INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

MATHEMATICAL AND STATISTICAL TECHNIQUES FOR MODELING AND ANALYSIS OF MEDICAL DATA

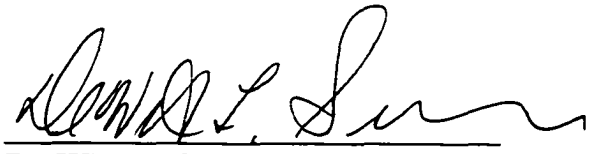
by

IVAYLO D. DINOV

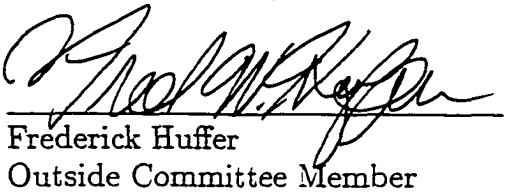
A Dissertation submitted to the
Department of Mathematics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 1998

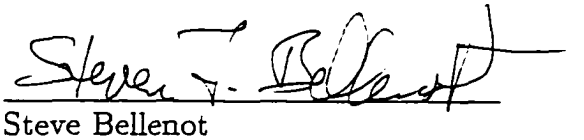
The members of the Committee approve the dissertation of Ivaylo D. Dinov defended on March 16, 1998.



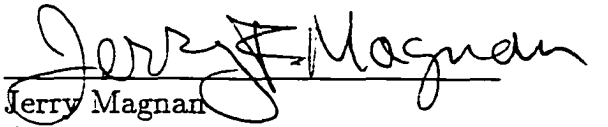
De Witt L. Sumners
Professor Directing Dissertation



Frederick Huffer
Outside Committee Member



Steve Bellenot
Committee Member



Jerry Magnan
Committee Member

ACKNOWLEDGMENTS

I would like to express my deep appreciation to my major professor, Dr. De Witt Sumners, for his helpful guidance and direction throughout my graduate study at Florida State University. I have learned a lot from Prof. Sumners about teaching, writing, conducting and presenting my research. Many of the techniques presented in this manuscript were suggested by Prof. Sumners.

I also wish to thank all members of my mathematics/statistics graduate committees; Prof. DW Sumners, Prof. F. Huffer, Prof. S. Bellenot, Prof. J. Magnan, Prof. X. Niu and Prof. I. McKeague for their supervision, ideas and suggestions in regard to the direction and the content of my dissertation.

I am indebted to Dr. A. Toga, Dr. P. Thompson, Dr. R. Woods and Dr. M. Mega from the UCLA Medical School; Dr. D. Rottenberg, K. Schaper and Dr. S. Strother at the PET Imaging Service of the VA Medical Center in Minneapolis; Dr. C-T Chen at the Medical School of the University of Chicago; Dr. M. Barnsley from the Iterated Systems Inc., and Prof. S. Awoniyi from the School of Engineering at FSU for their vital contributions to my education, for their support and encouragement.

My warm thanks extend to all faculty that lectured me in the Department of Mathematics and the Department of Statistics at FSU for motivating me, for helping me improve on my teaching and research strategies and for their mentorship.

This research was partially supported by NIH grant 1P20-MH57180.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Analysis and interpretation of medical data is oftentimes challenging due to intrinsic and extrinsic reasons. Two of these are presence of noise (random or deterministic) in the data, and the infinite anatomical, functional and physiological variability of the data obtained from seemingly identical sources. In particular, the study of the human brain, its structure and metabolism is cluttered by vast differences between any two brains and the unavoidable existence of noise, with (usually) unknown characteristics, mixed with the real signal.

In this manuscript we exploit two main problems; Identifying the regions of activation in single or multi-subject human brain functional studies, and quantifying the performance of various image-registration (warping) methods based of functional or structural single or group data. To address the first problem we develop a Sub-Volume Thresholding (SVT) technique that determines the statistically significant regions of activation in functional volumetric data (PET, SPECT, fMRI). The second problem is approached from a "transform-analytic" point of view; We convert the image-space warp-characterization problem into a wavelet or fractal (transform) space question. The reasons behind quantifying warp performance in transform space is that we can view various discrete transformations as good compression and denoising tools. For example, we propose a frequency-adaptive wavelet thresholding nonlinearity which exhibits some optimality properties in terms of stripping off the noise components of the signals and at the same time extracting the essence of the data in a robust and concise manner. We show

a number of examples illustrating the methodology and the results of applying our models for analyzing image-alignment of anatomical MRI data and determining the activation sites, under different stimulation paradigms, in functional PET and SPECT images.

INTRODUCTION

A common problem in the field of computerized medical (anatomical or functional) imaging is the automated interpretation and comparison of images acquired from different subjects. To facilitate the process of identifying the corresponding anatomical features, or regions of activation (for functional data), one co-registers and aligns (warps) the images to a standard reference image (target). When both images have been "warped" to the standard reference image, they can be compared with each other. In chapter I, we propose a quantitative scheme for evaluation of the performance of different image alignment techniques (warps). "Good" warps are ones which maintain a maximum amount of local image similarity between the initial image and its warp (thus minimizing the geometric distortion of the initial data), yet transforming it to an image similar to the target (of the warp). Our examples show quantitatively the uniform advantages of non-linear to the linear warps. We apply these methods to study various polynomial and other highly non-affine displacement fields.

In many clinical studies based on functional imaging there is a need for accurate, fast and robust techniques for extracting the "pure" signal from noisy data (image with low SNR - Signal-to-Noise-Ratio). For example, in Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI) the real signal may represent about 10% of the information content (Worsley *et al*, 1992), making data interpretation very difficult.

In the second chapter we introduce, implement and test a new technique (Sub-Volume/Sub-Image Thresholding) that allows for (spatially) *variable*

thresholding of the data based on prior anatomical and/or functional information about the images. This algorithm is less conservative than the Bonferroni's procedure test, but more conservative than a uniform $T - test$ on the entire image. We consider an image as being composed as the union of a number of sub-images (such a decomposition can reflect certain structural or physiological prior information about the data). For each sub-image we derive an estimate of the variance of its average, and use this estimate to determine the statistical significance of the activation over the sub-image of interest. For the sub-images where this significance level is high enough we search for the location(s) of the activation site(s). We also derive close mathematical forms for the correction factors of the variance estimates for rectangular (geometric type) partitioning. We discuss classes of permissible covariograms studied by Christacos [1984], Matern [1986], Cressie [1991] and others. We prove that the class of continuous functions we use in our Sub-Volume Thresholding (SVT) model induce valid covariance functionals. Then we present a number of examples illustrating the use of the SVT methodology.

A new wavelet thresholding scheme is introduced in the third chapter. This approach yields "almost-optimal" function estimators in the sense of achieving close approximations of the ideal risk. Combining knowledge from the areas of wavelet analysis, decision theory and parameter estimation we address the problem of numerical evaluation of different image-alignment algorithms on groups of volumetric data sets. Our "frequency-adaptive" soft-thresholding nonlinearity induces estimators whose risk values are within $\ln^2 N$ of the ideal polynomial risk using an oracle. This upper bound is similar to the one of Donoho and Johnstone, but it is obtained by a different "frequency-adaptive" wavelet shrinkage strategy.

CHAPTER I

TRANSFORM METHODS FOR ANALYZING STRUCTURAL IMAGES

An increasingly important problem in the area of medical imaging is pattern and feature recognition. Radiologists, neurologists, and cardiologists who study human anatomical or functional scans are faced with the same questions: What are the similarities and the differences between two images of the same modality? How can differences in these images be detected (identified and quantified), when the PSNR (Peak Signal to Noise Ratio) is very low (often the real signal contributes $\leq$ 10% of the data set, the rest is noise)? How do we find and match the statistically significant regions of activation in different functional data sets? This chapter we introduce a mathematical framework for approaching some of these questions.

We begin by reviewing the theory behind the discrete fractal transform (DFT) and the discrete wavelet transform (DWT). (Observe that we use the notation DFT for the fractal transform, which differs from its usual use for the Fourier transform). More mathematical detail is found in the following references: Barnsley, 1988; Daubechies, 1988; Dinov and Sumners 1996; Fisher, 1995; Mallat, 1989. In Sections 1 and 2 we discuss analysis and synthesis of signals using these discrete transforms. Our transform-based models are developed in Section 3. This section also contains the first application of the transform methods - quantitative comparison and identification of image similarities and dissimilarities. To illustrate these ideas we present two examples: One containing a set of 4 positron emission tomography (PET) brain scans of three different subjects; the second example contains 10 mag-

netic resonance imaging (MRI) brain scans of two subjects. In both cases we wish to see "grouping" and clustering effects induced by the numeric characterization of the transform models, and to detect which scans belong to which subjects.

Resolution enhancement and image magnification techniques are discussed in Section 4. We give examples to show the utility of the fractal based technique for blow-up of low-resolution images (like PET scans).

The most important application of our transform-based models is the quantitative evaluation of image registration techniques. Section 5 is devoted to classifying various 2D and 3D polynomial and other non-affine warping algorithms.

Before we begin, however, we would like to comment on the rationale behind using discrete transform approaches for the study of medical images: First, some transforms are "resolution independent" - the DFT, for instance. This means that once we encode (analyze) a signal we can recover (synthesize) it at any (higher or lower) resolution. The second point is that transforms of images allow comparison of unregistered signals. For example, a "more complex" version of the DFT could be developed which is rotation, scale and translation invariant. Third, both the DWT and the DFT are useful data compression techniques, so applying such transforms to the data can be viewed as a tool for reduction of data complexity. These transforms extract biologically relevant information from images in a rapid and concise manner which allows numeric quantization of the images and their similarities.

1. The Discrete Fractal Transform

1.1. Mathematical Preliminaries

To make this chapter self-contained we begin by stating a few definitions and some well-known results that provide the foundation for fractal approximations of signals.

Definitions 1.1. *Let (X, d) be a complete metric space and $U : X \longrightarrow X$ be a mapping, then:*

(1) $Lip(U) = \sup_{x_1 \neq x_2, x_i \in X} \frac{d(U(x_1), U(x_2))}{d(x_1, x_2)}$ *is called the* **Lipschitz constant** *of U;*

(2) U is called a **Lipschitz map** *if $Lip(U) < \infty$;*

(3) U is called a **Contractive map** *if $Lip(U) < 1$, Figure 1;*

(4) For $x \in X$ and $A \subseteq X$ the **distance** *between x and A is defined by $dist(x, A) = inf\{d(x, a) : a \in A\}$;*

(5) If $A, B \subseteq X$, the **Hausdorff metric**

$$h_d(A, B) = \sup\{dist(a, B), dist(b, A) \mid a \in A, b \in B\} =$$

$$= \max\{\inf\{\epsilon \mid B \subseteq Nbhd(A, \epsilon)\}, \inf\{\epsilon \mid A \subseteq Nbhd(B, \epsilon)\}\}$$

where $Nbhd(A, \epsilon) = \{x \in X | dist(x, A) < \epsilon\}$ is an $\epsilon-$neighborhood of the set $A \subseteq X$:

(6) U is a **Similarity map** *(similitude), if*

$$d(U(x_1), U(x_2)) = sd(x_1, x_2), \forall x_1, x_2 \in X$$

and the similarity factor s is strictly between zero and one.

Note that the Hausdorff metric (acting on subsets of X) can be very sensitive to small changes. For example, if $A = B \cup \{a\}$, where $a \notin Nbhd(B, \epsilon)$, then $h_d(A, B) \geq \epsilon$. Thus a difference of a single point could effect h_d. On the other hand, the usual metrics (acting on X, like the L^2 metric) are not so sensitive.

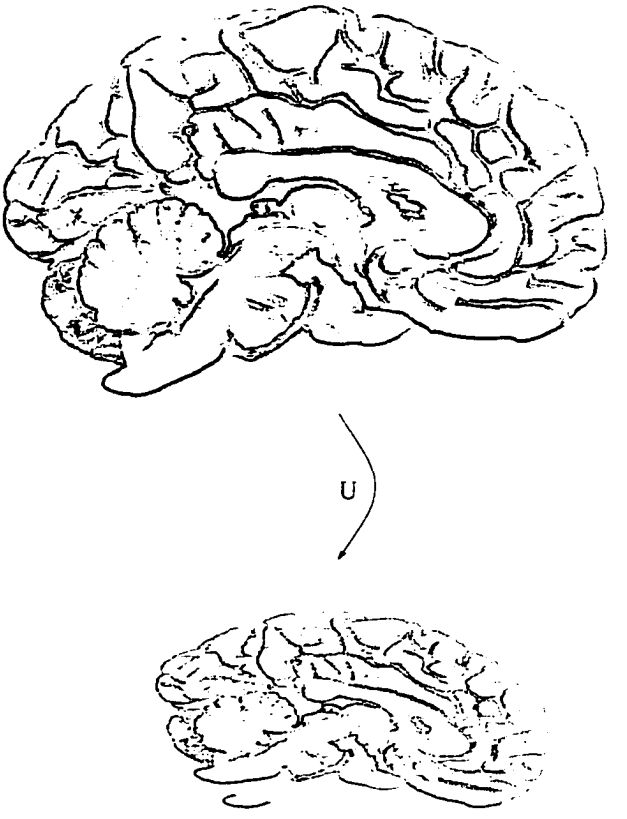


Figure 1. Example of a contractive map.

Lemma 1.2. *If (X, d) is a complete metric space and $H \equiv H(X) = \{S \subseteq X | S$ is compact$\}$, then (H, h_d) is a complete metric space, where h_d is the Hausdorff metric.*

Proof: [See Barnsley, 1988, Theorem 1, p.37].

Theorem 1.3. *[The Contractive Mapping Fixed-Point Theorem] If (X, d) is a complete metric space and $U : X \longrightarrow X$ is a contractive map then there exists a unique $\widehat{x} \in X$ such that $\widehat{x}$ is the fixed point (the attractor) of U, i.e.*

$$\widehat{x} = U(\widehat{x}) = \lim_{n \longrightarrow \infty} U(U(...(U(x_o))...))$$

for any $x_o \in X$.

Proof: [Fisher, 1995].

Theorem 1.4. *[Hutchinson, 1981, Theorem 3.2.(1)] Let (X, d) be a complete metric space and (H, h_d) be the set of all closed and bounded subsets of X, endowed with the Hausdorff*

metric, h_d. If $w_i : \Re^n \longrightarrow \Re^n$, $i = 1, 2, \cdots, k$, are contractive with Lipschitz constants $s_i < 1$, for all i, then $W = \bigcup_{i=1}^{k} w_i : H \longrightarrow H$ will also be contractive with Lipschitz constant $s = \max\{s_i : 1 \leq i \leq k\}$.

Note that $\forall A \in H$ $W(A) = \bigcup_{i=1}^{k} w_i(A)$, where $w_i(A) = \{w(a) | \ a \in A\}$. Also, because all w_i are contractions they are continuous mappings and $w_i(A) \in H$, $\forall i$. Thus $W(A) \in H$.

Theorem 1.5. *[Collage Theorem] Suppose (X, d) is a complete metric space and $U : X \longrightarrow X$ is a contractive map with a Lipschitz constant s and fixed point $\widehat{x}$. Then, for any $x \in X$*

$$d(x, \widehat{x}) \leq \frac{1}{1-s} d(x, U(x))$$

Proof: [Barnsley, 1993].

