_{m=1}^{M}\Big[\mathrm{erf}\big(\tfrac{b_m-\langle x,\tilde{\varphi}_m^*\rangle}{\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma}\big) - \mathrm{erf}\big(\tfrac{a_m-\langle x,\tilde{\varphi}_m^*\rangle}{\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma}\big)\Big].
\end{aligned}$$

□

7.4.1 Bounds on the Total Classification Error

Given $x \in \mathbb{R}^N$, a frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$ and an additive noise η, $P(x+\eta \notin (\tilde{\Phi}^*)^{-1}(\Omega))$ represents the probability that the additive noise η perturbs x into being an element of this frame set (see (7.1)).

In the following, we derive an upper bound on $P(x+\eta \notin (\tilde{\Phi}^*)^{-1}(\Omega))$. This bound depends on the expected value of the decision function $D_{\tilde{\Phi}^*,\Omega}(x+\eta)$. Namely, we have:

PROPOSITION 7.9. Given a frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$ with $\Omega = \prod_{m=1}^{M}[a_m, b_m]$, any $x \in \mathbb{R}^N$ and any additive noise η,

$$P(x+\eta \notin (\tilde{\Phi}^*)^{-1}(\Omega)) \leq M(1 - \mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta))). \tag{7.37}$$

Proof. We demonstrate the above using two different arguments: the first one relies on the subadditivity of probability functions, whereas the second one is based on

the expression of the expected value of the decision function. The first argument is:

$$\begin{aligned}
P(x+\eta \notin (\tilde{\Phi}^*)^{-1}(\Omega)) &= P(\tilde{\Phi}^*(x+\eta) \notin \Omega) \\
&= P(\bigcup_{m=1}^{M} \{\eta : \langle x+\eta, \tilde{\varphi}_m^* \rangle \notin [a_m, b_m]\}) \\
&\leq \sum_{m=1}^{M} P(\langle x+\eta, \tilde{\varphi}_m^* \rangle \notin [a_m, b_m]) \\
&= M - \sum_{m=1}^{M} P(\langle x+w, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]) \\
&= M - \sum_{m=1}^{M} \int_{\mathbb{R}^N} \chi_{[a_m,b_m]}(\langle x+\eta, \tilde{\varphi}_m^* \rangle) p(\eta)\, \mathrm{d}\eta \\
&= M - M\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta)).
\end{aligned}$$

For the second method, recall that the decision function $D_{\tilde{\Phi}^*,\Omega}$ takes values in $\{\frac{m}{M}\}_{m=0}^{M}$, where $P(D_{\tilde{\Phi}^*,\Omega}(x+\eta)=1) = P(x+\eta \in (\tilde{\Phi}^*)^{-1}(\Omega))$. Thus,

$$\begin{aligned}
\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta)) &= \sum_{m=0}^{M} \tfrac{m}{M} P(D_{\tilde{\Phi}^*,\Omega}(x+\eta) = \tfrac{m}{M}) \\
&\leq P(D_{\tilde{\Phi}^*,\Omega}(x+\eta)=1) + \tfrac{M-1}{M} \sum_{m=0}^{M-1} P(D_{\tilde{\Phi}^*,\Omega}(x+\eta) = \tfrac{m}{M}) \\
&= P(x+\eta \in (\tilde{\Phi}^*)^{-1}(\Omega)) + \tfrac{M-1}{M} P(x+\eta \notin (\tilde{\Phi}^*)^{-1}(\Omega)) \\
&= 1 - \tfrac{1}{M} P(x+\eta \notin (\tilde{\Phi}^*)^{-1}(\Omega)),
\end{aligned}$$

which is equivalent to the desired result. □

We presented two arguments above, as each may be generalized in a different way; the first using the inclusion/exclusion principle, and the second using the variance and other higher-order moments. Each method shows some promise of obtaining a bound tighter than (7.37).

When Ω is convex, the total classification error of $(\tilde{\Phi}^*)^{-1}(\Omega)$ is:

$$\mathrm{Er}((\tilde{\Phi}^*)^{-1}(\Omega), p) = m(\{x \in (\tilde{\Phi}^*)^{-1}(\Omega) : P(x+\eta \notin (\tilde{\Phi}^*)^{-1}(\Omega)) > \tfrac{1}{2}\}.$$

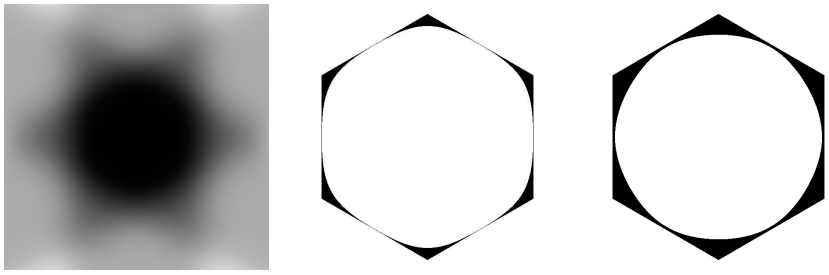
Using Proposition 7.9, we derive the following bounds on the classification error of a frame set:

COROLLARY 7.10. Given a convex decision set Ω and any noise model p,

$$\mathrm{Er}((\tilde{\Phi}^*)^{-1}(\Omega), p) \leq m(\{x \in (\tilde{\Phi}^*)^{-1}(\Omega) : \mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta)) < \tfrac{2M-1}{2M}\}).$$

If p is a radially symmetric Gaussian distribution, then

$$\mathrm{Er}((\tilde{\Phi}^*)^{-1}(\Omega), p) \leq m(\{x \in \mathbb{R}^N : \sum_{m=1}^{M} \Big[\mathrm{erf}\big(\tfrac{b_m - \langle x, \tilde{\varphi}_m^* \rangle}{\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma}\big) - \mathrm{erf}\big(\tfrac{a_m - \langle x, \tilde{\varphi}_m^* \rangle}{\|\tilde{\varphi}_m^*\|\sqrt{2}\sigma}\big)\Big] < 2M-1\}).$$



(a) (b) (c)

Figure 7.5: Classification error set and estimating bound for Gaussian noise with $\sigma = 1$. (a) The expected value of the decision function $\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+\eta))$, (b) The error set $\mathcal{E}_p((\tilde{\Phi}^*)^{-1}(\Omega))$ (right-hand side of (7.38)), (c) Upper bound on $\mathcal{E}_p((\tilde{\Phi}^*)^{-1}(\Omega))$ (left-hand side of (7.38)). We clearly see that the set of points in (b) is a subset of the points in (c).

As illustrated in the example below, Corollary 7.10 provides a quick numerical mechanism for estimating the set of those points for which our frame set-based classifier will fail more often than not. However, we note that the bounds in Corollary 7.10 will become increasingly useless as M grows large. Indeed, as seen in the proof of Proposition 7.9, one should only expect these bounds to be good whenever for most x, the addition of the noise η causes misclassification due to having $x+\eta$ leave but a single hyperband of the form $\{x \in \mathbb{R}^N : \langle x, \tilde{\varphi}_m^* \rangle \in [a_m, b_m]\}$. In particular, one should expect these bounds to perform poorly in the corners of the frame set, namely wherever two or more of the boundary-defining hyperplanes meet.

EXAMPLE 7.4. We continue our running example here and assume p is a Gaussian noise model of mean zero and standard deviation $\sigma = 1$, and let the sequence $\tilde{\varphi}_m^* = [\cos\frac{(m-1)\pi}{M} \ \sin\frac{(m-1)\pi}{M}]^{\mathrm{T}}$ be the frame vectors corresponding to the analysis frame operator $\tilde{\Phi}^*$. Let us choose $\Omega = [-a_0, a_0]^M$, where $a_0 = 2(4 + tan^2(\frac{\pi}{2M}))^{-\frac{1}{2}}$ is the parameter that optimizes the error due to the approximation of $\mathcal{C}$ by the frame set $(\tilde{\Phi}^*)^{-1}(\Omega)$ (7.26). Then for any $\eta \sim \mathcal{N}(0, \sigma^2 I)$, we use (7.36) to obtain:

$$\mathbb{E}(D_{\tilde{\Phi}^*,\Omega}(x+w)) = \frac{1}{2M}\sum_{m=1}^{M}\Big[\mathrm{erf}\big(\tfrac{a_0-\langle x,\tilde{\varphi}_m^*\rangle}{\sqrt{2}\sigma}\big) - \mathrm{erf}\big(\tfrac{-a_0-\langle x,\tilde{\varphi}_m^*\rangle}{\sqrt{2}\sigma}\big)\Big],$$

for any $x \in \mathbb{R}^N$. The set of points satisfying the equation above is shown in Fig. 7.5(a). Note that here, Corollary 7.10 may also be written as

$$\mathcal{E}_p((\tilde{\Phi}^*)^{-1}(\Omega)) \subseteq \{x \in \mathbb{R}^N : \sum_{m=1}^{M}\Big[\mathrm{erf}\big(\tfrac{a_0-\langle x,\tilde{\varphi}_m^*\rangle}{\sqrt{2}\sigma}\big) - \mathrm{erf}\big(\tfrac{-a_0-\langle x,\tilde{\varphi}_m^*\rangle}{\sqrt{2}\sigma}\big)\Big] < 2M-1\}), \tag{7.38}$$

which is illustrated in Fig. 7.5. ∎

7.5 Summary

We investigated a single-class classification problem where the class itself is a compact convex subset of $\mathbb{R}^N$. We introduced a classification scheme based on frames and proved that frame sets can approximate the class in question to within an arbitrary degree of precision. We also introduced a measure-theoretic framework for the study of classification errors in this context and showed that our classification scheme performs well in the presence of radially symmetric noise.

Chapter 8

Lapped Tight Frame Transforms

Contents

Examining the results we obtained with our MR classification system, we found the trends to be similar: MR significantly outperforms no MR and the best results are invariably obtained by frames. Whether it is classifying biomedical or biometric images, frames, and in particular the SWT, always outperformed any other transform. However, the redundancy of frames is only bought at an additional computational cost. Taking this fact into account, it is important to have a system that is efficient in addition to being accurate. The SWT is the most accurate here but also the most redundant. Therefore, to allow for a trade-off between accuracy and cost, we would like to create new frame transforms that are less redundant but still afford very good accuracies when it comes to classification.

A known issue with MR bases is that they are not translation invariant (rather, they are periodically translation invariant). This is due to downsampling being used and can create problems as translated versions of data can lead to different features in MR subspaces. As for the fingerprint data set, we conducted an experiment on the protein subcellular location images to test the sensitivity of our classification system to translations. Our hypothesis is that translations in the testing set produce reduced classification accuracy. We tested this hypothesis by training the system with the original data and tested with images that were translated by some number of pixels. We ran the algorithm with Haralick texture features T_3 alone and with translations of $t = 0, 1, 2, 3$ horizontally and vertically in the testing set (these translations were chosen because we use 2 levels of the MR transform, so it is translation

invariant to translations of 2^2t, but not to translations of $2^2t+1, 2^2t+2, 2^2t+3$). As expected, the classification accuracy dropped by 0.22%. Both experiments for fingerprints and protein subcellular location patterns strongly indicate the use of MR techniques which are translation invariant (or almost translation invariant).