Example 1.6. *Skew Koch-type Curve. This is an simple example of a Hutchinson operator consisting of 3 similarity maps, each of which is a composition of translation, rotation and re-scaling, with (different) contraction coefficients all less then one, Figure 2. The three contractions map the set of the previous iteration onto a new set, according to the rotation, translation and re-scaling, induced by the three maps from the Initiator to the Generator. The final attractor looks very similar to the set obtained on the 7-th iteration.*

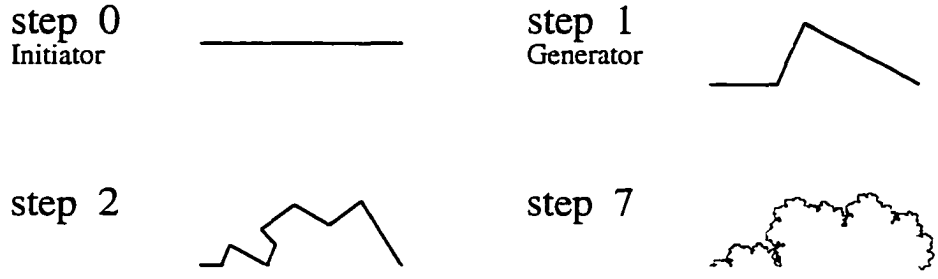


Figure 2. A Skew Koch-type curve.

Theorem 1.7. *Let E be a compact subset of R^n, $0 < s < 1$ and $\widehat{s} = \min(s, 1-s)$. Then, for any $\epsilon > 0$, there exists a positive integer m, and a collection of similarity maps $w_i : R^n \longrightarrow R^n$ with contraction coefficients $s_i < \widehat{s}$, $(d(w_i(x), w_i(y)) = s_i d(x, y), \forall x, y \in R^n)$, such that if $F = W(F)$ is the fixed point of the Hutchinson operator $W = \bigcup_{i=1}^{m} w_i$, then $h_d(E, F) < \epsilon$.*

Proof: Let $0 < s < 1$, $\widehat{s} = \min(s, 1-s)$, $\epsilon > 0$ and $E \in H$. Choose $m = m(\epsilon)$ so that there exists a collection of m balls $\{E_i = B(e_i, r_i)\}_{i=1}^{m}$, centered at $e_i \in E$ of radius $r_i < \frac{\widehat{s}\epsilon}{4}$ for $1 \leq i \leq m$ that cover E. Then

$$E \subseteq \bigcup_{i=1}^{m} E_i \subseteq Nbhd\left(E, \frac{\widehat{s}\epsilon}{4}\right). \tag{1}$$

For $1 \leq i \leq m$ let $w_i : E \longrightarrow E_i$ be any similarity with Lipschitz constant $s_i < \widehat{s}$ (note that we can choose such similarity maps since $0 < s_i < \widehat{s}$ and w_i need not be onto). Then $d(w_i(x), w_i(y)) = s_i d(x, y)$; $\quad x, y \in E$ and

$$w_i(E) \subseteq E_i \subseteq Nbhd\,(w_i(E), \widehat{s}\epsilon)\,. \tag{2}$$

The first inclusion in (2) is clear. To show the second one we use the triangle inequality for the metric h_d. Suppose $a \in E_i \setminus Nbhd\,(w_i(E), \widehat{s}\epsilon)$. Then two things follow: first $d(a, e_i) < r_i < \frac{\widehat{s}\epsilon}{4}$, where $E_i = B(e_i, r_i)$; and second - for all $e \in E$, $d(a, w_i(e)) \geq \widehat{s}\epsilon$ (remember the metrics d and h_d coincide for singleton sets). We obtain the following contradiction

$$\widehat{s}\epsilon \leq d(a, w_i(e_i)) = h_d(a, w_i(e_i)) \leq h_d(w_i(a), w_i(e_i)) + h_d(w_i(a), a) <$$

$$< 2r_i + 2r_i < \frac{\widehat{s}\epsilon}{2} + \frac{\widehat{s}\epsilon}{2} = \widehat{s}\epsilon,$$

since $a, e_i, w_i(a), w_i(e) \in E_i$ and $diam(E_i) = 2r_i < 2\frac{\widehat{s}\epsilon}{4}$. This shows that $E_i \subseteq Nbhd\,(w_i(E), \widehat{s}\epsilon)$. Finally, using (2) and (1) we have that

$$\bigcup_{i=1}^{m} w_i(E) \subseteq \bigcup_{i=1}^{m} E_i \subseteq Nbhd\left(E, \frac{\widehat{s}\epsilon}{4}\right) \tag{3}$$

$$E \subseteq \bigcup_{i=1}^{m} E_i \subseteq \bigcup_{i=1}^{m} Nbhd\,(w_i(E), \hat{s}\epsilon) \tag{4}$$

Using (3) and (4) and the properties of the Hausdorff metric we obtain $h_d\left(E, \bigcup_{i=1}^{m} w_i(E)\right) < \hat{s}\epsilon$. Thus, the Collage theorem yields:

$$h_d(E, F) \leq \frac{1}{1-s} h_d\left(E, \bigcup_{i=1}^{m} w_i(E)\right) \leq \frac{1}{1-s}\hat{s}\epsilon \leq \epsilon,$$

where $F = W(F) = \bigcup_{i=1}^{m} w_i(F)$ is the invariant set of W.

Δ

These results provide the theoretical foundation for approximating sets in R^n by fractal sets (Hutchinson, 1981, points out that often these approximating sets (attractors) would have non-integer dimension and thus we call them "fractal approximation" sets).

We conclude this section with an example illustrating the power of the contractive iterative dynamical systems as a signal approximation tool.

Figure 3. An iterative contractive dynamical system.

Figure 3 shows part of the dynamical system, induced by the Fractal transform (reverse quadtree partitioning), the attractor of which is a fractal image approximating the original magnetic resonance image (MRI) in the top-left corner. The relatively simple description of the fractal signal is the basis of the numeric image characterization we will develop in the later

sections. (Top Row, left-to-right: Original, Initial fractal approximation (null-signal), first step of the fractal approximation, third step; Bottom Row, left-to-right: tenth iteration, 50-th iteration, final fractal approximation, and difference between the original image and its fractal approximation.)

1.2 The "Inverse Problem"

So far we discussed the problem of identifying the invariant sets of contractive maps. Now, we will address what has come to be known as the "Inverse Problem": Given a set (curve, image) find a (small) collection of contractive maps, the fixed point (set) of which is a (fractal) set approximating closely the given set. M. Barnsley was among the first to identify and partially solve this problem in a Collage theorem setting: If $W = \{w_i\}_{i=1}^n$ are the desired contractive mappings and x is the given set, then by Theorem 1.5,

$$d(x, \widehat{x}) \leq \frac{1}{1-s} d(x, W(x))$$

where $\widehat{x}$ is the attractor of W ($W(\widehat{x}) = \widehat{x}$) and s is the contractivity of W. So, loosely speaking, we minimize the "collage" difference (on the right), to find W, and thus find $\widehat{x}$ close to x. Theorem 1.7 can be used as a foundation and a motivation for the set-approximation problem. For approximating images (pictures) we use the following two claims as a starting point. In this manuscript, the collection of maps W that minimizes the collage difference $d(x, W(x))$, (whether x is a set or an image) according to some optimization technique, is called **the Discrete Fractal Transform** (DFT) of x, and the attractor, $\widehat{x}$, of the dynamical system induced by W is called **the Inverse Fractal Transform** (IFT) of x.

Our study continues with 2 dimensional images and their fractal transforms (this is readily generalizable to n-dimensional images).

Definition 1.8. *The* **Image Space:** $S_1 = L_2(I \times I)$ *is the space of the square integrable (measurable) real-valued functions in the unit square, endowed with the norm* $\delta(f) = ||f||_2 = \sqrt{\int_{I\times I} f^2 dx}$ *[Folland, 1984].*

Definition 1.9. *The* **Transform Space:**

$$S_2 = \{\textit{continuous mappings} : W : S_1 \longrightarrow S_1\}.$$

Definition 1.10. $R = \{R_i\}_{i=1}^m$ *is called a* **partition** *of* $I^2 = I \times I$ *if the* R_i *are non-empty measurable subsets of the unit square* I^2 *that intersect (pairwise) only in sets of measure zero, and* $I^2 = \bigcup_{i=1}^{m} R_i$.

Definition 1.11. *If* $A : R^n \longrightarrow R^n$ *is a linear operator (a matrix) and* $||.||$ *denote the usual Euclidean norm on* R^n*, then the* **norm of** A *is*

$$||A|| = \sup_{||\mathbf{x}||\neq 0} \frac{||A\mathbf{x}||}{||\mathbf{x}||}.$$

Claim 1.12. *If* $A = diag\{\lambda_i\}_{i=1}^n$*, then* $||A|| = \max_{1\leq i\leq n}\{|\lambda_i|\}$.

Proof: If $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, then $A\mathbf{x} = (\lambda_1 x_1, \lambda_2 x_2, \cdots, \lambda_n x_n)$ and

$$\sup_{||\mathbf{x}||\neq 0} \frac{||A\mathbf{x}||}{||\mathbf{x}||} = \sup_{||\mathbf{x}||\neq 0} \sqrt{\frac{\lambda_1^2 x_1^2 + \lambda_2^2 x_2^2 + \cdots + \lambda_n^2 x_n^2}{x_1^2 + x_2^2 + \cdots + x_n^2}} \geq \sqrt{\lambda_i^2} = |\lambda_i|$$

for all $1 \leq i \leq n$ (by letting $x_i = 1$ and $x_j = 0, j \neq i$). However, if $|\lambda_{i_o}| = \max_{1\leq i\leq n}\{|\lambda_i|\}$,

$$\sup_{||\mathbf{x}||\neq 0} \frac{||A\mathbf{x}||}{||\mathbf{x}||} \leq \sup_{||\mathbf{x}||\neq 0} \sqrt{\frac{\lambda_{i_o}^2 (x_1^2 + x_2^2 + \cdots + x_n^2)}{x_1^2 + x_2^2 + \cdots + x_n^2}} = |\lambda_{i_o}|$$

Hence, $||A|| = \sup_{||\mathbf{x}||\neq 0} \frac{||A\mathbf{x}||}{||\mathbf{x}||} = \max_{1\leq i\leq n}\{|\lambda_i|\}$.

Δ

Claim 1.13. *If* M *is an orthogonal matrix, then* $||M|| = 1$.

Proof: $< M\mathbf{x}, M\mathbf{y} > = < \mathbf{M^T M x}, \mathbf{y} > = < \mathbf{x}, \mathbf{y} >$, in particular, $||M\mathbf{x}||^2 = ||\mathbf{x}||^2$. Thus, $||M|| = \sup_{||\mathbf{x}||\neq 0} \frac{||M\mathbf{x}||}{||\mathbf{x}||} = \sup_{||\mathbf{x}||\neq 0} \frac{||\mathbf{x}||}{||\mathbf{x}||} = 1$.

Δ

Claim 1.14. *If A is a symmetric matrix then $||A|| = \max_{1 \le i \le n} \{|\lambda_i|\}$, where $\{\lambda_i\}_i$ are the eigenvalues of the matrix A.*

Proof: This follows from the fact that a symmetric matrix is diagonalizable by an orthogonal matrix [Schneider, 1987]. Thus, $A = M \Lambda M^T$, where M is orthogonal ($||M|| = ||M^{-1}|| = 1$) and $\Lambda = diag(\lambda_i)$, where $\{\lambda_i\}$ are the eigenvalues of A. Since M is a bijective linear transformation

$$||A|| = \sup_{||\mathbf{x}|| \neq 0} \frac{||A\mathbf{x}||}{||\mathbf{x}||} = \sup_{||M\mathbf{x}|| \neq 0} \frac{||AM\mathbf{x}||}{||\mathbf{M}\mathbf{x}||} =$$

$$= \sup_{||\mathbf{x}|| \neq 0} \frac{||M\Lambda\mathbf{x}||}{||\mathbf{x}||} = \sup_{||\mathbf{x}|| \neq 0} \frac{||\Lambda\mathbf{x}||}{||\mathbf{x}||} = \max_i \{|\lambda_i|\}$$

Δ

Definition 1.15. *The non-negative number $\sigma(A) = \max_i \{|\lambda_i|\}$, where $\{\lambda_i\}$ are the eigenvalues of the (arbitrary) matrix A, is called spectral radius of A.*

The spectral radius of a linear operator plays an important role in functional analysis when one wants to investigate iterative methods [Zeider, 1985].

Claim 1.16. *For any matrix A, $||A|| = \sqrt{\sigma(A')} = \sqrt{\max_i \{|\lambda_i'|\}}$, where $\{\lambda_i'\}$ are the eigenvalues of the matrix $A' = A^T A$.*

Proof: First note that A' is symmetric: $(A')^T = (A^T A)^T = A^T (A^T)^T = A^T A = A'$. Let $\{u_i\}$ be an orthonormal set of eigenvectors of $A' = A^T A$. Then $\forall \mathbf{x} \in \mathbf{R^n}$, $\mathbf{x} = \sum_{i=1}^{n} \alpha_i(\mathbf{x})\mathbf{u_i}$ and $A^T A\mathbf{x} = \mathbf{A'x} = \sum_{i=1}^{n} \alpha_i(\mathbf{x})\mathbf{A'u_i} = \sum_{i=1}^{n} \alpha_i(\mathbf{x})\lambda_i\mathbf{u_i}$. Expanding the quotient $\frac{||A\mathbf{x}||^2}{||\mathbf{x}||^2} = \frac{\mathbf{x^T A^T A x}}{\mathbf{x^T x}} = \frac{\sum_{i=1}^{n} |\alpha_i(\mathbf{x})|^2 |\lambda_i| ||\mathbf{u_i}|^2}{\sum_{i=1}^{n} |\alpha_i(\mathbf{x})|^2} = \frac{\sum_{i=1}^{n} |\alpha_i(\mathbf{x})|^2 |\lambda_i|}{\sum_{i=1}^{n} |\alpha_i(\mathbf{x})|^2}$. Therefore, $||A||^2 = \sup_{||\mathbf{x}|| \neq 0} \frac{||A\mathbf{x}||^2}{||\mathbf{x}||^2} = \sigma(A')$, since for all i, if $\mathbf{x} = \mathbf{u_i}$, then $||A||^2 \geq |\lambda_i|$. And if $|\lambda_{i_o}|^2 = \max\{|\lambda_i|\}$, then obviously $||A||^2 \leq |\lambda_{i_o}|$.

Δ

These simple results will play an important role when we talk in details about the contractivity of the maps $w_{[m,n]}$, which will turn out to depend only on the norms of their linear factors ($A_{[m,n]}$).

Let $R = \{R_i\}_{i=1}^k$ be a fixed partition on I^2. Suppose we are given $D_i \subseteq I^2$ and affine maps $v_i : D_i \longrightarrow R_i$, $v_i(x) = A_i(x) + b$, which are contractive, ($0 < ||A_i|| < 1, \forall i$), one-to-one and onto. Then for all i, (based on the L_2-metric), we showed that $||A_i|| = \sqrt{\max_{1 \le j \le n} \{|\lambda'_{j,i}|\}}$, where $\{\lambda'_{j,i}\}$ are the eigenvalues of the symmetric matrix $A'_i = A_i^T A_i$. So, to insure that the maps v_i bring points closer together in space we only require that all of the eigenvalues of the matrices A'_i are less than one (in absolute value).

Note: This contractivity requirement

$$\sigma(A') = \max_{\{1 \le j \le n\}, \{1 \le i \le k\}} \{|\lambda'_{j,i}|\} < 1 \tag{5}$$

yields also that $1 > ||A_i|| > |det(A_i)|$ for each i. For clearly,

$$1 > ||A_i||^2 = \max_{1 \le j \le n} \{|\lambda'_{j,i}|\} \ge |\lambda'_{j_{max},i}| \prod_{j \ne j_{max}} |\lambda'_{j,i}| = \left| \prod_{1 \le j \le n} \lambda'_{j,i} \right| = |det(A'_i)| =$$
$$= |det(A_i^T)||det(A_i)| = |det(A_i)|^2$$

This fact will be used in Proposition 1.18 to show that the Hutchinson operator W (defined below) is contractive on S_1. In practice, the matrices A_i are often symmetric (or even diagonal) and (5) is easy to verify.

Definition 1.17. *Let $\{s_i\}_{i=1}^k$ be a collection of contraction* **scaling** *coefficients ($0 < s_i < 1, \forall i$), and $\{o_i\}_{i=1}^k$ be a family of* **offset** *factors. For $f \in S_1$ define $W : S_1 \longrightarrow S_1$ by*

$$W = \bigcup_{i=1}^k w_i \quad \text{and} \quad W(f) = \bigcup_{i=1}^k w_i(f)$$

where $w_i(f_{/D_i})_{/R_i} = s_i f(v_i^{-1})_{/R_i} + o_i$.

Note: W is well-defined since a composition of measurable functions is measurable and $W(f)$ has a finite L_2 norm because $f \in S_1$. In fact, to avoid

misinterpretation in the above definition we use the "lowest-index-maximum-priority" scheme when we evaluate the resulting image ($W(f)$) along the intersections (boundaries) of partitioning sets. With this restriction it is obvious that W is well-defined and (by Proposition 1.18) contractive. Hence, W has a unique fixed point in the image space.

Proposition 1.18. *The Hutchinson operator W defined above is a contraction on the image space, S_1.*

Proof: Let $f, g \in S_1$. Then $\delta^2(W(f), W(g)) =$

$$= \|W(f) - W(g)\|_2^2 = \int_{I^2} (W(f)(x) - W(g)(x))^2 \, dx =$$

$$= \sum_{i=1}^{k} \int_{R_i} (W(f)(x) - W(g)(x))^2 \, dx = \sum_{i=1}^{k} \|s_i f(v_i^{-1})_{/R_i} + o_i - s_i g(v_i^{-1})_{/R_i} - o_i\|_2^2 =$$

$$= \sum_{i=1}^{k} |s_i| . \|f(v_i^{-1})_{/R_i} - g(v_i^{-1})_{/R_i}\|_2^2 \leq s_m \sum_{i=1}^{k} \|f(v_i^{-1})_{/R_i} - g(v_i^{-1})_{/R_i}\|_2^2,$$

where $s_m = \max\limits_{1 \leq i \leq k} |s_i| < 1$. Therefore,

$$\delta^2(W(f), W(g)) < s_m \sum_{i=1}^{k} \int_{R_i} \left(f(v_i^{-1}(x)) - g(v_i^{-1}(x))\right)^2 dx$$

Changing the variables, $x = v(y)$,

$$\delta^2(W(f), W(g)) < s_m \sum_{i=1}^{k} \int_{D_i} (f(y) - g(y))^2 \, |det(A_i)| dy =$$

$$= s_m \sum_{i=1}^{k} |det(A_i)| \sum_{j=1}^{k} \int_{D_i \bigcap R_j} (f(y) - g(y))^2 \, dy \leq$$

$$\leq s_m A_m \sum_{j=1}^{k} \sum_{i=1}^{k} \int_{D_i \bigcap R_j} (f(y) - g(y))^2 \, dy,$$

where $A_m = \max\limits_{1 \leq i \leq k} |det(A_i)| < \max\limits_{1 \leq i \leq k} \|A_i\| < 1$, by the choice of the maps v_i. Finally,

$$\delta^2(W(f), W(g)) < s_m A_m \sum_{j=1}^{k} \int_{R_j \bigcap \left(\bigcup\limits_{i=1}^{k} D_i \right)} (f(y) - g(y))^2 \, dy \leq$$

$$\leq s_m A_m \sum_{j=1}^{k} \int_{R_j} (f(y) - g(y))^2 \, dy = s_m A_m \delta^2(f, g)$$

This also yields that the contractivity factor of W is $s_m A_m$.

Δ

Probably the best known example of an image encoded by a small collection of contractive maps is the Fern of M. Barnsley [Barnsley, 1988]. Using only four affine similarity maps [Fisher, 1995] Barnsley was able to produce an self-symmetric attractor resembling a real fern. Figure 4, shows the first 3 recursive iterations of the four contractions and the final (stochastically obtained) fractal fixed point.

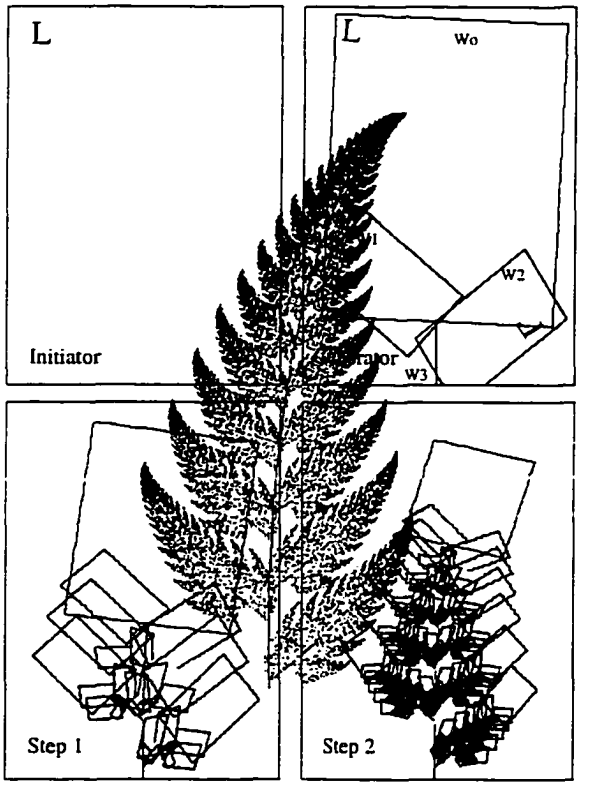


Figure 4. Barnsley's Fern.

Claim 1.19. *Fractal signals (images that could be realized as attractors of contractive (Hutchinson) operators on S_1) are dense in the signal space. That is, for every image in S_1, there is an image that is the fixed point of a contractive mapping on S_1, and which is arbitrarily close to the initial image.*

Proof: This is a trivial result, based on the fact that for any $f \in S_1$ and any $\epsilon > 0$, there exists a partition $R = \{R_i\}_{i=1}^n$ and a step function ϕ_n, approximating f, with

$$\phi_n = \sum_{i=1}^{n} \alpha_i \chi_{R_i}$$

where χ_{R_i} is the indicator function on the set R_i, and $\delta(f, \phi_n) < \epsilon$. To realize ϕ_n as an attractor of a contractive map on S_1 we use the induced partitioning $R = \{R_i\}$. Let $W : S_1 \longrightarrow S_1$ be $W(g) = \bigcup_{i=1}^{n} w_i(g) = \phi_n, \forall g \in S_1$. So, the spatial part (v_i) of each w_i contracts the domain of g onto R_i. The intensity part of w_i has scaling and offset coefficients $s_i = 0$ and $o_i = \alpha_i$.

Δ

Note: The above claim is only used as a motivational example of what follows. The trivial contractive map we constructed requires a very large number of partitioning sets $\{R_i\}$, and is not interesting for either contractivity, fractal image representation or image compression. In practice, for every R_i we search for a non trivial contractive (covering) map that involves fewer (unknown) parameters (D_i, v_i, s_i, and o_i). Also, one can observe that Claim 1.19 is s special case of Theorem 1.7, where $f \in S_1 = L_2[I^2]$, $E = \{f\}$ is a singleton (compact) subset of S_1 and we are searching for $\widehat{f} \in S_1$ so that $\delta(f, \widehat{f})$ is small.

The following two results are interesting from the mathematical point of view. They amount to saying that a signal and its inverse fractal transform have the same fractal representation. This justifies the name of the inverse fractal transform since it undoes what the fractal transform does.

Claim 1.20. *Let W be a deterministic Fractal Transform based on a particular partitioning scheme. Let S_1 be the space of all functions f with compact graphs in $\Re^n$ such that Wf is*

contractive. Then the following diagram commutes:

$$\begin{array}{ccc} & & S_2 \times S_1 \\ & \widehat{FD} \swarrow & \downarrow p \\ S_1 & \underset{W}{\longrightarrow} & S_2 \end{array}$$

where: $W(f) = Wf$, the fractal transform of f, contains the symmetries and self similarities of f; $\widehat{FD}(Wf, h) = \hat{f}$, with $\hat{f}$ being the fixed point (attractor) of the iterative fractal decoding map, $(Wf(\hat{f}) = \hat{f})$; And p is the projection on S_2 ; $p(Wf, h) = Wf$, for all $(Wf, h) \in S_2 \times S_1$.

Proof: First we note that every fractal transform can be slightly "modified" to induce a unique deterministic $n-$tuple of symmetries for every signal $f \in S_1$. This can be done, for instance, by minimizing the distance between $R[m, n]$ and all possible coverings $D[i, j]$ and then out of the remaining ones choosing the one that is the "closest" to the line determined by the upper-left corner of $R[m, n]$ and $(0, 0)$.

Let (Wf, h) be in $S_2 \times S_1$, then $p(Wf, h) = Wf$, and $W\left(\widehat{FD}(Wf, h)\right) = W(\hat{f}) = W\hat{f}$. Thus, to show that the diagram commutes we need only show that the fractal transforms of f and $\hat{f}$ are equal, that is $Wf = W\hat{f}$.

For, since $\hat{f}$ is the fixed point of the iterative fractal decoding (FD) of f we have $Wf(\hat{f}) = \hat{f}$. By the definition of Wf, see section 2,

$$\hat{f}_{/R[m,n]} = s[m][n]\hat{f}(v^{-1})_{/R[m,n]} + r[m][n]$$

Therefore, $\delta(\hat{f}_{/R[m,n]}, s[m][n]\hat{f}(v^{-1})_{/R[m,n]} + r[m][n]) = 0$. Recall, that δ is a metric and hence $\delta \geq 0$. We obtain that the $5-$tuple of the symmetries of f, (i, j, r, s, r), minimizes the functional δ. But that corresponds (by the uniqueness of the $5-$tuple symmetries) to finding the fractal transform of $\hat{f}$. Thus, the symmetries and self-similarities of $\hat{f}$ are the same as those of f and $W\hat{f} = Wf$.

Δ

Claim 1.20 can be called the "Fractal Idempotency Lemma", because it implies that such fractal transform techniques are idempotent (recall that an operator P is idempotent if $P^2 = P$). And this leads us to the following somewhat interesting result:

Theorem 1.21. *Suppose we fix a particular FT partitioning scheme. Let* $f \in S_1 = \{f \mid G(f) \subseteq \Re^n,$ *compact,* $Wf -$ *contractive*$\}$. *Define the equivalence class of* f *in* S_1 *by:*

$$[f] = \{h \in S_1 \mid Wh \equiv Wf \}.$$

Let $[S_1] = \{[f] | f \in S_1\}$ *be the space of equivalence classes. Then* $[f] = [\widehat{f}]$, *where as usual* $\widehat{f}$ *is the unique attractor of* Wf $(Wf(\widehat{f}) = \widehat{f})$. *Moreover,* $IFT : S_2 \longrightarrow S_1$ *defined by:* $IFT(Wf) = [\widehat{f}]$ *is the inverse of* $FT : [S_1] \longrightarrow S_2$.

Proof: The proof is a one-liner using *Claim 1.20*:

$$\begin{array}{ccccccc} & W & & IFT & & W = FT & \\ [S_1] & \longrightarrow & S_2 & \longrightarrow & [S_1] & \longrightarrow & S_2 \\ [f] & \longrightarrow & Wf & \longrightarrow & [f] \equiv [\widehat{f}] & \longrightarrow & Wf \equiv W\widehat{f}. \end{array}$$

Δ

The above results extend trivially to the case of non-affine maps $\{w_i\}_{i=1}^N$ as long as these are still contractive.

1.3 Practical Implementation - Reverse Quadtree Partitioning

In this section we discuss the fractal transform, from a practical point of view, for discrete digital images. A Discrete Fractal Transform (DFT) of a signal is a technique that searches for the self-similarities and self-symmetries of an image. It can be thought of as an encoding map $DFT : S_1 \longrightarrow S_2$, where S_1 is the space of the n–dimensional images (think of $L^p([0,1] \times [0,1])$) and S_2

is the symmetry space of the signals (see below). This mapping considers an image as being formed by copies of parts of itself (possibly rescaled and translated), up to some intensity-level corrections. Those similar copies along with measures of self-similarities and self-symmetries are detected and stored in the symmetry space S_2 by the DFT. Figure 5 shows an example of an *affine* $(w(x)Ax + b : D \longrightarrow R)$ self-similarity on a PET image.

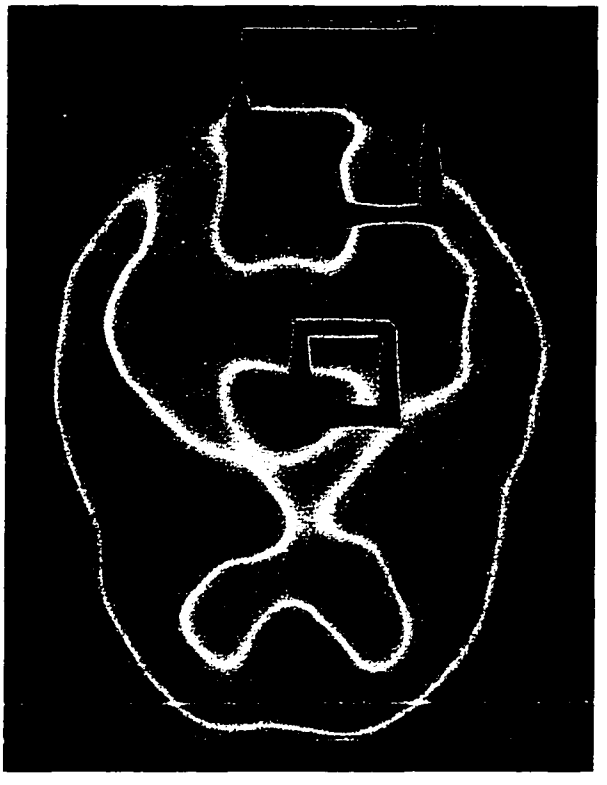


Figure 5. Example of an affine self-similarity.

To find the DFT of a signal we need to model the signal as a function. For simplicity we will think of a signal as being a map

$$f : I^2 \longrightarrow \Re$$

where I^2 is (say) a $2^N \times 2^N$ discrete square lattice. The functional values $f(i,j)$ represent the intensities of the signal at the $(i,j)^{th}$ pixel.