These considerations lead us to conjecture that properties provided by frames, on top of the MR ones, are crucial requirements in some applications. Motivated by the need of having frame families dedicated to a spectrum of applications not considered before, we seek to design new classes of frames.

The question now is: How do we go about constructing new families and what do we look for? Most of the known frame families (though not all) are block based ones (finite-dimensional) leading to blocking effects. We want efficient implementations as well as the flexibility to decide on the requisite amount of redundancy. These requirements reminded us of LOTs: As mentioned in Section 4.1.3, in addition to being computationally efficient, the LOTs have the advantage of processing blocks of overlapping data and hence eliminate blocking artifacts. So the question is: Could we construct a similar transform with frames? Our idea is to seed LOTs to obtain a new class of frames we name *Lapped Tight Frame Transforms (LTFTs)*. That is, we want to find filter-bank frames seeded from the LOTs in the hope they will inherit all the good properties LOTs possess. Obtained by seeding, the LTFTs could thus be seen both as the frame counterpart of LOT bases as well as the infinite-dimensional, filter-bank counterpart of the most famous frame family—Harmonic Tight Frames (HTFs, seeded from the DFT). These relationships are illustrated in the table below.

	finite-dimensional (block transforms)		infinite-dimensional (overlapped transforms)
ONBs	DFT	$\rightarrow$	LOT
	$\downarrow$		$\downarrow$
TFs	HTF	$\rightarrow$	**LTFT**

There has already been some work done in designing what we call LTFTs. In particular, in [46], the authors propose a LTFT derived from the extended lapped complex transform [131]. They use a change of parameters to derive their decomposition vectors from the extended lapped complex transform and ensure that the inverse of such a decomposition, that corresponds to the reconstruction matrix, exists. They also propose a construction of the inverse. These are not obtained by seeding (they start from a frame) and while they are in spirit similar to what we are proposing, they lead to a completely different family. The same authors have also developed a 2D nonseparable LTFT.

In this chapter, we first present general LTFTs, we then look at a specific cases of LTFTs and study four families that we derive from the LOTs presented in Section 4.4.6. We investigate equal-norm and maximal robustness properties for all families and explore window design procedures for the first family. Some results from this chapter appear in [25].

8.1 Lapped Tight Frame Transforms

We previously mentioned that the HTFs are the counterpart of the DFT, that is, they are obtained by seeding the DFT. As we said in Section 4.4.5, the HTFs are finite-dimensional frames and thus equivalent to block transforms. For the same reasons LOTs were introduced, we would like to find filter-bank frames seeded from the LOTs in the hope they will inherit all the good properties LOTs possess. Recall that in filter-bank parlance, seeding is done on the polyphase matrix. Suppose that $\Psi_p(z)$ is the $M \times M$ polyphase matrix associated with the DFT of size M. Then $\Psi_p(z) = \Psi_0$ (see (4.2) and Section 4.4.5) and

$$\Phi_p^*(z) = \Phi_0^* = \Psi_p[\mathbb{J}]$$

is the transpose of the HTF matrix. It turns out that the indices in $\mathbb{J}$ do not have to be contiguous for the following discussion to hold, that is, we can erase any subset of $M - N$ columns from $\Psi_p(z)$ and still get an HTF. However, to simplify the discussion, we take $\mathbb{J} = [0, \ldots, N-1]$. Note that for $M = 3$ and $N = 2$, this procedure leads to the Mercedes-Benz frame [74] (within unitary equivalence).

Now, let us start with $\Psi_p(z)$ being the $M \times M$ polyphase matrix associated with the LOT of size M. Then (4.11) holds and $\Phi_p^*(z) = \Phi_0^* + z^{-1}\Phi_1^* = \Psi_p[0, \ldots, N-1]$. The matrices Φ_r are now rectangular of size $N \times M$. For $r = 0, 1$, we have

$$\Phi_r = \begin{pmatrix} \psi^*_{0,Mr} & \cdots & \psi^*_{0,Mr+M-1} \\ \psi^*_{1,Mr} & \cdots & \psi^*_{1,Mr+M-1} \\ \vdots & \cdots & \vdots \\ \psi^*_{N-1,Mr} & \cdots & \psi^*_{N-1,Mr+M-1} \end{pmatrix}. \tag{8.1}$$

By the Naimark Theorem, we know that this family is a TF, which implies that $\Phi_p(z)\Phi_p^*(z) = cI$ (c is a constant). Note that as opposed to the LOT case, the matrix products no longer commute.

All of the above is general and can be applied to any type of LOT. Let us now go through an example what happens when the obtained LTFT has been seeded by the Princen-Johnson-Bradley (PJB) LOT in (4.14). We then turn our attention to some of the LOT families presented in Sections 4.1.3 and study their frame counterparts. Figure 8.1 shows all four LTFT families produced via a seeding of the LOT families defined in Section 4.1.3.

8.2 The Princen-Johnson-Bradley LTFT

We use the PJB family as our main LTFT example and study here its properties. Namely, we investigate equal-norm and maximal robustness properties. We also explore appropriate solutions for a modulating window analytically and through optimization techniques. The frequency response of the PJB LTFT filters resulting from consecutive seeding of the first $N = 5$ columns with $M = 5$ are depicted in Fig 8.1(a).

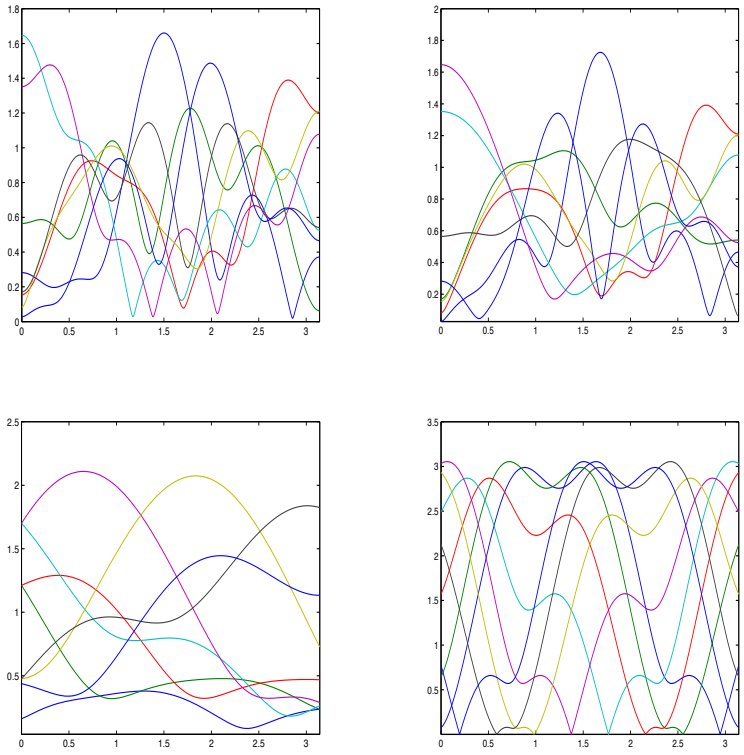


Figure 8.1: LTFT families resulting from consecutive seeding with $M = 8$ and $N = 5$. Namely, in each case there are $M = 8$ filters of length $2N = 10$. (a) Princen-Johnson-Bradley, (b) oddly modulated DCT, (c) Young-Kingsbury, (d) Malvar.

8.2.1 Equal-Norm

To investigate the equal norm property, we need to compute the norm of the frame elements of the PJB-LTFT $\|\varphi_m\|$, for $m = 0, \ldots, M-1$ and prove that they are all equal. In fact, the mth element of the diagonal of the matrix $\Phi_0^*\Phi_0 + \Phi_1^*\Phi_1$ is $\text{diag}(\Phi_0^*\Phi_0 + \Phi_1^*\Phi_1)_m = \|\varphi_m\|^2$. Note that

$$\|\varphi_m\|^2 = \sum_{n=0}^{N-1} \psi^{*2}_{n,m} + \psi^{*2}_{n,m+M}. \tag{8.2}$$

Using this expression, we can prove the following:

PROPOSITION 8.1. If $\{\varphi_m\}_{m=1}^M$ are the frame vectors of the PJB-LTFT, then

$$\|\varphi_m\|^2 = \frac{N}{M}, \qquad m = 0, \cdots, M-1,$$

that is, the PJB-LTFT is an equal norm frame.

We prove this results using two proofs. The first one computes directly $\|\varphi_m\|^2$, whereas the second one relies on expressing the frame coefficients as cosine and sine functions of the same angles, which leads to a straightforward way to prove equal-norm.

Proof.

1. Let us write the $2N \times M$ LTFT matrix Φ explicitly as

$$\Phi = \begin{pmatrix} \psi_{0,0} & \cdots & \cdots & \psi_{0,M-1} \\ \psi_{1,0} & \cdots & \cdots & \psi_{1,M-1} \\ \vdots & \cdots & \cdots & \vdots \\ \psi_{N-1,0} & \cdots & \cdots & \psi_{N-1,M-1} \\ \psi_{0,M} & \cdots & \cdots & \psi_{0,2M-1} \\ \vdots & \cdots & \cdots & \vdots \\ \psi_{N-1,M} & \cdots & \cdots & \psi_{N-1,2M-1} \end{pmatrix},$$

where the mth column of Φ is φ_m for $m = 0, \ldots, M-1$. This allows us to write

$$\begin{aligned} \|\varphi_m\|^2 &= \sum_{n=0}^{2N-1} \varphi_{m,n}^2 \\ &= \sum_{n=0}^{N-1} \psi_{n,m}^2 + \psi_{n,m+M}^2, \end{aligned}$$

where we have from (4.14) that

$$\begin{aligned} \psi_{n,m} &= \tfrac{1}{\sqrt{M}} \cos\left(\tfrac{\pi(2n+1)}{2M}(m - \tfrac{M}{2} + \tfrac{1}{2})\right) \\ &= \tfrac{1}{\sqrt{M}} \cos\left(q_n(m - + \tfrac{M}{2} + \tfrac{1}{2})\right) \text{ and} \\ \psi_{n,m+M} &= \tfrac{1}{\sqrt{M}} \cos\left(q_n(m + + \tfrac{M}{2} + \tfrac{1}{2})\right), \end{aligned}$$

with $q_n = \frac{\pi(2n+1)}{2M}$.