We define a metric on the space S_1 (of the signals), called the RMS - root

mean square -

$$\delta(f,g) = \sqrt{\left(\frac{1}{2^{2N}}\right) \sum_{(i,j)\in I^2} (f(i,j) - g(i,j))^2}.$$

One way to define a discrete algorithm for identifying image similarities is the following: Let $R = \{R[m,n,s]\}$ be a quadtree family of partitions (*ranges*) of (the domain) I^2, Figure 6, where $m = m_1 2^s, n = n_1 2^s, 1 \leq m_1, n_1 \leq 2^{N-s}$. So, that for a fixed s the $R[m,n,s]$'s are squares of (edge)-size 2^s with upper-left corner at $[m,n]$. They cover I^2, and intersect at most along their boundaries. Also, let $D = \{D[i,j,s]\}$ be the collection of squares in I^2 of (edge)-size 2^{s+1} (*domains*), with upper-left corners at $[i,j]$. Then for any $R[m,n,s]$, in R, we will find a $D[i,j,s]$ in D, so that the image of f over $D[i,j,s]$ resembles closely the image of f over $R[m,n,s]$. This can be done (using affine transformations, for example) by minimizing the functional:

$$\delta(f_{/R[m,n,s]}, w_{[m,n,s]}(f_{/R[m,n,s]})), \tag{6}$$

where $w_{[m,n,s]}\begin{pmatrix} i \\ j \\ z \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & s_{[m,n,s]} \end{pmatrix}\begin{pmatrix} i \\ j \\ z \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ o_{[m,n,s]} \end{pmatrix}$. The matrix $A = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & s_{[m,n,s]} \end{pmatrix}$ is called the <u>linear part</u> of the affine map $w_{[m,n,s]} : D[i,j,s] \times \Re \longrightarrow R[m,n,s] \times \Re$. Note that $z = f(i,j)$ and $w_{[m,n,s]}(f) = w_{[m,n,s]}(i,j,f)$. As before, $v_{[m,n,s]} : D[m,n,s] \longrightarrow R[m,n,s]$ is the affine spatial part of the map $w_{[m,n,s]}$, so that $w_i(f)_{/R[m,n,s]} = s_{/R[m,n,s]} f(v^{-1}_{[m,n,s]})_{/R[m,n,s]} + o_{/R[m,n,s]}$. A note of caution: the fractal transform of the image f, (W_f), would be well-defined provided we take special care to define the operator $W_f : S_1 \longrightarrow S_1$, so that over the set of intersection points of the ranges R_i (boundary intersection, set of measure zero) $W_f(f_o)$ makes sense as an element of S_1. The map $v_{[m,n,s]}$ that minimizes the functional (6) is called a <u>cover</u> of $R[m,n,s]$, sometimes the domain $D[m,n,s]$ (or even $w_i(f)_{/R[m,n,s]}$) is loosely referred to as the cover of $R[m,n,s]$. Note that the identity maps (w_i) trivially minimize (6), however, they are not useful for our purposes since they are not contractions and

amount only to a step function approximation, see Claim 1.19. We now concentrate on finding non-trivial solutions to the optimization problem (6).

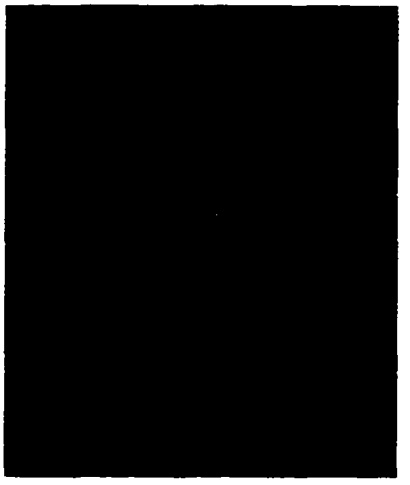


Figure 6. A contour plot of a PET image
with background Quadtree partitioning.

The Reverse Quadtree Partitioning targets best possible encoding (decoding), inducing a convergent (contractive) fractal transform. We start with the smallest size partition $\{R[m,n,s]\}$ - usually 2×2 ($s = 1$) - and attempt to cover each $R[m,n,1]$ by a $D[i,j,1]$ of size 4×4 with scaling coefficients $|s_{[m,n,1]}| < 1$. If this can not be done for some $R[m,n,1]$ we replace the $R[m,n,1]$ along with its 3 immediate "closest" neighbors by a new square $R' = R[m,n,2]$ of size 4×4 containing the 4 smaller neighbors. This way we are going from smaller in size $R[m,n,s]$'s to larger ones. Eventually, we will get all existing $R[m,n,s]$'s (of different sizes) covered by $D[i,j,s]$'s (of twice their size), with all scaling (contrast) coefficients $|s_{[m,n,s]}| < 1$. This guarantees contractivity of the DFT and best possible encoding (based on Quadtree scheme), since the smaller the $R[m,n,s]$'s the smaller the difference $||x - \hat{x}||_2$, where $\hat{x}$ is the attractor (fixed point) of the induced dynamical system.

The rationale behind using the reverse quadtree partitioning algorithm can be explained by the fact that, in general, the smaller the ranges the

better the fractal transform (smaller collage difference). So, when we go from smaller to larger ranges (as opposed to going from larger to smaller, as the regular quadtree does) we obtain better transforms without worrying about loosing contractivity. (Often, as the size of the ranges decreases the scaling coefficients increase and we could loose contractivity).

A disadvantage of the reverse quadtree partitioning is that it is more computationally intensive and can easily take 3 times longer than the (regular) quadtree scheme. However, we obtain an attractor $\hat{x}$ closer to x than the attractor of the regular quadtree scheme.

To insure that the induced map $W = \bigcup\{w_{[m,n,s]}\}$ is contractive we require each of the maps $w_{[m,n,s]}$ to be contractive, which is equivalent (in the affine mapping case) to $||A|| < 1$, where $||A||$ is the norm of the linear operator A. Then the DFT of the signal f consists of the collection of contractive maps $w_{[m,n,s]}$, (obtained by the minimization procedure described above), that contain the self-similarities and symmetries of the signal f. In more general terms, contractivity of W could be guaranteed by requiring that the collection of maps $\{w_{[m,n,s]}\}$ is "eventually contractive". [Fisher, 1995], in which case we would use the "Generalized Collage Theorem" instead of Theorem 1.5.

Example 1.22. *Let I^2 be a 128×128 square lattice and R be a specific fixed size partition of I^2. Say, $R = \{R[m,n] : size(R[m,n]) = 4 \times 4;\ R[m,n]'s\ are\ disjoint;\ \cup R[m,n] = I^2\}$. Similarly, let $D = \{D[i,j] \subseteq I^2 : size(D[i,j]) = 8 \times 8\}$. Observe that: $|R| = 2^{10}$, $|D| = 122^2$, the $R[m,n]$'s are disjoint, however, the $D[i,j]$'s could (and sometimes do) overlap.*

In this case, the coordinates of the upper left corners of $R[m,n]$ and $D[i,j]$ are $[4m, 4n]$ and $[i,j]$, respectively. Then, we let $D1[i,j] = \{[i + 2k + 1, j + 2l +$

1]}$_{0\leq k,l\leq 3}$ be the sub-sampled version of $D[i,j]$, size($D1[i,j]$) = size($R[m,n]$) = 4×4. Our goal is to minimize:

$$\delta(f_{/R[m,n]}, w_{[m,n]}(f_{/R[m,n]})) = \delta(f_{/R[m,n]}, s[m][n]f(v^{-1}_{/R[m,n]}) + r[m][n]),$$

where $w_{[m,n]}\begin{pmatrix} i \\ j \\ z \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & s_{[m][n]} \end{pmatrix}\begin{pmatrix} i \\ j \\ z \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ r_{[m][n]} \end{pmatrix}$. And $v : D[i,j] \longrightarrow R[m,n]$ is a composition of translations, rotations and flips (rigid motions on a square), $v\begin{pmatrix} i \\ j \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} i \\ j \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, Figure 7. Note that $z = f(i,j)$ and $w_{[m,n]}(f) = w_{[m,n]}(i,j,f)$. In practice, we could choose the maps $w_{[m,n]}$ of the form:

$$w\begin{pmatrix} i \\ j \\ z \end{pmatrix} = \begin{pmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & s_{[m][n]} \end{pmatrix} + \begin{pmatrix} i \\ j \\ z \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ r_{[m][n]} \end{pmatrix} \quad or$$

$$w\begin{pmatrix} i \\ j \\ z \end{pmatrix} = \begin{pmatrix} 0 & a_1 & 0 \\ a_2 & 0 & 0 \\ 0 & 0 & s_{[m][n]} \end{pmatrix} + \begin{pmatrix} i \\ j \\ z \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ r_{[m][n]} \end{pmatrix}$$

and then the contractivity of w would be guaranteed by: $|a_1|, |a_2|, |s_{[m][n]}| < 1$. Note that the condition $size(D[m,n]) > size(R[m,n])$ is equivalent to $|a_1|, |a_2| < 1$. Therefore, if we pick $a_1 = a_2 = \pm 1/2$, $size(D[m,n]) = 2size(R[m,n])$, the only requirement for contractiveness of the affine mappings $w_{[m,n]}$ (inducing the Fractal Transform) is $|s_{[m][n]}| < 1$. If we use more general affine maps, allowing shearing and rotations at arbitrary degrees, we would need to use Claim 1.6 to find the norm of the linear part (A) of W. Then contractiveness will depend on the maximum positive eigenvalue of a certain matrix.

So, for each $R[m,n]$ the FT assigns a 5−tuple (i,j,r,c,b) in the symmetry space S_2, where (i,j) are the coordinates of the upper-left corner of the square $D[i,j]$ that "covers" the $R[m,n]$, that is minimizes the above functional δ. There are eight ways to map a $D[i,j]$ over the $R[m,n]$, these are the 8 rigid motions of a square. So, $r \in \{0,1,2,3,4,5,6,7\}$ carries this "rotation" information for the mappings $w_{[m,n]}$ and $v_{[m,n]}$, look at Figure 7. Lastly, the c and b

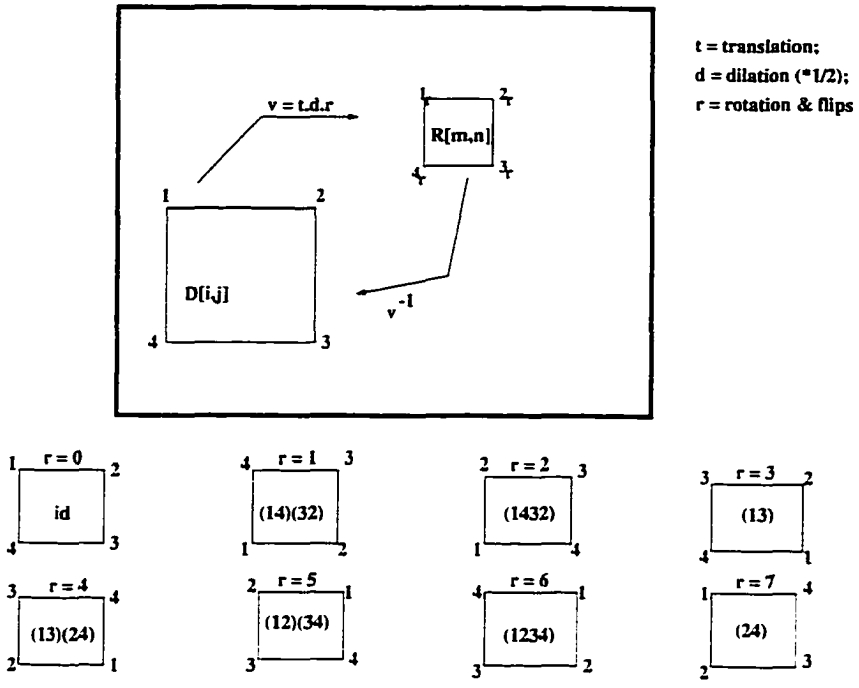


Figure 7. Incorporating 90-degree rotations and flips into the FT.

carry the contrast (scaling, $s[m][n]$) and brightness (offset, $r[m][n]$) coefficients of the map $w_{[m,n]} : D[i,j] \times \Re \longrightarrow R[m,n] \times \Re$. A way to figure out those real coefficients is to use Least-Squares Linear Regression.

$$\delta = \delta(f_{/R[m,n]}, w_{[m,n]}(f_{/R[m,n]})) =$$

$$\sum_{i,j=0}^{3} (f(4m+i, 4n+j) - s[m][n] f(v^{-1}(i,j)) - r[m][n])^2,$$

where $v^{-1} : R[m,n] \longrightarrow D1[i,j]$ $(D[i,j])$ is given by:

$$Case\ r = 0\ (r_o = id): \quad v^{-1}\begin{pmatrix} 4m+k \\ 4n+l \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} k \\ l \end{pmatrix} + \begin{pmatrix} i+1 \\ j+1 \end{pmatrix}$$

$$Case\ r = 1\ (r_1 = \tau): \ v^{-1}\begin{pmatrix} 4m+k \\ 4n+l \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 3-k \\ l \end{pmatrix} + \begin{pmatrix} i+1 \\ j+1 \end{pmatrix} \equiv$$

$$\equiv \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\left[\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} k \\ l \end{pmatrix} + \begin{pmatrix} 3 \\ 0 \end{pmatrix}\right] + \begin{pmatrix} i+1 \\ j+1 \end{pmatrix}$$

$$Case\ r = 2\ (r_2 = \rho): \quad v^{-1}\begin{pmatrix} 4m+k \\ 4n+l \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} l \\ 3-k \end{pmatrix} + \begin{pmatrix} i+1 \\ j+1 \end{pmatrix}$$

$$Case\ r=3\ (r_3=\rho\cdot\tau):\quad v^{-1}\begin{pmatrix}4m+k\\4n+l\end{pmatrix}=\begin{pmatrix}2&0\\0&2\end{pmatrix}\begin{pmatrix}3-l\\3-k\end{pmatrix}+\begin{pmatrix}i+1\\j+1\end{pmatrix}$$

$$Case\ r=4\ (r_4=\rho^2):\quad v^{-1}\begin{pmatrix}4m+k\\4n+l\end{pmatrix}=\begin{pmatrix}2&0\\0&2\end{pmatrix}\begin{pmatrix}3-k\\3-l\end{pmatrix}+\begin{pmatrix}i+1\\j+1\end{pmatrix}$$

$$Case\ r=5\ (r_5=\rho^2\cdot\tau):\quad v^{-1}\begin{pmatrix}4m+k\\4n+l\end{pmatrix}=\begin{pmatrix}2&0\\0&2\end{pmatrix}\begin{pmatrix}k\\3-l\end{pmatrix}+\begin{pmatrix}i+1\\j+1\end{pmatrix}$$

$$Case\ r=6\ (r_6=\rho^3==\rho^{-1}):\quad v^{-1}\begin{pmatrix}4m+k\\4n+l\end{pmatrix}=\begin{pmatrix}2&0\\0&2\end{pmatrix}\begin{pmatrix}3-l\\k\end{pmatrix}+\begin{pmatrix}i+1\\j+1\end{pmatrix}$$

$$Case\ r=7\ (r_7=\rho^3\cdot\tau):\quad v^{-1}\begin{pmatrix}4m+k\\4n+l\end{pmatrix}=\begin{pmatrix}2&0\\0&2\end{pmatrix}\begin{pmatrix}l\\k\end{pmatrix}+\begin{pmatrix}i+1\\j+1\end{pmatrix},$$

where ρ is a 90^o rotation (counterclockwise), $\rho = (1432)$. And τ is a flip with respect to a horizontal line through the middle of the square, $\tau = (14)(32)$. The identity of the dihedral group D_4, acting on the square, is denoted by $r_o = id$.

To minimize δ we set the "partial derivatives" of δ equal to zero and solve for $s[m][n]$ and $r[m][n]$.

$$s[m][n]=$$

$$\frac{16\sum_{i,j=0}^{3} f(4m+i,4n+j)f(v^{-1}(i,j))-\left(\sum_{i,j=0}^{3} f(4m+i,4n+j)\right)\left(\sum_{i,j=0}^{3} f(v^{-1}(i,j))\right)}{16\sum_{i,j=0}^{3}\left(f(v^{-1}(i,j))\right)^2-\left(\sum_{i,j=0}^{3} f(v^{-1}(i,j))\right)^2}$$

$$r[m][n]=\frac{1}{16}\left(\sum_{i,j=0}^{3} f(4m+i,4n+j)-s[m][n]\sum_{i,j=0}^{3} f(v^{-1}(i,j))\right)$$

Remember, that $f(i,j)$ stands for the intensity of the signal at the $(i,j)^{th}$ pixel. For every $R[m,n]$ we have to search through finitely many $D[i,j]$'s, trying to minimize δ. And, for each $D[i,j]$ we have explicit formulas for the scaling and offset coefficients. We should remark, however, that the 5-tuple (i,j,r,c,b) may not be unique. If we insist on having a unique minimizer of δ that we will have to impose some additional constrains on the possible covers. Some authors use fixed scaling coefficients (say $s = 0.9$) which insures uniform contractiveness for all images, however, it also reduces the degree of freedom of the system.

In 3D the group S_4, having 24 elements, gives all possible rigid motions of a cube. One can compute the action of S_4 on the domains and write explicitly the matrix form for the affine contractive maps $v_{[m,n,k]}$.

1.4 A New Classification Scheme for Matching Domains and Ranges

The computationally intensive step of the Fractal Encoding is the domain-range comparison and matching. For each range $R[m,n]$ we search through the whole collection D of domains, trying to cover the $R[m,n]$ "closely" by some $D[i,j]$, and find the mapping $w : D[i,j] \times \Re \longrightarrow R[m,n] \times \Re$. Therefore, one way to speed-up the encoding would be to reduce the number of domain-range comparison steps. This can be done by classifying all domains and ranges, so that at each step an $R[m,n]$ of a certain class is compared only with $D[i,j]$'s of the same or the "near-by" classes.

How can we choose a good classification criterion? Remember, that minimizing the functional δ is up to scaling and offset coefficients (when using affine transformations). If a $D[i,j]$ is a good "cover" of $R[m,n]$ and

$g(w) = s_{[m][n]} f(w)_{/D[i,j]} + r_{[m][n]}$, then $\delta(f_{/R[m,n]}, g(v^{-1})_{/R[m,n]})$ would be small. Since g is an affine functional of f we would like to have a classification criterion that is affinely invariant, or close to being linearly invariant. First thing that comes to mind is using the variance of the signal, which is translation invariant: $Var(f) = Var(f + r)$. Unfortunately, the variance is not dilation invariant: $Var(sf) = s^2 Var(f)$.

In the one dimensional case, a functional that is dilation invariant is the "number of up-crossings" of f over the expectation of f, $\mu = E(f)$. In a more general setting, this is the characteristic of the excursion set of f above the level $\mu = E(f)$. More details on random fields, excursion sets and their characteristics above a given threshold value, and the expectation of those characteristics can be found in Adler, 1981. It would be interesting to see if there is an easy way to obtain a classifying-functional based on the characteristics of the excursion sets and the variance, that is "close" to being linearly invariant. Other methods for classifying domains and ranges can be found in Fisher, 1995.

Now we present a new classification scheme that we used to speed-up the fractal encoding algorithm based on reversed-Quadtree partition. The *skewness* of a signal is defined by:

$$skewness(f) = \frac{E(f-\mu)^3}{(Var(f))^{3/2}},$$

where $Var(f)$ is the variance and $\mu = E(f)$ is the expectation (average) of f.

Claim 1.23. *The functional skewness(•) is linearly invariant.*

Proof: Let $g(w) = sf(w) + r$, then the mean and the variance of g can be expressed as

$$E(g) = sE(f) + r = s\mu + r, \qquad Var(g) = s^2 Var(f)$$

$$skewness(g) = \frac{E(g-E(g))^3}{(Var(g))^{3/2}} = \frac{E(sf+r-s\mu-r)^3}{(s^2Var(f))^{3/2}} = \frac{E(s(f-\mu))^3}{(s^2Var(f))^{3/2}} =$$
$$= s^3 \frac{E(f-\mu)^3}{s^3\,(Var(f))^{3/2}} = skewness(f)$$

Δ

An implementation of the skewness-classification is the following: Let $K+1$ be the number of classes we like to divide the domains and the ranges in. Because of "border-line" cases we always attempt to cover the $R[m,n]$ with a $D[i,j]$ satisfying:

$$Class(R[m,n]) - 1 \leq Class(D[i,j]) \leq Class(R[m,n]) + 1,$$

where the classes are ordered between 0 and K. We define $Class(R[m,n]) = i$ and $Class(D[k,l]) = j$ if and only if

$$\frac{iMax}{K} \leq skewness(f_{/R[m,n]}) < \frac{(i+1)Max}{K},$$
$$\frac{jMax}{K} \leq skewness(f_{/D[k,l]}) < \frac{(j+1)Max}{K},$$

respectively, with $Max = Max\{skewness(f_{/R[m,n]}), skewness(f_{/D[k,l]}) | m,n,k,l\}$.

A caution should be exercised in such classification because it is possible that for some $R[m,n]$ the neighboring classes $\{-1,0,1\}$ could be domain-empty. In these cases we search through the complete library of domains, D, to find the "best" cover of $R[m,n]$.

The *skewness* domain-range classification provides a good way to reduce the computationally intensive block-comparison step of the DFT. If we assume that every range (R) has a "perfect" covering domain (D), i.e. $\exists s, r$ such that $\delta\left(f_{/R}, sf(v_{/R}^{-1}) + r\right) = 0$, then $D = v^{-1}(R)$ and R fall into the same (skewness) class because $f_{/R} = sf(v^{-1}(r)) + o = sf_{/D} + o$ and the skewness is affinely invariant. This assumption is not really restrictive since in the search for covers we march through various (range) sizes and thus have a lot of flexibility in selecting and matching the right self-affine pair.

In practice, the skewness-based classification could reduce the time for computing the DFT by more than 100 times, depending on the number of classes one employs. The main advantage of this preprocessing step is that it does not limit the "essential" space of reasonable covers.

2. The Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) is certainly better known and understood than the DFT because of its continuous and discrete interpretations. We will now review the overall idea behind the DWT. The complete details could be found in Daubechies [1988, 1989], Mallat [1989] and others. The DWT is a method that decomposes signals into superposition of small waves called *wavelets* (similar to the Fourier decomposition of signals into *cosine* waves), enabling us to do analysis and synthesis of the information contained in the data. It is a fast linear operation that maps data vectors of length a power of 2 to a (numerically different) vector of the same length. Also, the DWT is invertible and in fact *orthogonal*, so that if we view it as a matrix, the inverse is simply the transpose of the linear operator. Thus, we can regard the DWT as a rotation in the function space (S_1), from the unitary basis e_i to a *wavelet* basis.

2.1 The Fourier and Cosine Transforms

Let $f(x)$ be a function defined on $[0,1]$ with $||f||_2^2 = \int_0^1 |f(x)|^2 dx$. Then one can write f in terms of an infinite (Fourier) series [Folland, 1984]

$$f(x) = a_o + \sum_{n=1}^{\infty} (a_n \cos(2\pi n x) + b_n \sin(2\pi n x)).$$

Using Euler's representation $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, where i is the imaginary unit

($i^2 = -1$), one writes the Fourier expansion of f using

$$\cos(2\pi nx) = \frac{1}{2}\left(e^{i2\pi nx} + e^{-i2\pi nx}\right) \qquad \sin(2\pi nx) = \frac{1}{2i}\left(e^{i2\pi nx} - e^{-i2\pi nx}\right) \tag{7}$$

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{i2\pi nx},$$

where the (complex) Fourier coefficients $c_n = \int_0^1 f(x)e^{-i2\pi nx}dx$.

Definition 1.24. *A collection of functions* $\{\phi_n| \quad n = 0, \pm 1, \pm 2, \pm 3, \cdots\}$ *in* L^2 *on* $[a, b]$ *is an orthonormal basis of* $L^2([a, b])$ *if*

(1) Orthogonality: $< \phi_k, \phi_l >= \int_a^b \phi_k(x)\overline{\phi_l(x)}dx = 0, \quad \forall k \neq l;$

(2) Normality: $||\phi_k||^2 =< \phi_k, \phi_k >= \int_a^b \phi_k(x)\overline{\phi_k(x)}dx = \int_a^b |\phi_k(x)|^2 dx = 1, \forall k;$

(3) Completeness: For every $g(x) \in L^2([a, b])$, there exist complex coefficients $\{c_n\}$ such that $g(x) = \sum_{n=-\infty}^{\infty} c_n \phi_n(x)$.

Observe that if $\{\phi_n\}$ is an orthonormal basis of $L^2([a, b])$, then the (Fourier) coefficients of the Fourier expansion series of $g(x) \in L^2([a, b])$, with respect to the basis $\{\phi_n\}$, are given by

$$c_n = \sum_{k=-\infty}^{\infty} c_k \int_a^b \phi_k(x)\overline{\phi_n(x)}dx = \int_a^b g(x)\overline{\phi_n(x)}dx.$$

One can also verify easily that the exponential family

$$\{e^{i2\pi nx}| \quad n = 0, \pm 1, \pm 2, \pm 3, \cdots\}$$

indeed forms an orthonormal basis of $L^2([0, 1])$.

There are two main limitations of the Fourier transform; It is well localized in frequency domain, but not in time (space) domain; And there are "edge artifacts" at the end points of the domain. These are caused by slow convergence of the Fourier series at the boundary points.

The good localization in frequency and the poor localization in time domain of the Fourier transform are features inherited by the nature of the

basic functions ($\{\sin, \cos\}$) used in the image decomposition. The problematic end point convergence, however, is a property that can be improved on.

Suppose $f(x) \in L^2([0,1])$ has a Fourier series expansion $f(x) = \sum_{n=-\infty}^{\infty} c_n e^{i2\pi nx}$ on the interval $[0,1]$. If we extend the definition of $f(x)$ to $[0,2]$ by letting $f(2-x) = f(x)$ the new extension function, $\tilde{f}(x)$, would be an $L^2([0,2])$ function whose graph represents two simple mirror image reflections of the graph of $f(x)$ with respect to the vertical line $x = 1$, Figure 8.

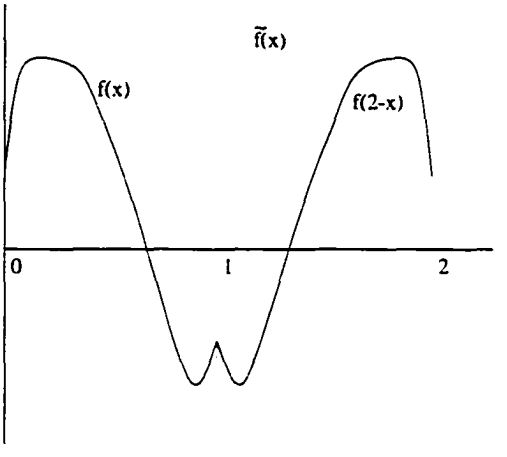


Figure 8. Mirror image extension of a signal.

The Fourier series expansion of $\tilde{f}(x)$ has some nice properties;

$$\tilde{f}(x) = \sum_{n=-\infty}^{\infty} a_n e^{i2\pi nx}, \quad 0 \leq x \leq 2$$

Using equations (7) and the fact that $f(2-x) = f(x)$ we can simplify $\tilde{f}(x)$ to

$$\tilde{f}(x) = \sum_{n=0}^{\infty} b_n \cos(2\pi nx)$$

This representation of $\tilde{f}(x)$, restricted to $[0,1]$, is called the Cosine transform of $f(x)$. Note that because of the fact that the right end of the domain of $f(x)$, the point 1, turns out in the middle of the domain of the extension $\tilde{f}(x)$, the Fourier representation of $\tilde{f}(x)$ at 1 is exact (and $\tilde{f}(1) = f(1)$). Thus we have eliminated the edge point (boundary) artifacts

of the representation of $f(x)$ at 1. Similarly, one can expand $f(x)$ on $[-1, 1]$ and avoid the slow convergence of the Fourier series at the left end of the domain, 0.

Thus, we have a way (Cosine transform) to solve the problem associated with the edge effects of the Fourier representation at the end points of the domains of the signals. Can we find a way around to construct a function representation that is as well localized in times as it is in frequency domain? This, of course, would ease dramatically the function interpretation and would yield very fast convergence (of the representation series). Recall that if $\{\phi_n\}$ is an orthonormal basis for $L^2([0,1])$ then every $g(x) \in L^2([0,1])$ has an infinite series representation $g(x) = \sum_k < \phi_k, g > \phi_k(x)$, where the coefficients $< \phi_k, g >= \int_0^1 \phi_k(x) g(x) dx$. The problem is that for some choices of orthonormal bases (including the $\{\sin, \cos\}$ basis) local perturbations of the function of interest, g, would affect all of the representation coefficients $< \phi_k, g >$. One can fix that by using "windowed" transforms, like the Windowed Fourier Transform (WFT). If $g(x) \in L^2(R)$ the windowed Fourier transform of $g(x)$ is defined by

$$\widehat{g}(w, s) = \int_R g(u) \overline{W(u-s)} e^{-i2\pi w u} du,$$

where the variable s (shifting factor) determines the position of the (non-trivial) window function $W(u) \in L^2(R)$.

Then we can recover g from its WFT by

$$g(x) = \frac{1}{||W||^2} \int_R \int_R \widehat{g}(w, s) W(u-s) e^{i2\pi w u} dw ds.$$

Now small changes of $g(x)$ may only affect the Fourier coefficients $\widehat{g}(w, s)$ with s localized within the domain of change of x.

This, however, introduces other problems. For instance, windowing the basis functions is likely to yield discontinuities at the edges of the windowed

basis. In turn this may cause a significant amount of the representation coefficients $\widehat{g}(w,s)$ to be large, reflecting the "high-frequency" components artificially introduced to the representation of $g(x)$ through windowing the basis of the space. If the Fourier type transforms can not be easily fixed to avoid the above mentioned problems, can we find other bases of $L^2(R)$ that are both localized in the time and the frequency domain that provide efficient, compact and robust function representations?

2.2 The Discrete Wavelet Transform

2.2.1 Orthogonal decomposition of spaces. We now describe the theory of Multi-Resolution Analysis (MRA) in the case of the Haar wavelet basis. Let $\phi(x)$ be a step function of magnitude one and step-size one

$$\phi(x) = \chi_{[0,1)}(x) = \begin{cases} 0 & x \notin [0,1) \\ 1 & x \in [0,1) \end{cases}$$

Denote $V_o = \left\{ f \in L^2(R) | \quad f(x) = \sum_{n=-\infty}^{\infty} c_n \phi(x-n) \right\}$, V_o contains all step functions on R^1 of step-size $= 1$. Also, let $V_1 = \left\{ f \in L^2(R) | \quad f(x) = \sum_{n=-\infty}^{\infty} d_n \phi(2x-n) \right\}$, these are all square-integrable step functions of step-size $= \frac{1}{2}$. One can see that $V_o \subseteq V_1$. We want a representation of every function $f(x) \in V_1$ as $f(x) = f_o(x) + g_1(x)$, where $f_o(x) \in V_o$ and $g_1(x) \in (V_o)^{\perp}$, $(V_o)^{\perp}$ is the perpendicular space of V_o, containing all functions in V_1 that are orthogonal to all functions in V_o. That is $(V_o)^{\perp} = \left\{ f_1(x) \in V_1 | \quad < f_1, f_o >= \int_R f_1(x)\overline{f_o(x)}dx = 0 \quad \forall f_o \in V_o \right\}$. For simplicity of notation let $W_o = (V_o)^{\perp}$. Can we explicitly identify the subspace W_o? For the Haar scaling function $\phi(x)$ it turns out [Daubechies, 1998] that $W_o = \left\{ f_1 \in V_1 | \quad f_1(x) = \sum_{n=-\infty}^{\infty} c_n \psi(x-k) \right\}$, where

$$\psi(x) = \begin{cases} 1 & 0 \leq x < 1/2 \\ -1 & 1/2 \leq x < 1 \\ 0 & otherwise \end{cases}$$

Clearly, $\psi(x) \in V_1$, every function of V_o is orthogonal to every function in W_o and $\frac{1}{2}(\psi(x)+\phi(x)) = \phi(2x)$. This last equality yields a disjoint decomposition of the space $V_1 = V_o \oplus W_o$.