Let $b_m = m - \frac{M}{2} + \frac{1}{2}$ and $c_m = m + \frac{M}{2} + \frac{1}{2}$ and now write

$$\begin{aligned} \psi_{n,m} &= \tfrac{1}{2\sqrt{M}} \left(e^{jq_n b_m} + e^{-jq_n b_m}\right) \\ \psi_{n,m+M} &= \tfrac{1}{2\sqrt{M}} \left(e^{jq_n c_m} + e^{-jq_n c_m}\right). \end{aligned}$$

Then,

$$\begin{aligned}
& M\|\varphi_m\|^2 \\
= \quad & \frac{1}{4}\sum_{n=0}^{N-1}(e^{jq_nb_m}+e^{-jq_nb_m})^2+(e^{jq_nc_m}+e^{-jq_nc_m})^2 \\
= \quad & \frac{1}{4}\sum_{n=0}^{N-1}e^{2jq_nb_m}+e^{2jq_nc_m}+e^{-2jq_nb_m}+e^{-2jq_nc_m}+2+2 \\
= \quad & N+\frac{1}{4}\sum_{n=0}^{N-1}e^{2jq_nb_m}+e^{2jq_nc_m}+e^{-2jq_nb_m}+e^{-2jq_nc_m} \\
\overset{b_m=c_m-M}{=} \quad & N+\frac{1}{4}\sum_{n=0}^{N-1}e^{2jq_nc_m}e^{-2jq_nM}+e^{2jq_nc_m}+e^{-2jq_nc_m}e^{2jq_nM}+e^{-2jq_nc_m} \\
= \quad & N+\frac{1}{4}\sum_{n=0}^{N-1}e^{2jc_m}[\cos(\pi(2n+1))+1]+e^{-2jc_m}[\cos(\pi(2n+1))+1] \\
= \quad & N,
\end{aligned}$$

where the last equality results from the fact that $q_nM=\frac{\pi}{2}(2n+1)$ and $e^{-2jq_nM}=\cos(\pi(2n+1))=e^{2jq_nM}=-1$, for all $n=0,\ldots,N-1$. □

2. For this second proof, we use the polyphase version of the PJB-LOT filters writing the elements of $\Psi_p(z)$ as

$$\begin{aligned}
\psi_{k,m}(z) &= \psi_{k,m}+z^{-1}\psi_{k,m+M} \qquad \text{for } k,m=0,\ldots,M-1 \\
&= \tfrac{1}{\sqrt{M}}\cos\tfrac{(2k+1)(2m+1-M)}{4M}\pi \\
&\quad +\tfrac{1}{\sqrt{M}}z^{-1}\cos\left(\tfrac{(2k+1)(2(m+M)+1-M)}{4M}\pi\right) \\
&= \tfrac{1}{\sqrt{M}}\cos\tfrac{(2k+1)(2m+1-M)}{4M}\pi \\
&\quad +\tfrac{1}{\sqrt{M}}z^{-1}\cos\left(\tfrac{(2k+1)(2(m+M)+1-M)}{4M}\pi+\tfrac{(2k+1)\pi}{2}\right).
\end{aligned}$$

Therefore,

$$\psi_{k,m}(z) \;=\; \tfrac{1}{\sqrt{M}}\left(\cos\tfrac{(2k+1)(2m+1-M)}{4M}\pi+(-1)^{k+1}z^{-1}\sin\tfrac{(2k+1)(2m+1-M)}{4M}\pi\right), \tag{8.3}$$

for $k,m=0,\ldots,M-1$. Then, by using (8.2), it is easy to see that

$$M\|\varphi_m\|^2=N,$$

for all $m=0,\ldots,M-1$.

□

8.2.2 Maximal Robustness

Since LTFTs are the counterparts of HTFs, we expect them to be maximally robust as well. This requirement arose in using frames for robust transmission where the loss of up to $M-N$ transform coefficient over the transmission channel would not be fatal. The loss of coefficients translates into removal of the corresponding set of $M-N$ columns in $\Phi_p(z)$ and the ability to reconstruct despite the loss translates into the remaining matrix being invertible. Hence, a LTFT is maximally robust if and only if any $N \times N$ submatrix of $\Phi_p(z)$ is of full rank on the unit circle.

To study necessary and sufficient conditions for a LTFT to be maximally robust, we need to go back to the LOT family from which it originated and consider the type of seeding that would lead us to maximal robustness.

For $z=-j$, $\psi_{k,m}(z)$ in (8.3) can be expressed in terms of the roots of unity $W_M = e^{-j\frac{2\pi}{M}}$ as

$$\psi_{k,m}(-j) = \tfrac{1}{\sqrt{M}} W_{8M}^{(-1)^k(2k+1)(1-M)} W_{4M}^{(2k+1)m}.$$

The row-scaling factors $W_{8M}^{(-1)^k(2k+1)(1-M)}$, $k=0,\ldots,M-1$, can be collected into an invertible diagonal matrix, so that

$$\Psi_p(-j) \;=\; U \cdot \hat{\Psi}_p(-j),$$

$$U = \tfrac{1}{\sqrt{M}} \mathrm{diag}\left(W_{8M}^{(-1)^k(2k+1)(1-M)} \right)_{0\le k\le M-1}, \quad \hat{\Psi}_p(-j) = \left[W_{4M}^{(2k+1)m} \right]_{0\le k,m\le M-1}.$$

Note that $\hat{\Psi}_p(-j)$ is a submatrix of the DFT matrix DFT_{4M} of size $4M$ and is constructed from the first M *consecutive* columns and the first M odd rows of DFT_{4M}. Therefore, any seeding of $\hat{\Psi}_p(-j)$ is a seeding of a submatrix of a DFT. That is, a choice of N consecutive columns in $\Psi_p(-j)$ to build the $M \times N$ LTFT matrix $\hat{\Phi}_p(-j)$ is the same choice of N consecutive columns in a submatrix of the DFT_{4M}.

Using the results in [98], we know that any choice of *consecutive* seeding of a DFT matrix produces a maximally robust frame. Hence, any *consecutive* seeding of a submatrix (here, $\hat{\Psi}_p(-j)$) built using *consecutive* columns of a DFT, produces a maximally robust frame. In particular, let us assume that $\hat{\Phi}_p(-j)$ results from consecutive seeding of $\hat{\Psi}_p(-j)$. Then, any $N \times N$ submatrix of $\hat{\Phi}_p(-j)$ is also a submatrix of a frame consecutively-seeded from DFT_{4M}, which is invertible. That is $\hat{\Phi}_p(-j)$ is a maximally robust frame. Now, since U is a diagonal invertible matrix, we can use the invariance properties of frames described in Section 4.4.4 to conclude that the PJB LTFT $\Phi_p(-j) \;=\; U\hat{\Phi}_p(-j)$ is a maximally robust frame. Note that proving this result for $z=-j$ is sufficient to deduce that $\Phi_p(z)$ is also maximally robust. We summarize our result as follows:

PROPOSITION 8.2. The PJB LTFT resulting from consecutive seeding of the PJB LOT is a maximally robust frame.

Note that consecutive seeding is a sufficient condition only and unlike the results in [98], the seeding cannot be cyclically contiguous (only contiguous).

8.2.3 Window Design

Since LTFTs are determined by the LOT basis functions, we have little design freedom in the frames construction. To circumvent this, we can design a modulating window that provides additional degrees of freedom in the design as well as improves the frequency response of the filters. Ideally, a single window would modulate all filters at once. Namely, a diagonal window matrix would multiply $\Phi_p(z)$ to produce a modulated LTFT. We tackled the design of this window in two different ways. The first, analytical using perfect reconstruction conditions for the filter bank. The second uses optimization techniques where the goal is to approximate the frequency behavior of harmonic tight frames, as these are narrow band-pass filters evenly spread across the frequency domain.

Analytical Design If we start with the Princen-Johnson-Bradley LOT with a window Δ, and seed $\Delta\Psi$, the tight frame obtained would loose its equal-norm property since $\|\varphi_m\|^2 = (N/M)\delta_m^2$. To preserve equal norm, we have to modulate directly the LTFT after seeding the LOT. In the Princen-Johnson-Bradley LOTs, the window chosen was symmetric, that is, $\delta_m = \delta_{2M-1-m}$. We lift this restriction initially and assume a general window represented by a matrix Δ, a $2N \times 2N$ diagonal matrix. We can write $\Delta = [\Delta_0 \quad \Delta_1]$ and Δ_r is a $N \times N$ diagonal matrix. Unlike for the LOTs, the matrix product $\Phi_0\Phi_0^*$ has no particular structure, in fact,

$$(\Phi_0\Phi_0^*)_{i,n} \;=\; a_{i,n} \;=\; \tfrac{1}{2M}\frac{\sin(\frac{\pi(i+n+1)}{2})}{\sin(\frac{\pi(i+n+1)}{2M})} + \frac{1}{2M}\frac{\sin(\frac{\pi(i-n)}{2})}{\sin(\frac{\pi(i-n)}{2M})},$$

for $i, n, = 0, \ldots, N-1$.

Proof. We have that $(\Phi_0\Phi_0^*)_{i,n} = a_{i,n} = \sum_{k=0}^{M-1} \psi_{i,k}\psi_{n,k}$.

$$\begin{aligned} Ma_{i,n} &= \sum_{k=0}^{M-1} \cos\left(\frac{2i+1}{4M}(2k-M+1)\pi\right)\cos\left(\frac{2n+1}{4M}(2k-M+1)\pi\right) \\ &= \frac{1}{2}\sum_{k=0}^{M-1} \cos\left(\frac{2i+2n+2}{4M}(2k-M+1)\pi\right) + \cos\left(\frac{2i-2n}{4M}(2k-M+1)\pi\right) \\ &= \underbrace{\frac{1}{4}\sum_{k=0}^{M-1} e^{jlv_k\pi} + e^{-jlv_k\pi}}_{\alpha} + \underbrace{\frac{1}{4}\sum_{k=0}^{M-1} e^{jov_k\pi} + e^{-jov_k\pi}}_{\beta}, \end{aligned}$$

where $l = 2i+2n+2, o = 2i-2n$, and $v_k = 2k-M+1$. We now compute α and

obtain:

$$\begin{aligned}\alpha &= \frac{1}{4}\sum_{k=0}^{M-1} e^{2jlk\pi}e^{j(-M+1)l\pi} + e^{-2jl\pi}e^{-j(-M+1)l\pi} \\ &= \frac{1}{4}(e^{-jlM\pi}e^{jl\pi}\frac{1-e^{2jlM\pi}}{1-e^{2jl\pi}} + e^{jlM\pi}e^{-jl\pi}\frac{1-e^{-2jlM\pi}}{1-e^{-2jl\pi}}) \\ &= \frac{1}{2}\frac{\sin(lM\pi)}{\sin(l\pi)}.\end{aligned}$$

Similarly, $\beta = \frac{1}{2}\frac{\sin oM\pi}{\sin o\pi}$ and we finally obtain the desired result:

$$Ma_{i,n} = \frac{1}{2}\left(\frac{\sin\frac{(i+n+1)\pi}{2}}{\sin\frac{(i+n+1)\pi}{2M}} + \frac{\sin\frac{(i-n)\pi}{2}}{\sin\frac{(i-n)\pi}{2M}}\right). \tag{8.4}$$

□

Substituting this into (4.12), we obtain the following:

$$a_{n,n}\delta_n^2 + (1-a_{n,n})\delta_{N+n}^2 = 1, \tag{8.5}$$

$$\delta_n\delta_i = \delta_{N+n}\delta_{N+i}, \quad i = 0,\cdots,N-1, i \neq n. \tag{8.6}$$

The set of solutions to (8.5)-(8.6) is infinite. Of course, the constant window with $\delta_n = 1$, for $n = 0,\ldots,2N-1$ is also a solution to the above. Finding the best window amongst all the possible solutions is part of our future work.

If the window is symmetric, then (4.12) becomes:

$$\Delta_0\Phi_0\Phi_0^*\Delta_0 + J\Delta_0 J\Phi_1\Phi_1^* J\Delta_0 J = I \tag{8.7}$$

$$\text{with} \quad \Phi_0\Phi_0^* + \Phi_1\Phi_1^* = I. \tag{8.8}$$

Using (8.7), we derive the following conditions on Δ:

$$a_{n,n}\delta_n^2 + (1-a_{n,n})\delta_{N-n-1}^2 = 1, \tag{8.9}$$

$$\delta_n\delta_i = \delta_{N-n-1}\delta_{N-i-1}, \quad i = 0,\cdots,N-1, i \neq n. \tag{8.10}$$

Proof. Using $\Phi_0\Phi_0^* = I - \Phi_1\Phi_1^*$, we can rewrite (8.7) as

$$\Delta_0\Phi_0\Phi_0^*\Delta_0 + J\Delta_0^2 J - J\Delta_0 J\Phi_0\Phi_0^* J\Delta_0 J = I, \tag{8.11}$$

where $\Delta_0 = \text{diag}\{\delta_n\}_{n=0}^{N-1}$. Note that $J\Delta_0 J = \text{diag}\{\delta_{\text{N-1-n}}\}_{\text{n=0}}^{\text{N-1}}$ and similarly for $J\Delta_0^2 J$. Also, $\Phi_0\Phi_0^*$ is a symmetric matrix. Assuming as previously that $(\Phi_0\Phi_0^*)_{i,n} = a_{i,n}$, we have

$$\Delta_0\Phi_0\Phi_0^*\Delta_0 = \begin{pmatrix} \delta_0^2 a_{0,0} & \cdots & \delta_j\delta_0 a_{0,n} & \cdots \\ \delta_0\delta_1 a_{1,0} & \cdots & \delta_j\delta_1 a_{1,n} & \cdots \\ \delta_0\delta_2 a_{2,0} & \cdots & \delta_j\delta_2 a_{2,n} & \cdots \\ \vdots & \cdots & \vdots & \cdots \\ \delta_0\delta_{N-1}a_{N-1,0} & \cdots & \delta_n\delta_{N-1}a_{N-1,n} & \cdots \end{pmatrix},$$

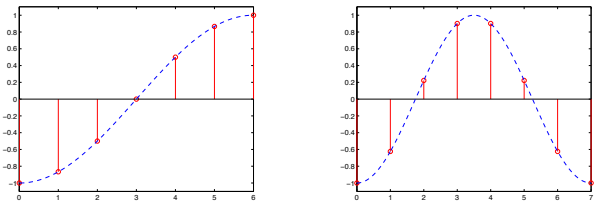


Figure 8.2: Window solution to (8.9)-(8.10) for $N = 7, 8$ (left to right).

and

$$J\Delta_0 J\Phi_0\Phi_0^* J\Delta_0 J = \begin{pmatrix} \delta_{N-1}^2 a_{0,0} & \cdots & \delta_{N-n-1}\delta_{N-1}a_{0,n} & \cdots \\ \delta_{N-1}\delta_{N-2}a_{1,0} & \cdots & \delta_{N-n-1}\delta_{N-2}a_{1,n} & \cdots \\ \delta_{N-1}\delta_{N-3}a_{2,0} & \cdots & \delta_{N-n-1}\delta_{N-3}a_{2,n} & \cdots \\ \vdots & \cdots & \vdots & \cdots \\ \delta_{N-1}\delta_0 a_{N-1,0} & \cdots & \delta_{N-n-1}\delta_0 a_{N-1,n} & \cdots \end{pmatrix}.$$

Hence, using (8.11), the conditions that Δ_0 have to satisfy are

$$\begin{cases} (\delta_n^2 - \delta_{N-n-1}^2)a_{n,n} + \delta_{N-n-1}^2 & = 1 \\ \delta_n\delta_i - \delta_{N-n-1}\delta_{N-i-1} & = 0 \text{ for } i \neq n, i = 0, \ldots, N-1. \end{cases},$$

which are the conditions written as (8.9) and (8.10). □

Fixing $\delta_0 = -1$, we have $\delta_{N-1} = \pm 1$ and $\delta_i = -\delta_{N-1}\delta_{N-i-1}$ for $i = 1, \ldots, N-2$. Note that the same conditions hold for an anti-symmetric window, that is, the half-windows can only be symmetric or antisymmetric. For a symmetric window, a possible solution, depicted in Fig. 8.2, is given by

$$\delta_n = \begin{cases} \cos(\frac{n\pi}{N-1} + \pi) & \text{if } N \text{ is even,} \\ \cos(\frac{2n\pi}{N-1} + \pi) & \text{if } N \text{ is odd,} \end{cases} \qquad n = 0, \ldots, N-1.$$

Optimization Techniques The first window design procedure using optimization tools we explore is through error minimization algorithms. This procedure finds the optimal window $\hat{\delta}$ that minimizes the weighted error between HTF and LTFT filters in the frequency domain as follows

$$\hat{\delta} = \underset{\delta}{\operatorname{argmin}}\ \kappa \odot (\Psi^{(HTF)} - \delta \circledast \Psi^{(PJB)})$$

where κ is a weight vector, $\odot$ denotes point-wise multiplication and $\circledast$ denotes column-wise convolution. Note that to make sizes compatible, we need to use a stacked version of the HTF. Namely, we build $\Psi^{(HTF)}$ by stacking two $N \times M$ HTF matrices on top of each other. Algorithms used to implement this procedure

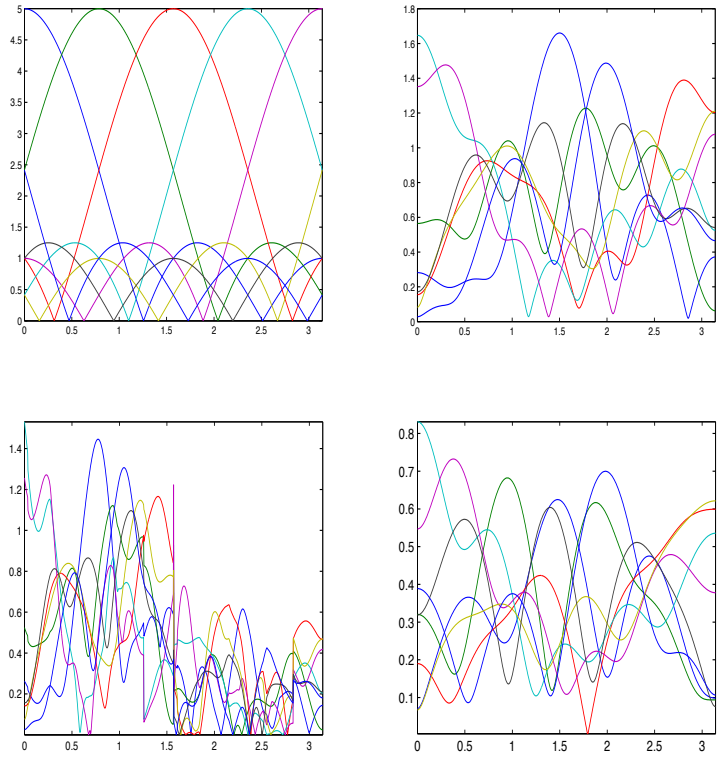


Figure 8.3: Window design for the PJB LTFT with $M = 8$ and $N = 5$. (a) HTF filters, (b) PJB LTFT filters, (c) Modulated PJB LTFT filters with $\hat{\delta}$, generated through error minimization techniques, (d) Modulated PJB LTFT filters with a randomly generated window.

include the trust region method [30] and sequential quadratic programming methods [17], which are gradient descent-based methods. The results we obtain with this procedure are not as promising as hoped. Indeed, when we randomly generate the modulating window, the modulated LTFT filters look better than the ones modulated with the optimized window $\hat{\delta}$. Fig 8.3 shows the results when using the optimized window (Fig 8.3 (c)) and the random window (Fig 8.3 (d)). The frequency response of the HTF filters is depicted for reference on Fig 8.3 (a) and the frequency response of the PJB LTFT filters is shown on Fig 8.3 (b).

The second type of design procedure we investigate relies upon the polar decomposition of matrices and the Fan and Hoffman theorem [41]. The polar decomposition of a matrix Ψ is defined as follows [39]: Given a $2N \times 2N$ matrix Ψ, its polar decomposition is

$$\Psi = \Delta\Sigma$$

where Σ is a $2N \times 2N$ Hermitian positive semi-definite matrix defined as $\Sigma = (\Psi\Psi)^{\frac{1}{2}}$, and Δ is a $2N \times 2N$ unitary matrix with singular value decomposition written as

$$\Delta = P \begin{pmatrix} I_r & 0 \\ 0 & \Theta \end{pmatrix} Q^*,$$

with r is the rank of Ψ, and the matrices P and Q originate from the singular value decomposition of Ψ itself. That is, the singular value decomposition of Ψ is $\Psi = P\Lambda Q^*$. Note that Δ is unique if Ψ has full rank. Then the best approximation theorem by Fan and Hoffman [41] states that

$$\|\Psi - \Delta\| = \min \{\|\Delta - Q\| : Q^*Q = I_{2N}\} \tag{8.12}$$

for any unitarily invariant norm.

By taking $\Psi = \Psi^{(HTF)} \times \Psi^{(PJB)^{-1}}$ and Δ the modulating window *matrix*, the problem becomes:

$$\hat{\Delta} = \|\Psi - \Delta\| = \underset{\Delta}{\operatorname{argmin}} \|\Delta - \Psi^{(HTF)} \times \Psi^{(PJB)^{-1}}\|. \tag{8.13}$$

Therefore, we obtain a $2N \times 2N$ modulating window matrix. As ideally, we would like to have one window vector for the entire set of LTFT filters, we can use $\hat{\Delta}$ in three different ways:

1. $\Delta_1 = \hat{\Delta}$. Each column of Δ_1 modulates one LTFT filter.
2. $\delta_{\mathbf{2}}$ is the vector of eigenvalues of $\hat{\Delta}$. This window modulates all LTFT filters.
3. $\delta_3 = \lambda_1 \times e_1$, where λ_1 is largest eigenvalue of $\hat{\Delta}$ and e_1 its corresponding eigenvector.

Fig 8.4 shows the frequency response of the PJB LTFT filters when modulated by Δ_1 (Fig 8.4(a)), $\delta_{\mathbf{2}}$ (Fig 8.4(b)), and $\delta_{\mathbf{3}}$ (Fig 8.4(c)). All results show a small improvement over the original PJB in that they have a somewhat better localization in the frequency band of the PJB frame vectors, with $\delta_{\mathbf{2}}$ being the best one.

8.3 The Oddly Modulated DCT LTFT

Figure 8.1(b) depicts the frequency response of the eight oddly modulated DCT LTFT filters obtained from contiguous seeding (choice of the first N columns) of the oddly modulated DCT LOT with $M = 8$ and $N = 5$.