In general, if $V_k = \left\{ f \in L^2(R) | \quad f(x) = \sum_{n=-\infty}^{\infty} d_n \phi(2^j x - n) \right\}$, then $V_k = V_{k-1} \oplus W_{k-1}$ for $W_k = \left\{ f \in L^2(R) | \quad f(x) = \sum_{n=-\infty}^{\infty} d_n \psi(2^k x - n) \right\}$. Therefore, for all $k \geq 1$

$$V_k = V_{k-1} \oplus W_{k-1} = V_{k-2} \oplus W_{k-2} \oplus W_{k-1} = \cdots =$$

$$= V_o \oplus W_o \oplus W_1 \oplus W_2 \oplus \cdots \oplus W_{k-1}$$

In a similar fashion one can extend this orthogonal space decomposition using step functions of "increasing" step-size. Let

$$V_{-1} = \left\{ f \in L^2(R) | \quad f(x) = \sum_{n=-\infty}^{\infty} d_n \phi(2^{-1} x - n) \right\}$$

be the $L^2(R)$ span of all step functions of step-size $= 2$. And, in general,

$$V_{-l} = \left\{ f \in L^2(R) | \quad f(x) = \sum_{n=-\infty}^{\infty} d_n \phi(2^{-l} x - n) \right\}$$

be the collection of all functions in $L^2(R)$ that can be written as linear combinations of the (shifted) step functions of step-size $= 2^l$. Again trivial calculations show that $V_o = V_{-1} \oplus W_{-1}$ and $V_{-l} = V_{-l-1} \oplus W_{-l-1}$, for $l > 0$, where

$$W_{-k} = \left\{ f \in L^2(R) | \quad f(x) = \sum_{n=-\infty}^{\infty} d_n \psi(2^{-k} x - n) \right\}.$$

Because, the family of all step functions (of different step-sizes) is "dense" in $L^2(R)$ (i.e. its span is $L^2(R)$), Folland 1984, we obtain the following orthogonal decomposition of the image space $L^2(R)$

$$L^2(R) = \bigcup_{n=0}^{\infty} V_k = V_o \oplus W_o \oplus W_1 \oplus W_2 \oplus W_3 \oplus \cdots$$

Here we implicitly used the fact that $V_k = V_{k-1} \oplus W_{k-1}$ is an orthogonal decomposition of the intermediate space V_k, for any positive integer k. In

addition, the starting space V_o can be expressed as an infinite sum of mutually orthogonal spaces

$$V_o = V_{-1} \oplus W_{-1} = (V_{-2} \oplus W_{-2}) \oplus W_{-1} = \cdots = \cdots \oplus W_{-4} \oplus W_{-3} \oplus W_{-2} \oplus W_{-1}.$$

Finally, $L^2(R) = \cdots \oplus W_{-3} \oplus W_{-2} \oplus W_{-1} \oplus W_0 \oplus W_1 \oplus W_2 \oplus W_3 \oplus \cdots$, where essentially W_k is the $L^2(R)$ span of $\{\psi(2^k x - s) | \quad s \in Z\}$, for all integer (step-sizes) k. The induced basis of $L^2(R)$, $\{\psi(2^k + s) | \quad k, s \in Z\} = \{2^{\frac{k}{2}} \psi(2^k + s) | \quad k, s \in Z\}$ is called the Haar wavelet basis and the function $\psi(x) = \phi(2x) - \phi(2x - 1)$ is called the Haar wavelet.

We observe the following three important properties of the Haar decomposition of $L^2(R)$

(1) V_o is the $L^2(R)$ span of $\{\phi(x - s) | \quad s = 0, \pm 1, \pm 2, \pm 3, \cdots\}$,

$$V_k = \left\{ f \in L^2(R) | \quad f(x) = \sum_{s=-\infty}^{\infty} d_n \phi(2^k x - s) \right\}$$

and $\{2^{\frac{k}{2}} \phi(2^k x - s) | \quad s \in Z\}$ is an orthonormal basis of V_k;

(2) $\bigcap_{n=-\infty}^{\infty} V_n = \{0\}$;

(3) If $\overline{A}$ denotes the topological closure of the set A, $\overline{\bigcup_{n=-\infty}^{\infty} V_n} = L^2(R)$.

Definition 1.25. *If there exists a function $\phi(x)$ such that the above three properties for the spaces V_k are satisfied, then the collection of spaces $\{V_k\}$ is called a Multi-Resolution Analysis (MRA), and the function $\phi(x)$ is termed the scaling function of the MRA.*

Why are we interested in MRA "framings" of $L^2(R)$ and how do we construct wavelets and wavelet representations of signals using MRA's? All scaling functions giving rise to MRA's have to satisfy certain properties. For example, because $\phi(x) \in V_o \subseteq V_1 = span\{2^{\frac{1}{2}} \phi(2x - s) | \quad s \in Z\}$ we obtain what is known as the dilation equation

$$\phi(x) = \sum_{s=-\infty}^{\infty} d_s \phi(2x - s)$$

Furthermore, the coefficients d_s need to satisfy $\sum_{s=-\infty}^{\infty} d_s = 2$. Since if we denote the expectation of the scaling function $\phi(x)$ by $\mu = \int_R \phi(x)dx$ ($\mu \neq 0$, Daubechies, 1988), then integrating both hand sides of the dilation equation we obtain the desired normalization relationship between the coefficients d_s

$$\mu = \sum_{s=-\infty}^{\infty} d_s \frac{\mu}{2}.$$

The function family $\{\phi(x-s)| \quad s \in Z\}$ forms an orthonormal basis of V_o. Therefore, if $\phi_{s_1} = \phi(x-s_1)$ and $\phi_{s_2} = \phi(x-s_2)$, then

$$< \phi_{s_1}, \phi_{s_2} >= \int_R \phi(x-s_1)\phi(x-s_2)dx = \begin{cases} 1, & s_1 = s_2 \\ 0, & s_1 \neq s_2 \end{cases}$$

Note that

$$\phi(x-s_1) = \sum_{s=-\infty}^{\infty} d_s \phi(2(x-s_1)-s) = \sum_{s=-\infty}^{\infty} d_s \phi(2x-(2s_1+s)) = \sum_{p=-\infty}^{\infty} d_{p-2s_1}\phi(2x-p)$$

Similarly, $\phi(x-s_2) = \sum_{q=-\infty}^{\infty} d_{q-2s_2}\phi(2x-q)$.

Putting these two facts together we obtain (after a change of variables)

$$\left.\begin{matrix} for\ s_1 = s_2, & 1 \\ for\ s_1 \neq s_2, & 0 \end{matrix}\right\} =< \phi_{s_1}, \phi_{s_2} >= \int_R \sum_{p,q} d_{p-2s_1} d_{q-2s_2} \phi(2x-p)\phi(2x-q)dx =$$

$$= \sum_{p,q} d_{p-2s_1} d_{q-2s_2} \int_R \phi(2x-p)\phi(2x-q)dx = \frac{1}{2}\sum_{p} d_{p-2s_1} d_{p-2s_2} = \frac{1}{2}\sum_{k} d_k d_{k+2m}$$

where $k = p = 2s_1 + s$, $k = q = 2s_2 + s$ and $2m = 2s_1 - 2s_2$.

The last equality, called the orthogonal coefficients condition, together with the dilation coefficient condition are "usually" employed to find a scaling function $\phi(x)$. For computational purposes it is helpful to search for scaling functions supported on a compact set. This is guaranteed whenever only finitely many of the dilation equation coefficients (d_s) are non-zero.

We used the Haar wavelet basis to introduce the theory of MRA, however, the Haar scaling function (and the Haar wavelet) are not smooth at

the edges. Thus we have not completely solved the problem of the end point artifacts in the windowed Fourier transform. So, we begin looking for continuous (not necessarily differentiable) scaling functions. To find such functions Daubechies [1988] proposed the following procedure; For simplicity we take $N = 4$, the number of non-zero dilation equation coefficients. Using the orthogonal coefficients and the dilation equation conditions we have

$$d_o^2 + d_1^2 + d_2^2 + d_3^2 = 2, \quad d_o d_2 + d_1 d_3 = 0, \quad d_o + d_1 + d_2 + d_3 = 2$$

The extra degree of freedom of the system is used to make $m_o(z) = d_o + d_1 z + d_2 z^2 + d_3 z^3 = 0$, where $z = e^{i\theta}$, factor as $m_o(z) = (1+z)^2 m_2(z)$, with $m_2(z)$ an affine trigonometric polynomial in z. One solution to these 4 equations is

$$d_o = \frac{1+\sqrt{3}}{4}, \quad d_1 = \frac{3+\sqrt{3}}{4}, \quad d_2 = \frac{3-\sqrt{3}}{4}, \quad d_3 = \frac{1-\sqrt{3}}{4}.$$

The first continuous scaling function (with expectation $\mu = 1$) was obtained using these 4 coefficients and the recursive relation we called "dilation equation"

$$\phi(x) = \sum_{k=0}^{3} d_k \phi(2x - k).$$

In addition, Daubechies [1988, p. 951] showed that if $\nu_o = \chi_{[0,1]}(x)$ and $\nu_l(x)$ is defined recursively by $\nu_l(x) = \sum_{k=-\infty}^{\infty} d_k \nu_{l-1}(2x-k)$, then $\nu_l(x)$ converges pointwise, as $l \longrightarrow \infty$, to a continuous function we call $\phi(x)$.

Once we have the scaling function $\phi(x)$ we can determine the induced continuous (Daubechies, DAUB4) wavelet by

$$\psi(x) = \sum_{s=-1}^{2} (-1)^s d_{s+1} \phi(2x+s)$$

In general, Daubechies shows the constrain that $m_o(z) = \sum_{n=0}^{N} d_n z^n$ is divisible by $(1+z)^K$, where $z = e^{i\theta}$, imposes regularity conditions on the limit function $\phi(x) = \lim_{l \to \infty} \nu_l(x)$. The bigger the exponent K the smoother the induced wavelet and the larger its support.

Daubechies also point out another, Fourier, approach of obtaining the wavelet filter coefficients. Transforming the orthogonal coefficient condition, $\sum_k d_k d_{k+2m} = 2\delta_{om} = \begin{cases} 2, & m = 0 \\ 0, & m \neq 0 \end{cases}$, (in terms of $\{d_k\}$) into a condition on the function $m_o(z)$, $|m_o(z)|^2 + |m_o(z+\pi)|^2 = 1$, and replacing the normalization condition, $\sum_k d_k = 2$, by $m_o(0) = 1$, Daubechies proves that a trigonometric polynomial satisfying these two conditions and such that $m_o(z) \neq 0$, for $|z| < \epsilon$ (for some $\epsilon > 0$), induces an orthonormal set of functions $\{\phi(x-k)\}_k$, where $\widehat{\phi}(w) = FT(\phi)(w) = \prod_{l=1}^{\infty} m_o(\frac{w}{2^l})$. As for the DAUB4 example, having the function ϕ we can obtain the corresponding wavelet function ψ by

$$\psi(x) = \sum_{s=-\infty}^{\infty} (-1)^s d_{s+1} \phi(2x+s).$$

Indeed, $\{2^{\frac{k}{2}}\psi(2^k x - s) | \quad k, s \in Z\}$ is an orthonormal basis for $L^2(R)$, Daubechies [1988].

Figure 9 shows some of the induced scaled wavelets (DAUB20 filter) and the corresponding wavelet decomposition of a function. It is a general rule that the more (non-zero) coefficients are used in the iterative definition of $\phi(x)$ the smoother the scaling function and the corresponding wavelet. The drawback of working with very smooth wavelets is that their support increases and we loose the time localization properties we were pursuing.

2.2.2 Discrete Pyramidal Algorithm for the DWT. The wavelet constructions and the induced orthogonal spatial decomposition we presented in the previous section are very suitable for digital signal analysis. If we work with a compactly supported wavelet constructed from an MRA and the discretized signals are of length 2^N the DWT is a linear functional on R^{2^N}.

We describe the analysis and synthesis of functions using DWT based on the Haar wavelet, however the same procedures are involved under any

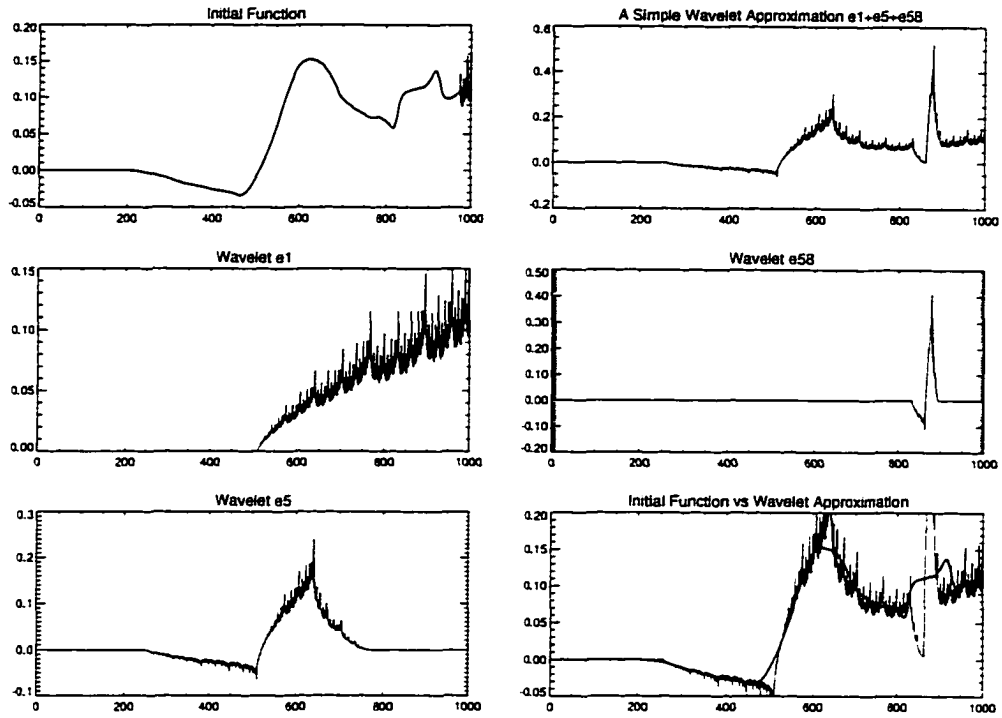


Figure 9. Wavelet representation of signals.

other wavelet basis. Let $F(x)$ and $G(x)$ be two signals defined on $[0,1]$. Let $f(k) = F(\frac{k}{2^N})$ and $g(k) = G(\frac{k}{2^N})$ be their discretized (sampled) versions, for $k = 0, 1, 2, \cdots, 2^N - 1$. As before, let $\phi(x) = \chi_{[0,1)}(x)$, $V_k = span\{2^{k/2}\phi(2^k x - s) | \quad s \in Z\}$, and $\psi(x) = \phi(2x) - \phi(2x-1)$ (there are only two non-zero coefficients in the dilation equation).