8.3.1 Equal-Norm

Similarly to the PJB LTFT, to prove the equal norm property for the oddly modulated DCT LTFT, we rewrite (4.18) in its polyphase version as

$$\begin{aligned} \psi_{k,m}(z) &= \psi_{k,m} + z^{-1}\psi_{k,m+M} \\ &= \tfrac{1}{\sqrt{M}} \cos \tfrac{(2k+1)(2m+1+M)}{4M}\pi + \tfrac{1}{\sqrt{M}} z^{-1} \cos \tfrac{((2k+1)(2(m+M)+1+M)}{4M}\pi) \\ &= \tfrac{1}{\sqrt{M}} \cos (\tfrac{(2k+1)(2m+1+M)}{4M}\pi) + \tfrac{1}{\sqrt{M}}(-1)^{k+1} z^{-1} \sin (\tfrac{(2k+1)(2m+1+M)}{4M}\pi) \end{aligned} \tag{8.14}$$

Using (8.2), we deduce the following:

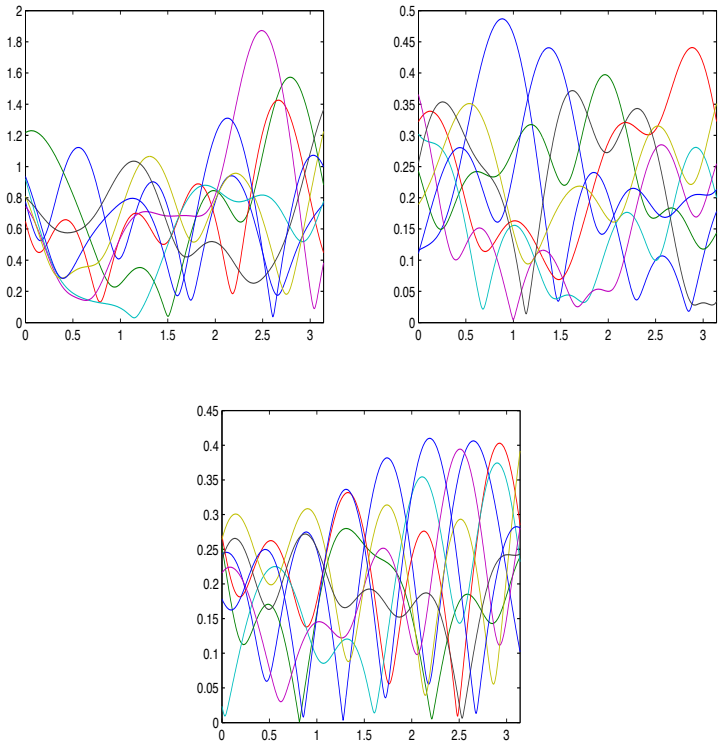


Figure 8.4: Modulated PJB LTFT filters using the polar decomposition method ($M = 8, N = 5$). (a) PJB LTFT filters modulated with Δ_1, (b) LTFT PJB filters modulated with $\delta_{\mathbf{2}}$, (c) PJB LTFT filters windowed with $\delta_{\mathbf{3}}$.

PROPOSITION 8.3. If $\{\varphi_m\}_{m=1}^{M}$ are the frame vectors of the oddly modulated DCT LTFT, then

$$\|\varphi_m\|^2 = \frac{N}{M}, \qquad i = m, \ldots, M-1,$$

that is, the oddly modulated DCT LTFT is an equal norm frame.

8.3.2 Maximal Robustness

For $z = -j$, $\psi_{k,m}(-j)$ in (8.14) can be expressed in terms of the roots of unity W_M as

$$\psi_{k,m}(-j) = \tfrac{1}{\sqrt{M}} W_{8M}^{(-1)^k(2k+1)(1+M)} W_{4M}^{(2k+1)m}.$$

The row-scaling factors $\frac{1}{\sqrt{M}} W_{8M}^{(-1)^k(2k+1)(1+M)}$, $k = 0, \ldots, M-1$, can be collected into an invertible diagonal matrix, so that

$$\Psi(-j) = U \cdot \hat{\Psi}_p(-j),$$

where $U = \frac{1}{\sqrt{M}}\text{diag}\left(W_{8M}^{(-1)^k(2k+1)(1+M)}\right)_{0\le k\le M-1}$ and $\hat{\Psi}_p(j) = \left[W_{4M}^{(2k+1)m}\right]_{0\le k,m\le M-1}$.

Similarly to the PJB case, $\hat{\Psi}_p(-j)$ is a submatrix of DFT_{4M}, constructed from the first M *consecutive* columns and first M odd rows. Therefore, selecting any number N of *consecutive* columns of $\hat{\Psi}_p(-j)$ and constructing an $M\times N$ matrix $\hat{\Phi}_p(-j)$ from them corresponds to selecting N *consecutive* columns from DFT_{4M}. We then conclude that $\hat{\Phi}_p(-j)$ is a maximally robust frame and so is $\Phi_p(z)$ using the invariance of frame properties. We summarize this result as follows:

PROPOSITION 8.4. The oddly modulated DCT LTFT resulting from consecutive seeding of the oddly modulated DCT LOT is a maximally robust frame.

8.4 The Young-Kingsbury LTFT

The frequency responses of the Young-Kingsbury LTFT filters are shown in Fig. 8.1(c) for $M = 8$ and $N = 5$. As in the example for the previous LTFT families, these result from seeding contiguously the first N columns.

8.4.1 Equal-Norm

Unlike the previous two families, the Young-Kingsbury LOT has complex basis vectors (4.19). However, we can still use the same argument as for the other families to prove the equal norm property. Indeed, we can write

$$\psi_{k,m} \;=\; \tfrac{1}{\sqrt{M}}\cos\tfrac{\pi m}{2M}e^{-j\frac{(2k+1)m\pi}{2M}} \;=\; \tfrac{1}{\sqrt{M}}\cos\left(\tfrac{\pi l}{2M}\right)W_{4M}^{(2k+1)m},$$

and the polyphase elements as:

$$\begin{aligned}\psi_{k,m}(z) &= \tfrac{1}{\sqrt{M}}\cos\left(\tfrac{\pi m}{2M}\right)W_{4M}^{(2k+1)m} + \tfrac{2}{\sqrt{M}}z^{-1}\cos\left(\tfrac{\pi(m+M)}{2M}W_{4M}^{(2k+1)(m+M)}\right)\\ &= \tfrac{1}{\sqrt{M}}\cos\tfrac{\pi m}{2M}W_{4M}^{(2k+1)m} + \tfrac{2}{\sqrt{M}}z^{-1}\sin\left(\tfrac{\pi m}{2M}\right)W_{4M}^{(2k+1)M}W_{4M}^{(2k+1)m}\\ &= \tfrac{1}{\sqrt{M}}W_{4M}^{(2k+1)m}\left(\cos\left(\tfrac{\pi m}{2M}\right) + z^{-1}(-1)^{k+1}j\sin\left(\tfrac{\pi m}{2M}\right)\right). \qquad (8.15)\end{aligned}$$

Since $|W_{4M}^{(2k+1)m}| = 1$, we can conclude the following:

PROPOSITION 8.5. If $\{\varphi_m\}_{m=1}^{M}$ are the frame vectors of the Young-Kingsbury-LTFT, then

$$\|\varphi_m\|^2 = \frac{N}{M}, \qquad m = 0,\ldots,M-1,$$

that is, the Young-Kingsbury-LTFT is an equal norm frame.

8.4.2 Maximal Robustness

By taking $z = 1$ and using (8.15), we can write $\psi_{k,m}(z)$ as

$$\psi_{k,m}(1) = \tfrac{1}{\sqrt{M}}W_{4M}^{(2k+1+(-1)^{k+1})m}.$$

By the same reasoning we used for the previous families, we again observe that the Young-Kingsbury polyphase matrix taken at $z = 1$, $\Psi_p(1)$ is a submatrix of DFT_{4M}, constructed from the first M *consecutive* columns . By applying the same argument as for the PJB LOT, we can conclude the following:

PROPOSITION 8.6. The Young-Kingsbury LTFT resulting from consecutive seeding of the Young-Kingsbury LOT is a maximally robust frame.

8.5 The Malvar LTFT

The frequency response of the Malvar LTFT filters for $M = 8$ and $N = 5$ are shown in Fig. 8.1(d).

8.5.1 Equal-Norm

Similarly to the previous family, we write the polyphase elements of the Malvar LOT defined in (4.20) as:

$$\begin{aligned}\psi_{k,m}(z) &= \tfrac{2}{\sqrt{M}} W_{8M}^{-(2k+1)(2m+M+1)} + \tfrac{2}{\sqrt{M}} z^{-1} W_{8M}^{-(2k+1)(2(m+M)+M+1)} \\ &= \tfrac{2}{\sqrt{M}} \left(1 + (-1)^k j z^{-1}\right) W_{8M}^{-(2k+1)(M+1)} W_{4M}^{(2k+1)m}.\end{aligned}$$

PROPOSITION 8.7. If $\{\varphi_m\}_{m=1}^{M}$ are the frame vectors of the Malvar LTFT, then

$$\|\varphi_m\|^2 = \frac{2N}{M}, \qquad m = 0, \ldots, M-1,$$

that is, the Malvar LTFT is an equal norm frame.

8.5.2 Maximal Robustness

For the Malvar LTFT, the row-scaling factors $\frac{2}{\sqrt{M}} \left(1 + (-1)^k j z^{-1}\right) W_{8M}^{-(2k+1)(M+1)}$, $k = 0, \ldots, M-1$, can be collected into a diagonal matrix, so that

$$\Psi(z) \;=\; U(z)\hat{\Psi},$$

where $U(z) = \frac{2}{\sqrt{M}} \mathrm{diag}\left((1 + (-1)^k j z^{-1}) W_{8M}^{-(2k+1)(M+1)}\right)_{0 \le k \le M-1}$ and $\hat{\Psi} = \left[W_{4M}^{(2k+1)m}\right]_{0 \le k,m \le M-1}$.

Here again, $\hat{\Psi}$ is a submatrix of DFT_{4M}, constructed from the first M *consecutive* columns and first M odd rows. Hence, by using the same arguments as for the other LTFT families, we draw the following conclusion

PROPOSITION 8.8. The Malvar LTFT resulting from consecutive seeding of the Malvar LOT is a maximally robust frame.

8.6 Summary

Using a simple design procedure, we developed new frame families we termed lapped tight frame transforms. These can be viewed as the frame counterpart of lapped orthogonal transforms. Similarly to harmonic tight frames, LTFTs are tight, and we proved that they are equal-norm and maximally robust as well. Finally, LTFTs are efficient to implement and their construction provide flexibility and control over the desired amount of redundancy. In an MR classification setting this is important as it would allow for example to adjust the redundancy depending on the biomedical application at hand and reach a desired compromise between redundancy and cost.

Part V

Conclusions

Conclusions

We have provided a mathematical framework for redundant
= multiresolution classification and have designed an accurate and adaptive multiresolution classification algorithm for the classification of biomedical images. We divided our work into two themes: The first was the design of a classification algorithm for biomedical applications based on multiresolution techniques. The second was the development of a theory of frame multiresolution classification along with the design of new frame families tailored for biomedical applications.

Multiresolution Classification Algorithm We have developed an accurate, efficient and adaptive supervised classification algorithm based on multiresolution (MR) techniques, which aims to extract discriminative features within space-frequency localized MR subspaces. These features are obtained by MR decomposition; that is, rather than add MR features to existing features, we instead chose to compute these features in the MR-decomposed subspaces themselves. Thus, our system has an upfront MR decomposition block which is followed by feature computation and classification in each of the MR subspaces, which, in turn, are then combined through an adaptive weighting process. For the MR decomposition step, we used both MR bases and MR frames. The main features used were Haralick-based texture features as these seem to well characterize the biomedical data sets under consideration in this work. We tested our system on five applications obtaining excellent results in four of the five, as well as promising initial results in the remaining case. As proven by the high accuracies obtained on the fingerprint recognition problem, our MR classification algorithm is flexible and can be used for data sets other than the ones we considered during this thesis.

Theory of Frame Multiresolution Classification As the use of redundant MR transformations in our classification algorithm outperformed nonredundant ones, we explored deeper the frame classification question and provided a framework for the development of a rigorous understanding of why frames perform better than bases when it comes to the classification of certain classes of signals. We simplified the question and investigated a single-class classification problem where the class itself is a compact convex subset of $\mathbb{R}^N$. Convex sets may be approximated by convex polytopes, and may thus be regarded as the preimages of hyperrectangles under frame analysis operators, which we termed *frame sets*. We proposed a classification

scheme based on frames and akin to a majority voting. We proved that frame sets can approximate any compact convex class to within an arbitrary degree of precision. We introduced a measure-theoretic framework for the analysis of classification errors, and applied it to the study of our proposed classification scheme. We showed this scheme performs well in the presence of radially symmetric noise, and provided upper bounds on the measure of the set of points at which misclassification will frequently occur.

We also developed new frame families we termed lapped tight frame transforms. These can be viewed as the frame counterpart of lapped orthogonal transforms. We showed in four specific cases that in addition to being tight, lapped tight frame transforms possess many desirable properties, such as equal norm, maximal robustness and efficient implementation. In the MR classification algorithm, the frame representation that is the most accurate is also the most expensive in terms of computational cost. This new family of frames is simple to design and its construction provide flexibility and control over the desired amount of redundancy. This allows the user to customize the trade-off between efficiency and accuracy. In addition to providing custom-tailored frame transforms, the design of this new family enriches the frame toolbox and offers a larger choice on the menu of redundant MR representations.

Reproducible Research In the past few year there have been many efforts in the signal processing community to adopt the ideas of reproducible research (see for example [18, 7]). The goal here is to make freely available all the necessary tools and materials that led to a publication. This allows signal processing algorithms to be widely and freely accessible to the scientific community and it also permits easy integration of results into more larger projects as well as facilitates exchanges and collaborations within the community. In our work, we followed the reproducible research paradigm. Namely, we distributed via the web all the material used to produce journal papers issued from our work. We made our classification algorithm as well as the frame toolbox available freely. The code is be accompanied by a compendium containing all the necessary data to reproduce any of the results in our published papers, as well as additional material such as proofs and pseudo-code (see, for example, the compendium for [23]).

Future Research

Multiresolution Classification Algorithm

To improve the performance of our MR classifier, a few venues are possible. By examining each of the blocks of the system, we discuss several potential avenues for enriching our classification toolbox as well as ways to enhance its last three blocks: feature extraction, classifier and weighting procedure.

Feature Extraction Block The aim of this block is to extract numerical features that will best characterize the data at hand. In building the MR classification system, we not only intended it to be accurate and adaptive to the specific data

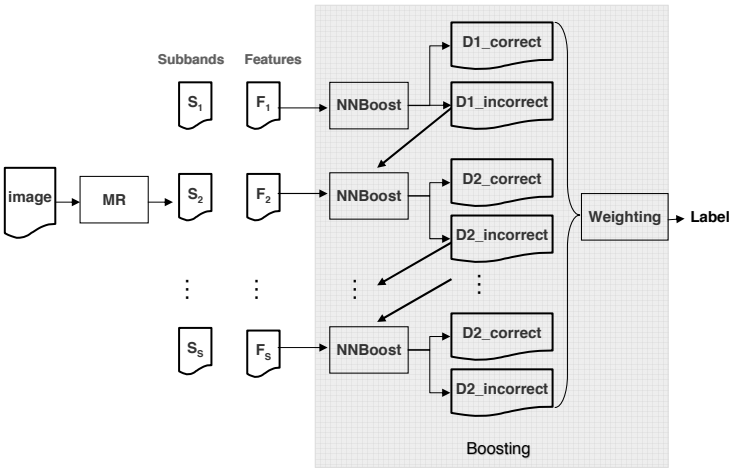


Figure 8.5: MR boosting classification system.

sets available to us, but to be versatile and applicable to other data sets as well. Having that in mind, we would like to have a pool of feature sets from which to choose depending on the input data.

It would also be important to examine how to make this block adaptive to the input signal. This is possible in two different ways: a "coarse" adaptivity per feature set and a "finer" adaptivity per subband. For the former, we would select the feature set that best suits each subband. Since each subband expresses different space-frequency content which lies within a signal, it is natural to think that some feature sets are more suitable for some subbands than for others. For example, it seems natural to have some morphological features (such as the number of objects) used in the coarse subbands, but in the detailed versions, one would rely on other types of features. As for the latter, we can describe a subband even better if within each feature set chosen for that MR subspace, we target only relevant features. This way, we avoid computing features that are not useful for classifying a data set. Thus reducing the size of the feature vector characterizing a subband at a given position in the MR tree would not only allow us to better describe the data but would improve the efficiency of the classification system as well. In a similar fashion to what we did with the weighting procedure, another variation on this theme would be to have the feature selection process per class and per data set.

Classifier and Weighting Blocks Currently, the classifier we use is composed of NNs that act independently on each subband to produce local decisions. The subsequent weighting procedure combines many local decisions into a single global one to finally assign a label to an image. We propose as a future venue to allow the subbands to work together to come up with a final decision by using a modified version of a boosting algorithm.

Boosting is a powerful technique for combining multiple base classifiers (here NNs) to form a committee whose performance can be significantly better than that of any of the base classifiers (called *weak learners*), as long as these achieve accuracies slightly higher than random. The most widely used form of boosting algorithm is called *AdaBoost* (for adaptive boosting) [44]. With boosting, the base classifiers are trained in sequence, and the boosting effect comes from the fact that data points that have been misclassified by one of the previous base classifiers are given greater weight when used to train the next classifier in the sequence. Once the classifiers are trained, their predictions are combined through a weighted majority voting process.

We propose an MR boosting algorithm where each base classifier (NN) represents a subband. The subbands would now work in sequence as opposed to the current system where they work in parallel. We train the first NN classifier (corresponding to the first subband) using weighting coefficients that are all equal. That means that we give each data point equal importance in reaching a correct classification decision. In the subsequent iterations, the weighting coefficients are increased for data points that were misclassified and decreased for data point that are correctly classified. Successive classifiers (subbands) are therefore forced to put greater emphasis on those points that have been misclassified by previous subbands, and data points that continue to be misclassified by successive classifiers receive ever greater weight. Once all subbands have been trained, their decisions are assembled by a weighted sum (similar to our current weighting procedure) where a greater weight will be given to the more accurate classifier.

Note that, unlike the usual boosting algorithm where the subsequent classifiers have as their input the set of misclassified data points, here (see Fig. 8.5) we propose to boost the performance of each base classifier based on the input from "its subband" as well as from the previous classifier (the misclassified points).

Theory of Frame Multiresolution Classification

Our work on frame classification establishes the foundations for a theory that allows to answer fundamental questions. We proved that, in a particular setting, frames outperform bases in regards to classification. In the future, this work can be generalized to more complicated classification problems. A top priority would be to extend our theory to multiple classes as well as classes with multiple clusters. It is also important to consider making the decision function a nonlinear function of the transform coefficients, as is usually the case in any real-world implementation of these ideas. By doing so, one would then be able to fit the present MR classification work in this model (multiple classes, nonlinear features) and truly establish theoretical results for the MR classification of biomedical images.

Another general issue, similar to the one we tackled in this work, is that of why is it always better to use multiresolution representations for classification as opposed to using only the original data. We believe that first gathering a complete understanding and answer to the first question "why are frames better than bases" will help better tackle this question. It is unclear how one would easily answer "why using MR is better than not using it?". We anticipate that here the modeling

of the signals at hand would play an important role in understanding the role of multiresolution representations in classification.

As for the new frame family we designed—lapped tight frame transforms, the immediate future step would be to integrate these into our adaptive multiresolution classification algorithm and study their performance for different biomedical data sets. Another possible venue is to find necessary conditions that would ensure the maximal robustness property of these frames.

Bibliography

[1] M. Acharyya, R. K. De, and M. K. Kundu. Extraction of features using M-band wavelet packet frame and their neuro-fuzzy evaluation for multitexture segmentation. *IEEE Trans. Patt. Anal. and Mach. Intelligence*, 25(12):1639–1644, Dec. 2003.

[2] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, and et al. The genome sequence of Drosophila melanogaster. *Science*, 287:2185–2195, 2000.

[3] N. I. Akhiezer and I. M. Glazman. *Theory of Linear Operators in Hilbert Spaces*, volume 1. Frederick Ungar, 1966.

[4] S. Arivazhagan, L. Ganesan, and T. G. S. Kumar. Texture classification using ridgelet transform. *Pattern Recogn. Letters*, 27(16):1875–1883, Dec. 2006.

[5] J. F. Aujol, G. Aubert, and L. Blanc-Féraud. Wavelet-based level set evolution for classification of textured images. *IEEE Trans. Image Proc.*, 12(12):1634–1641, Dec. 2003.

[6] R. R Bailey and M. Srinath. Orthogonal moment features for use with parametric and non-parametric classifiers. *IEEE Trans. Patt. Anal. and Mach. Intelligence*, 18(4):389–399, Apr. 1996.

[7] M. Barni and F. Perez-Gonzales. Pushing science into signal processing. *IEEE Signal Proc. Mag.*, Jul. 2005.

[8] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, second edition, 1993.

[9] G. Bi and Y. Zeng. *Transforms and Fast Algorithms for Signal Analysis and Representation*. Birkhäuser, Boston, MA, 2004.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[11] M. V. Boland, M. K. Markey, and R. F. Murphy. Classification of protein localization patterns obtained via fluorescence light microscopy. In *Proc. IEEE Int. Conf. EMBS Soc.*, pages 594–597, Chicago, IL, 1997.

[12] M. V. Boland, M. K. Markey, and R. F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, 33:366–375, Nov. 1998.

[13] M. V. Boland and R. F. Murphy. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics*, 17(12):1213–1223, 2001.

[14] H. Bölcskei and F. Hlawatsch. *Gabor Analysis and Algorithms: Theory and Applications*, chapter Oversampled modulated filter banks, pages 295–322. Birkhäuser, Boston, MA, 1998.

[15] H. Bölcskei and F. Hlawatsch. Oversampled cosine modulated filter banks with perfect reconstruction. *IEEE Trans. Circ. and Syst. II: Analog and Digital Signal Proc.*, 45(8):1057–1071, Aug. 1998.

[16] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger. Frame-theoretic analysis of oversampled filter banks. *IEEE Trans. Signal Proc.*, 46(12):3256–3269, Dec. 1998.