Since we discretized the functions on a 2^N-node grid of $[0,1]$

$$f(x) = \sum_{s=0}^{2^N-1} a_{N,s} 2^{N/2} \phi(2^N x - s) \quad \in \quad V_N$$

$$g(x) = \sum_{s=0}^{2^N-1} c_{N,s} 2^{N/2} \phi(2^N x - s) \quad \in \quad V_N \tag{8}$$

Using the orthogonal decomposition of V_N we have that $V_N = V_o \oplus W_o \oplus W_1 \oplus W_2 \oplus \cdots \oplus W_{N-1}$, with $W_k = span\{2^{k/2}\psi(2^k x - s) | \quad s \in Z\}$. Therefore, the collection $\{\phi(x-s), 2^{k/2}\psi(2^k x - s) | \quad k = 0, 1, 2, \cdots, N-1, s \in Z\}$ forms an orthonormal basis for V_N and we can write f and g as sums of orthogonal components

$$f(x) = \sum_{s \in Z} a_{o,s}\phi(x-s) + \sum_{k=0}^{N-1}\sum_{s \in Z} b_{k,s}2^{k/2}\psi(2^k x - s)$$

$$g(x) = \sum_{s \in Z} c_{o,s}\phi(x-s) + \sum_{k=0}^{N-1}\sum_{s \in Z} d_{k,s}2^{k/2}\psi(2^k x - s) \tag{9}$$

The main question now is how to find the wavelet representation coefficients $\{a_{o,s}, b_{k,s}, c_{o,s}, d_{k,s}\}$? We use the properties of orthogonal decomposition. Multiply both hand sides of equations (9) by $\psi(2^k x - s)$ or $\phi(x-s)$ and integrate over R, for all $\{k = 0, 1, 2, \cdots, N-1\}$ and $\{s \in Z\}$. After simplifying we get the $DWT(f) = \widehat{f}$ and $DWT(g) = \widehat{g}$

$$b_{k,s} = \int_R f(x)2^{k/2}\psi(2^k x - s)dx, \qquad d_{k,s} = \int_R g(x)2^{k/2}\psi(2^k x - s)dx$$

$$a_{k,s} = \int_R f(x)2^{k/2}\phi(2^k x - s)dx, \qquad c_{k,s} = \int_R g(x)2^{k/2}\phi(2^k x - s)dx \tag{10}$$

A fundamental property of the DWT is that it is a linear operation on the image space and can be computed recursively using the dilation and the wavelet equations, $\phi(x) = \phi(2x) + \phi(2x-1)$, $\psi(x) = \phi(2x) - \phi(2x-1)$ (for the Haar wavelets). For our example we have $2^{k/2}\phi(2^k x - s) = 2^{k/2}[\phi(2^k x - 2s) + \phi(2^k x - 2s - 1)]$ and $2^{k/2}\psi(2^k x - s) = 2^{k/2}[\phi(2^k x - 2s) - \phi(2^k x - 2s - 1)]$.

Using these facts we can determine recursive relations between the wavelet coefficients

$$a_{k-1,s} = \int_R f(x)2^{(k-1)/2}\phi(2^{k-1}x - s)dx =$$

$$= 2^{-1/2}\int_R f(x)2^{k/2}[\phi(2^k x - 2s) + \phi(2^k x - 2s - 1)]dx = \frac{1}{\sqrt{2}}[a_{k,2s} + a_{k,2s+1}]$$

$$b_{k-1,s} = \int_R f(x)2^{(k-1)/2}\psi(2^{k-1}x - s)dx =$$

$$= 2^{-1/2}\int_R f(x)2^{k/2}[\phi(2^k x - 2s) - \phi(2^k x - 2s - 1)]dx = \frac{1}{\sqrt{2}}[a_{k,2s} - a_{k,2s+1}] \tag{11}$$

And similarly, $c_{k-1,s} = \frac{1}{\sqrt{2}}[c_{k,2s} + c_{k,2s+1}]$ and $d_{k-1,s} = \frac{1}{\sqrt{2}}[c_{k,2s} - c_{k,2s+1}]$.

Another key observation is that if the $\widehat{f} = DWT(f)$ is known (the coefficients $a_{k,s}$ and $b_{k,s}$ are given) then we can quickly recover (synthesize) the coefficients $\{a_{N,s}\}$, that is we can reconstruct f. To do that we just invert equations (11) to get

$$a_{k,2s} = \frac{1}{\sqrt{2}}[a_{k-1,s} + b_{k-1,s}]$$
$$a_{k,2s+1} = \frac{1}{\sqrt{2}}[a_{k-1,s} - b_{k-1,s}] \tag{12}$$

Example 1.26. *The goal of this example is to illustrate the computations of the DWT and its inverse, IWT, for discretized signals F and G using the Haar wavelet basis.*

In addition, we will introduce a measure of image similarities based on comparing the "compressed" DWT's of f and g. This approach of estimating image variations will play an important role later on when we describe our "transform-based" model for image analysis. Suppose $N = 3$, F and G are two given functions on $[0, 1]$, f and g are their sampled versions on the lattice $\{0, 1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8\}$, $\widehat{f}$ and $\widehat{g}$ are the DWT's of f and g, respectively, $\widehat{f_c}$ and $\widehat{g_c}$ are the compressed DWT's and $f_c = IWT(\widehat{f_c})$ and $g_c = IWT(\widehat{g_c})$ are the IWT's. Figure 10.

$$\{f(k/8)| \quad k = 0, 1, \cdots, 7\} \quad = \quad \{1, 2, 3, 1, 5, 8, 8, 7\}$$

$$\{g(k/8)| \quad k = 0, 1, \cdots, 7\} \quad = \quad \{3, 2, 1, 1, 0, -1, 0, 1\}$$

Using equations (10) and (11) we iteratively determine

$$\{a_{3,s}| \quad s = 0, 1, \cdots, 7\} \quad = \left(\frac{1}{\sqrt{2}}\right)^3 \{1, 2, 3, 1, 5, 8, 8, 7\}$$

$$\{c_{3,s}| \quad s = 0, 1, \cdots, 7\} \quad = \left(\frac{1}{\sqrt{2}}\right)^3 \{3, 2, 1, 1, 0, -1, 0, 1\}$$

$$\{a_{2,s}| \quad s = 0, 1, 2, 3\} \quad = \left(\frac{1}{\sqrt{2}}\right)^4 \{3, 4, 13, 15\}$$

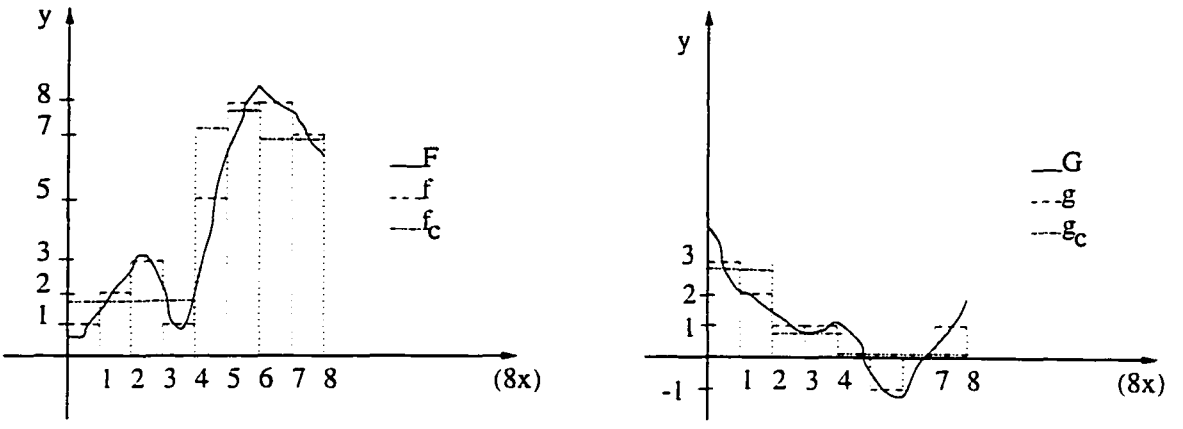


Figure 10. The Original signals (F, G), their discretization (f, g) and the IWT's of the compressed wavelet transforms (f_c, g_c).

$$\{c_{2,s} | \quad s = 0, 1, 2, 3\} \quad = \left(\frac{1}{\sqrt{2}}\right)^4 \{6, 2, -1, 1\}$$

$$\{b_{2,s} | \quad s = 0, 1, 2, 3\} \quad = \left(\frac{1}{\sqrt{2}}\right)^4 \{-1, 2, -3, 1\}$$

$$\{d_{2,s} | \quad s = 0, 1, 2, 3\} \quad = \left(\frac{1}{\sqrt{2}}\right)^4 \{1, 0, 1, -1\}$$

$$\{a_{1,s} | \quad s = 0, 1\} \quad = \left(\frac{1}{\sqrt{2}}\right)^5 \{7, 28\}$$

$$\{c_{1,s} | \quad s = 0, 1\} \quad = \left(\frac{1}{\sqrt{2}}\right)^5 \{8, 0\}$$

$$\{b_{1,s} | \quad s = 0, 1\} \quad = \left(\frac{1}{\sqrt{2}}\right)^5 \{-1, -2\}$$

$$\{d_{1,s} | \quad s = 0, 1\} \quad = \left(\frac{1}{\sqrt{2}}\right)^5 \{4, -2\}$$

$$\{a_{0,s} | \quad s = 0\} \quad = \left(\frac{1}{\sqrt{2}}\right)^6 \{35\}$$

$$\{c_{0,s} | \quad s = 0\} \quad = \left(\frac{1}{\sqrt{2}}\right)^6 \{8\}$$

$$\{b_{0,s} | \quad s = 0\} \quad = \left(\frac{1}{\sqrt{2}}\right)^6 \{-21\}$$

$$\{d_{0,s} | \quad s = 0\} \quad = \left(\frac{1}{\sqrt{2}}\right)^6 \{8\}$$

Thus, the $DWT(f) = \widehat{f} =$

$$= \{a_{o,o}, b_{o,o}, b_{1,0}, b_{1,1}, b_{2,0}, b_{2,1}, b_{2,2}, b_{2,3}\} =$$

$$= \{35/8, -21/8, -\sqrt{2}/8, -\sqrt{2}/4, -1/4, 1/2, -3/4, 1/4\}$$

And the $DWT(g) = \widehat{g} =$

$$= \{c_{o,o}, d_{o,o}, d_{1,0}, d_{1,1}, d_{2,0}, d_{2,1}, d_{2,2}, d_{2,3}\} = \{1, 1, -\sqrt{2}/2, -\sqrt{2}/4, 1/4, 0, 1/4, -1/4\}$$

We now compress the signals DWT's by setting to zero all coefficients of magnitudes less than or equal to $\frac{1}{2}$. Why $\frac{1}{2}$? Because less then 50% of the wavelet coefficients of the two signals are greater than $\frac{1}{2}$. From the function representations (8) and (9) we see that only the large wavelet coefficients capture the essence of the data content of the images. The places of jumps and high gradients of the functions (high frequencies) are encoded by the large representation coefficients. In practice, we use only the top 5% of the wavelet spectral coefficients. The compressed wavelet transforms of the discretized initial functions are $\widehat{f}_c = \{35/8, -21/8, 0, 0, 0, 0, -3/4, 0\}$ and $\widehat{g}_c = \{1, 1, -\frac{\sqrt{2}}{2}, 0, 0, 0, 0, 0\}$.

To recover the images using the "compressed" DWT's we use formulas (12)

$$\{\widehat{a_{3,s}} | \quad s = 0, 1, \cdots, 7\} \qquad = \left(\frac{1}{\sqrt{2}}\right)^3 \{14/8, 14/8, 14/8, 14/8, 50/8, 62/8, 56/8, 56/8\}$$

$$\{\widehat{c_{3,s}} | \quad s = 0, 1, \cdots, 7\} \qquad = \left(\frac{1}{\sqrt{2}}\right)^3 \{3, 3, 1, 1, 0, 0, 0, 0\}$$

$$\{\widehat{a_{2,s}} | \quad s = 0, 1, 2, 3\} \qquad = \left(\frac{1}{\sqrt{2}}\right)^2 \{14/8, 14/8, 56/8, 56/8\}$$

$$\{\widehat{c_{2,s}} | \quad s = 0, 1, 2, 3\} \qquad = \left(\frac{1}{\sqrt{2}}\right)^2 \{3, 1, 0, 0\}$$

$$\{\widehat{b_{2,s}} | \quad s = 0, 1, 2, 3\} \qquad = \left(\frac{1}{\sqrt{2}}\right)^2 \{0, 0, -3/4, 0\}$$

$$\{\widehat{d_{2,s}}| \quad s=0,1,2,3\} \qquad = \left(\frac{1}{\sqrt{2}}\right)^2 \{0,0,0,0\}$$

$$\{\widehat{a_{1,s}}| \quad s=0,1\} \qquad = \left(\frac{1}{\sqrt{2}}\right) \{14/8, 56/8\}$$

$$\{\widehat{c_{1,s}}| \quad s=0,1\} \qquad = \left(\frac{1}{\sqrt{2}}\right) \{2,0\}$$

$$\{\widehat{b_{1,s}}| \quad s=0,1\} \qquad = \left(\frac{1}{\sqrt{2}}\right) \{0,0\}$$

$$\{\widehat{d_{1,s}}| \quad s=0,1\} \qquad = \left(\frac{1}{\sqrt{2}}\right) \{1,0\}$$

$$\{\widehat{a_{0,s}}| \quad s=0\} \qquad = \left(\frac{1}{\sqrt{2}}\right) \{35/8\}$$

$$\{\widehat{c_{0,s}}| \quad s=0\} \qquad = \left(\frac{1}{\sqrt{2}}\right) \{1\}$$

$$\{\widehat{b_{0,s}}| \quad s=0\} \qquad = \left(\frac{1}{\sqrt{2}}\right) \{-21/8\}$$

$$\{\widehat{d_{0,s}}| \quad s=0\} \qquad = \left(\frac{1}{\sqrt{2}}\right) \{1\}$$

Finally, the recovered signals (IWT) are

$$\{f_c(k/8)| \quad k=0,1,\cdots,7\} \quad = \quad \{14/8, 14/8, 14/8, 14/8, 50/8, 62/8, 56/8, 56/8\}$$

$$\{g_c(k/8)| \quad k=0,1,\cdots,7\} \quad = \quad \{3,3,1,1,0,0,0,0\},$$

see Figure 10.

Observe that $||f-g||_{l_2} = 15$, $||\widehat{f_c} - \widehat{g_c}||_{l_2} = 5$. This will be discussed later when we present our "transform-based" model for image analysis.