[17] R. K. Brayton, S. W. Director, G. D. Hachtel, and L. Vidigal. A new algorithm for statistical circuit design based on quasi-newton methods and function splitting. *IEEE Trans. Circ. and Syst.*, 26:784–794, 1979.

[18] J. Buckheit and D. L. Donoho. *Wavelets and Statistics*, volume 103, chapter Wavelab and reproducible research, pages 55–81. Springer-Verlag, 1995.

[19] A. E. Carpenter and D. M. Sabatini. Systematic genome-wide screens of gene function. *Nat. Rev. Genet.*, 5:11–22, 2004.

[20] P. G. Casazza and J. Kovačević. Equal-norm tight frames with erasures. *Adv. Comp. Math., sp. iss. Frames*, 18:387–430, 2002.

[21] P. M. Cassereau. A new class of optimal unitary transforms for image processing. Master's thesis, Massachussetts Inst. of Technology, May 1985.

[22] T. Chang and C. C. J. Kuo. Texture analysis and classification with tree-structured wavelet transform. *IEEE Trans. Image Proc.*, 2(4):429–441, Oct. 1993.

[23] A. Chebira, Y. Barbotin, C. Jackson, T. E. Merryman, G. Srinivasa, R. F. Murphy, and J. Kovačević. A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, 8(210), 2007. http://www.andrew.cmu.edu/user/jelenak/Repository/-07_ChebiraBJMSMK/07_ChebiraBJMSMK.html.

[24] A. Chebira, L. P. Coelho, A. Sandryhaila, S. Lin, W. G. Jenkinson, J. MacSleyne, C. Hoffman, P. Cuadra, C. Jackson, M. Püschel, and J. Kovačević. An adaptive multiresolution approach to fingerprint recognition. In *Proc. IEEE Int. Conf. Image Proc.*, volume 1, pages 457 – 460, San Antonio, TX, Sep. 2007.

[25] A. Chebira and J. Kovačević. Lapped tight frame transforms. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, volume III, pages 857–860, Honolulu, HI, Apr. 2007.

[26] A. Chebira, J. A. Ozolek, C. A. Castro, W. G. Jenkinson, M. Gore, R. Bhagavatula, I. Khaimovich, S. E. Ormon, C. S. Navara, M. Sukhwani, K. E. Orwig, A. Ben-Yehudah, G. Schatten, G. K. Rohde, and J. Kovačević. Multiresolution identification of germ layer components in teratomas derived from human and nonhuman primate embryonic stem cells. In *Proc. IEEE Int. Symp. Biomed. Imaging*, pages 979–982, Paris, France, May 2008.

[27] X. Chen, M. Velliste, and R. F. Murphy. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. *Cytometry*, 69 A:631–640, 2006.

[28] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser. Signal processing and compression with wavelet packets. Technical report, Yale Univ., 1991.

[29] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Th., sp. iss. Wavelet Transforms and Multiresolution Signal Analysis*, 38(2):713–718, Mar. 1992.

[30] T.F Coleman and Y. Li. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journ. Optimization*, 6:418–445, 1996.

[31] Committee on Ways and Means. Facts and figures: Identity theft. http://waysandmeans.house.gov/media/pdf/ss/factsfigures.pdf.

[32] C. Conrad, H. Erfle, P. Warnat, N. Daigle, T. Lorch, J. Ellenberg, R. Pepperkok, and R. Eils. Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res.*, 14:1130–1136, 2004.

[33] Z. Cvetković. Oversampled modulated filter banks and tight Gabor frames in $\ell^2(\mathbb{Z})$. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, pages 1456–1459, Detroit, MI, May 1995.

[34] Z. Cvetković and M. Vetterli. Oversampled filter banks. *IEEE Trans. Signal Proc.*, 46(5):1245–1255, May 1998.

[35] A. Danckaert, E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes. Automated recognition of intracellular organelles in confocal microscope images. *Traffic*, 3:66–73, 2002.

[36] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Commun. Pure and Appl. Math.*, 41:909–996, Nov. 1988.

[37] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, Philadelphia, PA, 1992.

[38] A. Depeursinge, D. Sage, A. Hidki, A. Platon, P. A. Poletti, M. Unser, and H. Müller. Lung tissue classification using wavelet frames. In *Proc. IEEE Int. Conf. EMBS Soc.*, pages 6259–6262, Lyon, France, Aug. 2007.

[39] R. G. Douglas. On majorization, factorization, and range inclusion of operators on hilbert space. *Trans. Amer. Math. Soc.*, 17:413–415, 1966.

[40] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, Englewood Cliffs, NJ, 2001.

[41] K. Fan and A. J. Hoffman. Some metric inequalities in the space of matrices. *Trans. Amer. Math. Soc.*, 6(1):111–116, Feb. 1955.

[42] H. G. Feichtinger and T. Strohmer, editors. *Gabor Analysis and Algorithms: Theory and Applications*. Birkhäuser, Boston, MA, 1998.

[43] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391:806–811, 1998.

[44] Y. Freund and R. Schapire. A short introduction to boosting. *Journ. Japanese Soc. Artif. Intelligence*, 14(5):771–780, 1999.

[45] D. Gabor. Theory of communication. *Journ. IEE*, 93:429–457, 1946.

[46] J. Gauthier, L. Duval, and J.-C. Pesquet. Low redundancy oversampled lapped transforms and application to 3D seismic data filtering. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, volume II, pages 821–824, Toulouse, France, May 2006.

[47] W. J. Germann and C. L. Stanfield. *Principles of Human Physiology*. Benjamin Cummings, San Francisco, CA, 2004.

[48] E. Glory and R. F. Murphy. Automated subcellular location determination and high throughput microscopy. *Developmental Cell*, 12:7–16, 2007.

[49] L. Gong, M. Puri, M. Ünlü, M. Young, K. Robertson, S. Viswanathan, A. Krishnaswamy, S. R. Dowd, and J. S. Minden. Drosophila ventral furrow morphogenesis: A proteomic analysis. *Development*, 131(3):643–656, 2004.

[50] V. K Goyal, J. Kovačević, and J. A. Kelner. Quantized frame expansions with erasures. *Journ. Appl. and Comput. Harmonic Analysis*, 10(3):203–233, May 2001.

[51] G. M. Haley and B. S. Manjunath. Rotation-invariant texture classification using a complete space-frequency model. *IEEE Trans. Image Proc.*, 8(2):255–269, Feb. 1999.

[52] D. Han and D. R. Larson. *Frames, bases and group representations*. Number 697 in Memoirs AMS. American Mathematical Soc., Providence, RI, 2000.

[53] R. M. Haralick. Statistical and structural approaches to texture. *Proc. IEEE*, 67:786–804, 1979.

[54] R. M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. 1973.

[55] P. Hennings Yeomans, J. Thornton, J. Kovačević, and B. V. K. V. Kumar. Wavelet packet correlation methods in biometrics. *Appl. Opt., sp. iss. Biometric Recogn. Systems*, 44(5):637–646, Feb. 2005.

[56] A. Hoberman. Doctor Biography. http://www.chp.edu/bio2/hoberman_a.php.

[57] A. Hoberman, C. D. Marchant, S. L. Kaplan, and S. Feldman. Treatment of acute otitis media consensus recommendations. *Clin. Pediatrics*, 41(6):373–390, 2002.

[58] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets, Time-Frequency Methods and Phase Space*, pages 289–297. Springer-Verlag, Berlin, Germany, 1989.

[59] L. Hong, Y. Wan, and A. K. Jain. Fingerprint image enhancement: Algorithm and performance evaluation. *IEEE Trans. Patt. Anal. and Mach. Intelligence*, 20(8):777–789, 1998.

[60] K. Huang and S. Aviyente. Mutual information based subband selection for wavelet packet based image classification. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, volume 2, pages 241–244, Philadelphia, PA, Mar. 2005.

[61] K. Huang and S. Aviyente. Sparse representation for signal classification. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 609–616. MIT Press, Cambridge, MA, 2007.

[62] K. Huang and R. F. Murphy. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics*, 5(78), 2004.

[63] K. Huang and R. F. Murphy. From quantitative microscopy to automated image understanding. *Journ. Biomed. Optics*, 9:893–912, 2004.

[64] J. Johnson and M. Püschel. In Search of the Optimal Walsh-Hadamard Transform. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Istanbul, Turkey, Jun. 2000.

[65] Z. Kam, J. S. Minden, D. A. Agard, J. W. Sedat, and M. Leptin. Drosophila gastrulation: Analysis of cell shape changes in living embryos by three-dimensional fluorescence microscopy. *Development*, 112:365–370, 1991.

[66] R. A. Kellogg, A. Chebira, A. Goyal, P. A. Cuadra, S. F. Zappe, J. S. Minden, and J. Kovačević. Towards an image analysis toolbox for high-throughput D rosophila embryo RNA i screens. In *Proc. IEEE Int. Symp. Biomed. Imaging*, pages 288–291, Arlington, VA, Apr. 2007.

[67] A. Khotanzad and Y.H. Hong. Invariant image recognition by zernike moments. *IEEE Trans. Patt. Anal. and Mach. Intelligence*, 12(5):641–647, 1990.

[68] S. C Kim and T. J. Kang. Texture classification and segmentation using incomplete tree structured wavelet packet frame and Gaussian mixture model. In *Proc. IEEE Int. Workshop Imag. Syst. Tech.*, pages 46–51, May 2005.

[69] N. G. Kingsbury. The dual-tree complex wavelet transform: A new efficient tool for image restoration and enhancement. In *Proc. Europ. Sig. Proc. Conf.*, pages 319–322, 1998.

[70] N. G. Kingsbury. Image processing with complex wavelets. *Phil. Trans. Royal Soc. London A*, Sep. 1999.

[71] N. G. Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Journ. Appl. and Comput. Harmonic Analysis*, 10(3):234–253, May 2001.

[72] J. Kovačević and A. Chebira. Life beyond bases: The advent of frames (Part I). *IEEE Signal Proc. Mag.*, 24(4):86–104, Jul. 2007.

[73] J. Kovačević and A. Chebira. Life beyond bases: The advent of frames (Part II). *IEEE Signal Proc. Mag.*, 24(5):115–125, Sep. 2007.

[74] J. Kovačević and A. Chebira. *An Introduction to Frames*. Foundations and Trends in Signal Processing. Now Publishers, 2008.

[75] P. Kovesi. Matlab and octave functions for computer vision and image processing. http://www.csse.uwa.edu.au/ pk/research/matlabfns/.

[76] A. Laine and J. Fan. Texture classification by wavelet packet signatures. *IEEE Trans. Patt. Anal. and Mach. Intelligence*, 15(5):1186–1191, May 1993.

[77] R. P. Lanza. *Essentials of Stem-Cell Biology*. Elsevier, 2006.

[78] S. Mallat. Wavelets for a vision. *Proc. IEEE*, 33:604–614, 1996.

[79] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[80] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer-Verlag, 2003.

[81] H. S. Malvar. *Optimal pre- and post-filtering in noisy sampled data systems*. PhD thesis, Massachussetts Inst. of Technology, Aug. 1986.

[82] H. S. Malvar. *Signal Processing with Lapped Transforms*. Artech House, Norwood, MA, 1992.

[83] H. S. Malvar. A modulated complex lapped transform and its applications to audio processing. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, pages 1421–1424, Phoenix, AZ, Mar. 1999.

[84] T. E. Merryman, K. Williams, G. Srinivasa, A. Chebira, and J. Kovačević. A multiresolution enhancement to generic classifiers of subcellular protein location images. In *Proc. IEEE Int. Symp. Biomed. Imaging*, pages 570–573, Arlington, VA, Apr. 2006.

[85] J. S. Minden. The Minden Lab. http://www.andrew.cmu.edu/user/minden/.

[86] A. Mintos, G. Srinivasa, A. Chebira, and J. Kovačević. Combining wavelet features with PCA for classification of protein images. In *Proc. Annual Biomed. Res. Conf. for Minority Students*, Atlanta, GA, Nov. 2005.

[87] R. F. Murphy. The Murphy Lab. http://murphylab.web.cmu.edu.

[88] National Science and Technology Council (NSTC) , Subcommittee on Biometrics. Biometrics.gov. http://www.biometrics.gov.

[89] C. S. Navara, J. D. Mich-Basso, C. J. Redinger, A. Ben-Yehudah, E. Jacoby, E. Kovkarova-Naumovski, and et al. Pedigreed primate embryonic stem cells express homogeneous familial gene profiles. *Stem Cells*, 25:2695–2704, 2007.

[90] T. Ogawa, J. M. Aréchaga, M. R. Avarbock, and R. L. Brinster. Transplantation of testis germinal cells into mouse seminiferous tubules. *Int. J. Dev. Biol.*, 41:111–122, 1997.

[91] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Patt. Anal. and Mach. Intelligence*, 24(7):971–987, Jul. 2002.

[92] J. A. Ozolek. Doctor Biography. http://www.chp.edu/bio2/ozolek_j.php.

[93] P. Perner, H. Perner, and B. Muller. Mining knowledge for Hep-2 cell image classification. *Journ. Artif. Intelligence in Medicine*, 26:161–173, 2002.

[94] S. Petushi, C. Katsinis, M. M. Haber, F. U. Garcia, and A. Tozeren. Large scale computations on histology images revels grade differentiating parameters for breast cancer. *BMC Medical Imaging*, 6(14), 2006.

[95] C. W. Pouton and J. M. Haynes. Embryonic stem cells as a source of models for drug discovery. *Nat. Rev. Drug. Discov.*, 6:605–616, 2007.

[96] K. Price. Annotated Computer Vision Bibliography. http://www.visionbib.com/bibliography/twod327.html.

[97] J. Princen, A. Johnson, and A. Bradley. Subband transform coding using filter bank designs based on time domain aliasing cancellation. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, pages 2161–2164, Dallas, TX, Apr. 1987.

[98] M. Püschel and J. Kovačević. Real, tight frames with maximal robustness to erasures. In *Proc. Data Compr. Conf.*, pages 63–72, Snowbird, UT, Mar. 2005.

[99] M. Püschel and M. Rötteler. The discrete triangle transform. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, volume 3, pages 45–48, 2004.

[100] S. Ramakrishnan and S. Selvan. Image texture classification using wavelet based curve fitting and probabilistic neural network. *Int. Journ. Imag. Syst. Tech.*, 17(4):266–275, 2007.

[101] S. Richards, Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M. J. Hubisz, R. Chen, R. P. Meisel, and et al. Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. *Genome Res.*, 15:1–15, 2005.

[102] W. B. Richardson. Nonlinear filtering and multiscale texture discrimination for mammograms. In *Proc. SPIE Conf. Math. Meth. in Medical Imaging*, pages 293–305, 1992.

[103] R. M. Rosenfeld, L. Culpepper, K. J. Doyle, K. M. Grundfast, A. Hoberman, M. A. Kenna, A. S. Lieberthal, M. Mahoney, R. A. Wahl, C. R. Woods, Jr., and B. Yawn. Clinical practice guideline: Otitis media with effusion. *Pediatrics*, 130(5):1412–1429, 2004. American Academy of Pediatrics.

[104] H. L. Royden. *Real Analysis*. Macmillan, Third edition, 1968.

[105] G. M. Rubin, M. D. Yandell, J. R. Wortman, G. L Gabor Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann, and et al. Comparative genomics of the eukaryotes. *Science*, 287:2204–2215, 2000.

[106] N. Saito and R. R. Coifman. Local discriminant bases and their applications. *Journ. Math. Imag. Vis.*, 5:337–358, 1995.

[107] N. Saito, R. R. Coifman, F. B. Geshwind, and F. Warner. Discriminant feature extraction using empirical probability density estimation and a local basis library. *Pattern Recogn.*, 35:2841–2852, 2002.

[108] R. J. Schalkoff. *Artificial Neural Networks*. Computer Science. McGraw-Hill, 1997.

[109] Z. Schmilovitch, V. Alchanatis, M. Shachar, and Y. Holdstein. Spectrophotometric otoscope: A new tool in the diagnosis of otitis media. *Journ. of Near Infrared Spectroscopy*, 15(4):209–215, 2007.

[110] C. Scott and R. D. Nowak. TEMPLAR: A wavelet-based framework for pattern learning and analysis. *IEEE Trans. Signal Proc.*, 52(8):2264–2274, Aug. 2004.

[111] I. W. Selesnick. *Wavelets in Signal and Image Analysis*, chapter The double density DWT. Kluwer Academic Publishers, 2001.

[112] I. W. Selesnick. The double-density dual-tree DWT. *IEEE Trans. Signal Proc.*, 52(5):1304–1314, May 2004.

[113] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury. The dual-tree complex wavelet transform. *IEEE Signal Proc. Mag.*, Nov. 2005.

[114] N. Shaikh. Differentiation of AOM and OME- The COMPT mnemonic. http://www.eprom.pitt.edu/34_viewPage.asp?pageID=78175985.

[115] K. Skretting and J. H. Husøy. Texture classification using sparse frame-based representations. *EURASIP Journ. Appl. Signal Proc.*, 2006:1–11, 2006.

[116] G. Srinivasa, A. Chebira, T. E. Merryman, and J. Kovačević. Adaptive multiresolution texture features for protein image classification. In *Proc. BMES Annual Fall Meeting*, Baltimore, MD, Sep. 2005.

[117] G. Srinivasa, T. E. Merryman, A. Chebira, A. Mintos, and J. Kovačević. Adaptive multiresolution techniques for subcellular protein location image classification. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, volume V, pages 1177–1180, Toulouse, France, May 2006.

[118] T. Strohmer. *Modern Sampling Theory: Mathematics and Applications*, chapter Finite and infinite-dimensional models for oversampled filter banks, pages 297–320. Birkhäuser, Boston, MA, 2000.

[119] H. Thomson. Bioprocessing of embryonic stem cells for drug discovery. *Trends in Biotechnol.*, 25:224–230, 2007.

[120] M. Unser. Texture classification and segmentation using wavelet frames. *IEEE Trans. Image Proc.*, 4:1549–1560, Nov. 1995.

[121] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, Englewood Cliffs, NJ, 1992.

[122] R. Vale and S. Waldron. Tight frames and their symmetries. *Const. Approx.*, 21:83–112, 2005.

[123] G. Van De Wouwer, P. Scheunders, S. Livens, and D. Van Dyck. Wavelet correlation signatures for color texture characterization. *Pattern Recogn.*, 32(3):443–451, Mar. 1999.

[124] A. Van Nevel. Texture classification using wavelet frame decompositions. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, volume 1, pages 311–314, Philadelphia, PA, Nov. 1997.

[125] M. Vetterli and J. Kovačević. *Wavelets and Subband Coding*. Signal Processing. Prentice Hall, Englewood Cliffs, NJ, 1995. http://waveletsandsubbandcoding.org/.

[126] J. W. Wang, C. H. Chen, and J. S. Pan. Genetic feature-selection for texture classification using 2-D nonseparable wavelet bases. *Trans. Inst. Electr. Comm. Eng.*, E81-A(8):1635–1644, Aug. 1998.

[127] C. Watson. NIST Special Database 24: Live-Scan Digital Video Fingerprint Database. Natl. Institutes of Health, Gaithersburg, Md., 1998. http://www.nist.gov/srd/nistsd24.htm.

[128] A. R. Weeks and G. E. Hague. Color segmentation in the HSI color space using the k-means algorithm. *Proc. SPIE*, 3026:143–154, Apr. 1997.

[129] M. V. Wickerhauser. INRIA lectures on wavelet packet algorithms. Technical report, Yale Univ., Mar. 1991.

[130] E. Yeh, K. Gustafson, and G. L. Boulianne. Green fluorescent protein as a vital marker and reporter of gene expression in Drosophila. In *Natl. Acad. Sci.*, volume 92, pages 7036–7040, 1995.

[131] R. W. Young and N. G. Kingsbury. Frequency-domain motion estimation using a complex lapped transform. *IEEE Trans. Image Proc.*, 2(1):2–17, Jan. 1993.

[132] S. Zappe, M. Fish, M. P. Scott, and O. Solgaard. Automated MEMS-based Drosophila embryo injection system for high-throughput RNAi screens. *Lab Chip*, 6:1012–1018, 2006.

[133] Z. Zhou and H. Peng. Automatic annotation and recognition of gene expression patterns of fly embryos. *BMC Bioinformatics*, 23:589596, 2007.

Abstract

This thesis presents a mathematical framework and an algorithm for the classification of biomedical image data sets based on adaptive and redundant multiresolution representations---frames. We illustrate the results on several different biomedical applications.

Classification is a ubiquitous problem in image processing; many biomedical tasks are in essence classification problems. Examples of such problems include determining a specific protein from its subcellular location pattern, determining the developmental stage of Drosophila embryos, recognizing tissue types in histological images of stem-cell teratomas, as well as determining otitis media stages. Though cumbersome, some of the above tasks, and many similar ones, are performed simply by visual inspection. As our eyes are not trained to extract statistical measures or time-frequency behavior of the signal across scales, these characteristics often pass unnoticed, resulting in poorer performance. We hypothesize that classifying adaptively in multiresolution subspaces will increase classification accuracy. We develop a new classifier, based on adaptive multiresolution ideas, by adding a multiresolution block in front of a generic classifier. The system is completed with a weighting block at the end, which plays the role of an arbiter; it decides how to combine the ``subspace'' decisions into a common one. The classifier achieves remarkable results, with most of the applications having classification accuracy in the mid-to high 90s.

In all of the applications, redundant multiresolution transforms performed the best. This led us to ask the following question: Why do frames perform better than bases? This question is nontrivial in scope, to begin to answer it we propose a classification scheme which uses finite frames and introduce a measure-theoretic framework for the analysis of classification errors. We then use this framework to examine those classes of signals for which a bases-based classification scheme is sufficient, and those for which a frame-based scheme is superior. We also show the proposed classification scheme performs well in the presence of noise.

Finally, as there are very few frame families available in the literature, we embarked on developing our own. To that end, we introduce a new class of frames we call lapped tight frame transforms, obtained by seeding from higher-dimensional orthonormal bases. We prove several properties of such frames, such as tightness, equal norm and maximal robustness.