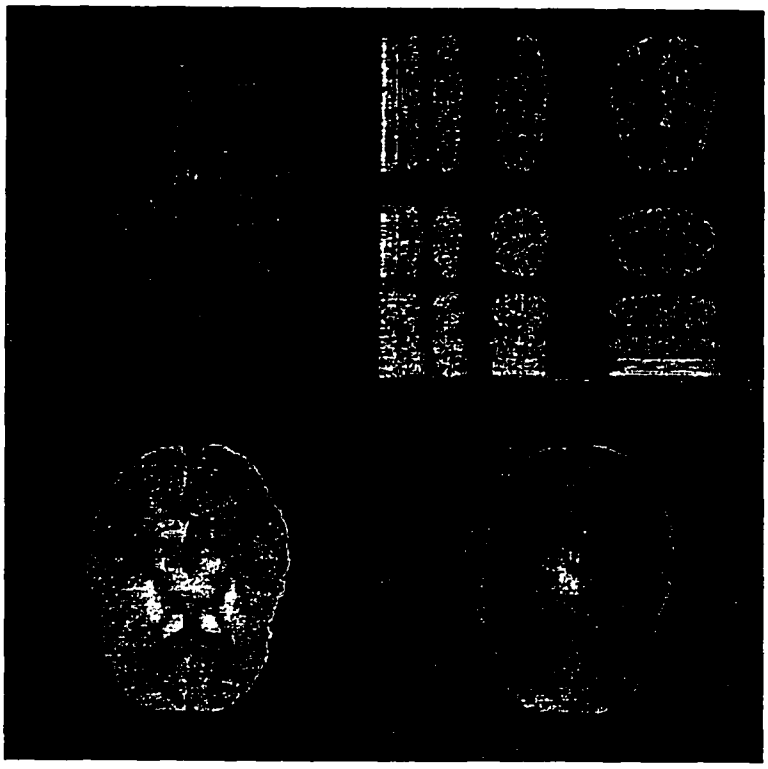
2.3 Practical Implementations

A particular set of wavelets is specified by a finite collection of numbers, called *wavelet filter coefficients*. For example, the Daubechies basis DAUB4,

[Daubechies, 1988], is determined by only four coefficients c_0, c_1, c_2, c_3 representing a solution of a system of dilation equations. Then the DWT would be induced by the orthogonal matrix

$$C = \begin{bmatrix} c_0 & c_1 & c_2 & c_3 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ c_3 & -c_2 & c_1 & -c_0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & c_0 & c_1 & c_2 & c_3 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & c_3 & -c_2 & c_1 & -c_0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & & & & & & \ddots & -c_2 & c_1 & -c_0 \\ c_2 & c_3 & 0 & \cdots & & & & & 0 & c_0 & c_1 \\ c_1 & -c_0 & 0 & \cdots & & & & & 0 & c_3 & -c_2 \end{bmatrix}$$

We think of the filter $\{c_0, c_1, c_2, c_3\}$ as being a *smoothing* filter, averaging four consecutive components of the data. Because of the minus signs in the second filter $\{c_3, -c_2, c_1, -c_0\}$ we view it as a *non-smoothing* filter. The DWT consists of iteratively applying the wavelet coefficient matrix (C) to the entire data $\mathbf{x_0}$ of length 2^N, then to the smooth vector $\mathbf{x_1}$ of length 2^{N-1} (consisting of the *odd* components of $\mathbf{Cx_0}$), then applying C to the smooth vector $\mathbf{x_2}$ of length 2^{N-2} (containing the *odd* components of $\mathbf{Cx_1}$) etc. This procedure is called the pyramidal algorithm for finding the DWT. At the end of the process, the collection of the "non-smooth" components (the residuals of applying the even rows of C, at each step) will contain the numerically different vector of length 2^N, representing the DWT of $\mathbf{x_0}$. To recover the original signal from its DWT, we reverse the process, applying C^T $(= C^{-1})$ to the DWT we have already computed. Figure 11 shows an example of a 2D MRI scan, its DWT, the IWT (Inverse Wavelet Transform), and the IFT (Inverse Fractal Transform). The DFT of the image is not included since it is not easy to visualize (recall that DFT is just a collection of contractive maps). This figure is not intended to be used as a comparison of the two types of transformations, rather it is an illustration to the ideas in these two sections.



Original MRI Scan - Upper-left, WT - Upper-right
IFT - Lower-left, and IWT - Lower-right

Figure 11. Examples of the DWT, IFT and IWT.

3. Transform-Based Image Analysis

Let $T : S_1 \longrightarrow S_2$ be a transformation mapping the *signal space* into the (*transform*) space S_2. In our framework T is either the DFT or the DWT. Neither of the two is uniquely defined (as the Fourier transform is, for instance), but once we select a partitioning and an appropriate Fractal encoding scheme (for the DFT) or a particular Wavelet filter bank (for the DWT), our transforms are mathematically well-defined. At this point we can define a family of metrics on the transforms of the signals. For the DFT, for example, some of these metrics measure "distances" between the DFT's of the signals applied to other signals (in the signal space, S_1). Others measure the "distances" between signals using only their self-similarities (in the symmetry space, S_2). Still others measure those distances combining the

"good" features of the first two metrics; these are mixed-metrics defined on the Cartesian product space, $S_2 \times S_1$. Recall that the DFT of an image is a collection of a partitioning $R = \{R_i\}_{i=1}^m$ (of the image domain), and cover maps $v_i : D_i \longrightarrow R_i$ for each R_i. So, there are a variety of metrics one could design using these (fractal) coefficients, giving different weights on different coefficients, for example. For the DWT, one typically uses the l_2 norm on the compressed wavelet coefficients. Both of these metric strategies are partitioning-scheme and wavelet-basis dependent, respectively.

In order to compare two images we compare their transforms instead, Figure 12.

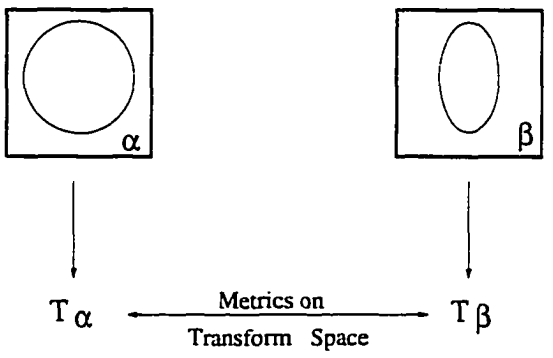


Figure 12. Quantitative comparison of images using their Transforms.

Again, there are at least three basic ways to compare two signals using their FT's. The first one compares the FT's applied to other signals. This takes place in the signal space, S_1. The second one compares the very FT's inside the symmetry space, S_2. The third one is a mixture of the first two. It compares the signals taking the "mixed-norm" of the difference in the product space $S_1 \times S_2$.

(FT.1) Let $||\bullet||$ be the L^2 or the *Sup* norm. [Remember: $Wf(f) \approx f$.]

To compare f and h in the signal space, S_1 one may look at:

◁ $||Wf(h) - f||$

- $||Wf(h) - h||$
- $||Wf(h) - Wh(f)||$
- $||(Wf)^k(h) - (Wh)^l(f)||$, for some $k, l \in \aleph$
- $||(Wf)^k(zero - signal) - (Wh)^l(zero - signal)||$
- Given an $\epsilon > 0$, find $k \in \aleph$ so that: $||(Wf)^k(h) - f|| < \epsilon$ and/or $||(Wh)^k(f) - h|| < \epsilon$,

where $(Wf)^k(h) = Wf(Wf(...(Wf(h))...))$ is the FT of 'f' applied to 'h' k times.

(FT.2) Inside the space S_2 the comparison depends upon the specific choice of a FT: For Example 1.22, presented in Section 1.3, one way to compare the signals 'f' and 'h' is to compute

$$\sqrt{\sum_{R[m,n]} \sum_{k=1}^{5} \frac{|H_k^f[m,n] - H_k^h[m,n]|^2}{KN}},$$

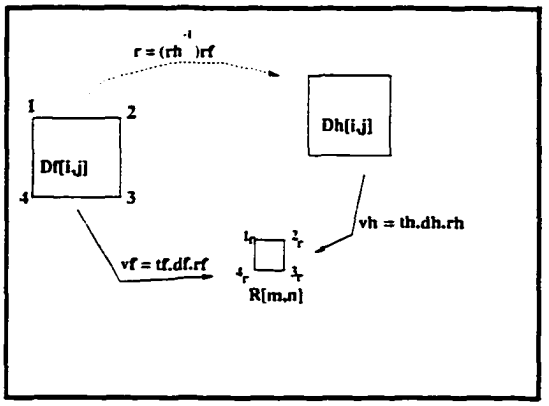
where KN is a constant of normalization, the $H_k^f[m,n]$'s determine the mapping $w_{[m,n]} : D[i,j] \times \Re \longrightarrow R[m,n] \times \Re$ for the signal 'f'. Similarly the $H_k^h[m,n]$'s encode the symmetries of the signal 'h'.

A better estimate would be to consider the 'i', 'j' and 'r' symmetries of the FT's. We compute Hf = "distance" between $R[m,n]$ and $Df[i,j]$ and Hh = "distance" between $R[m,n]$ and $Dh[i,j]$. Then subtract and normalize those difference, for all m and n, see Figure 13. We define:

$$Symm||f - h|| = \sum_{m,n} \frac{|Hf - Hh|}{KN}.$$

Another way to compare the signals is to compute the "distance" between $Df[i,j]$ and $Dh[i,j]$ directly, without explicitly going through $R[m,n]$. In this case we still use the 'i', 'j' and 'r' symmetries, only, but the results are more accurate, see Figure 13. [This measure is denoted by $Symm1||f - h||$]. More advanced metrics would involve weight coefficients produced by ANN's

(Artificial Neural Networks) - the ANN would be trained on a large enough set of images, and then it will assign valid weight coefficients to the different symmetry factors of the FT.



r = rf =	0	1	2	3	4	5	6	7	
0	rh=0	rh=1	rh=2	rh=3	rh=4	rh=5	rh=6	rh=7	big-small-big
1	rh=1	rh=0	rh=3	rh=2	rh=5	rh=4	rh=7	rh=6	s-b-s
2	rh=6	rh=3	rh=0	rh=5	rh=2	rh=7	rh=4	rh=1	b-s-b
3	rh=3	rh=6	rh=5	rh=0	rh=7	rh=2	rh=1	rh=4	s-b-s
4	rh=4	rh=5	rh=6	rh=7	rh=0	rh=1	rh=2	rh=3	b-s-b
5	rh=5	rh=4	rh=7	rh=6	rh=1	rh=0	rh=3	rh=2	s-b-s
6	rh=2	rh=7	rh=4	rh=1	rh=6	rh=3	rh=0	rh=5	b-s-b
7	rh=7	rh=2	rh=1	rh=4	rh=3	rh=6	rh=5	rh=0	s-b-s

Figure 13. Dependence between the "rotation" fractal coefficients.

(FT.3) A way to mix-compare the FT's in the space $S_2 \times S_1$ is to use only the contrast and brightness corrections (the 'c' and 'b' symmetries) of the FT's. For any pair $[m, n]$ we find the average of 'f' and 'h' on $R[m, n]$, linearly correct those, using the 'c' and 'b' symmetries, and then subtract and normalize them. The sum of all those subtracted and normalized differences we call the "mixed-norm1", MN1. The accuracy of this measure is surprisingly good.

In practice, we most often use another mixed-norm, MN2, defined by

$$||Wf - Wh||^2_{MN2} = \sum_{R \in A} |Hfh|^2 + \sum_{R \in B} (s_f f\_ave + o_f - (s_h h\_ave + o_h))^2 + \sum_{C} \chi_{\{Rf=Rh\}},$$

where Hfh is the distance between the covers Df and Dh of the range R in the

images f and h, respectively. Hfh is computed by averaging the the distances between the corresponding vertices of Df and Dh, paired together according to the action of the dihedral group D_4 (see the definition of $v_{[m,n]} : D \longrightarrow R$ in Section 1.2). The sets A, B and C we sum over are defined as follows: $A = \{$ all ranges common to both fractal decompositions $\}$, $B = \{$ all ranges that appear in the fractal partitioning of f or h, or both $\}$, and $C = \{$ all possible partitioning ranges of different sizes $\}$. The scaling and offset coefficients of f and h are denoted by s_f, o_f and s_h, o_h, respectively. And f_ave and h_ave are the averages of f and h over the corresponding ranges R. The MN2 measure is a pseudo-metric, since it does not necessarily satisfy the triangle-inequality, but it is very applicable because it uses all of the fractal coefficients in a meaningful way. One can construct real metrics on the fractal transforms of signals (for example, take only the last term of the definition of the MN2 metric, and define an induced equivalence class on the image space, S_1), but these are unlikely to incorporate all of the fractal similarity parameters.

Analogous types of measures could be defined on other different transforms of signals, for instance on the Wavelet Transform. Instead of the symmetry coefficients we will have to incorporate the differences of the compressed wavelet coefficients (details) at every level to compute the desired transform-based metric.

The importance of using such metrics, defined on transforms of signals, is that these metrics can be tuned to select features of interest. For instance, we can train an Artificial Neural Network (ANN), on a small test set of signals, to detect a specific feature (disease, abnormality) and then apply the algorithm to other real data sets.

We now apply these ideas to a set of four 2D slices of $[^{18}F]$ fluorodeoxyglucose PET scan datasets. The 2D slices correspond to anatomically equivalent

regions in the subjects. $M1$ and $M2$ are repeated 2D scans of a normal male subject. $F1$ is a scan of a normal female and $H1$ is a scan of a male subject with AIDS Dementia Complex [Rottenberg *et al.* 1996]. Figure 14 shows the four PET images and Table 1 contains the quantitative differences between the signals (based on their Fractal Transforms). In this example we show four different Fractal-Transform-based metrics. The magnitude of the metric-difference is inversely proportional to the degree of similarities between the images. Most of our metrics correctly group the PET scans to four clusters: $\{M1, M2\}$, $\{F1\}$, $\{M1, M2, F1\}$ and $\{H1\}$.

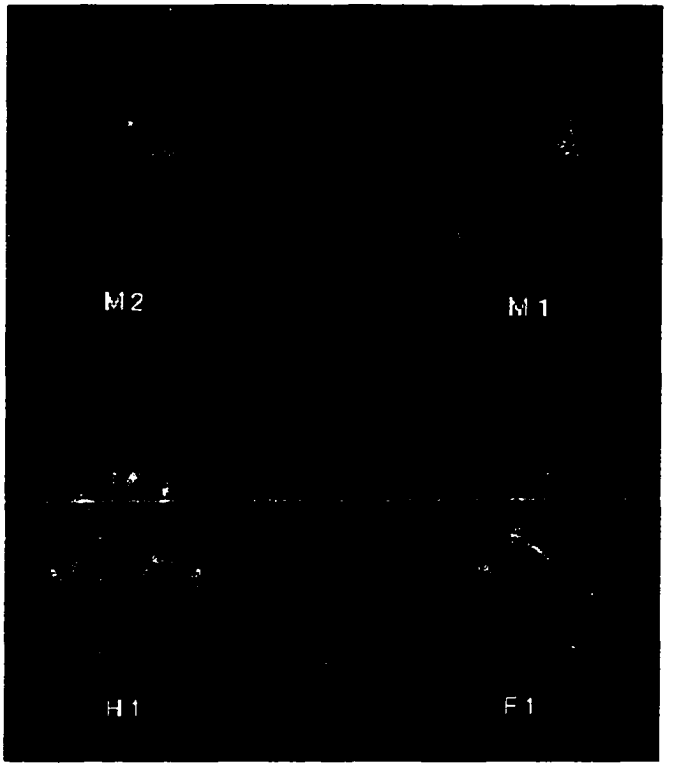


Figure 14. The four PET scans (test images).

We used a similar approach to study a set of 10 MRI brain scans. Since we had no prior knowledge about how many different subjects were scanned, we concentrated on finding the sibling to each of the MRI scans (i.e. finding

Table 1. Table of the Fractal Metrics applied to the 4 PET scans.

PET scans / METRICS	H1 - M1	H1 - M2	H1 - F1	M1 - M2	M1 - F1	M2 - F1	
SN - Symmetry Norm	706.59	741.53	733.10	472.03	748.50	721.60	
SN1 - Symmetry Norm 1	901.22	904.57	831.10	492.59	710.15	789.00	
MN1 - Mixed Norm 1	71.62	66.86	100.69	23.56	39.21	32.02	
MN2 - Mixed Norm 2	1126.23	1051.77	1579.86	372.07	618.33	506.50	

LEGEND: H1- HIV male; M1 - Normal Male 1-st scan; M2 - Normal Male 2-nd scan; F1 - Normal Female

the "closest" image for all of the images).

The MRI data are displayed on Figure 15, and Tables 2 and 3 show the "closest neighbor" tables for all signals with respect to the Fractal and Wavelet based metrics, respectively. We can clearly see that all of the four different DFT schemes and the DWT yield basically the same results: The data set groups the scans into $\{mri\_0, mri\_1, ..., mri\_5\}$ and $\{mri\_7, ..., mri\_10\}$. Tables 2 and 3 differ in that we use only one metric on the DWT space and four different metrics on the DFT space.

4. Image Magnification and Enhancement

When visualizing low-resolution images researchers frequently use interpolation methods to enhance the resolution and blow-up the picture. As the following example shows, to visualize (without "massaging") a low-resolution 128×128 PET scan and avoid the "blocking effect", one can *not* magnify the image more than $1 \times 1in^2$ (Figure 16).

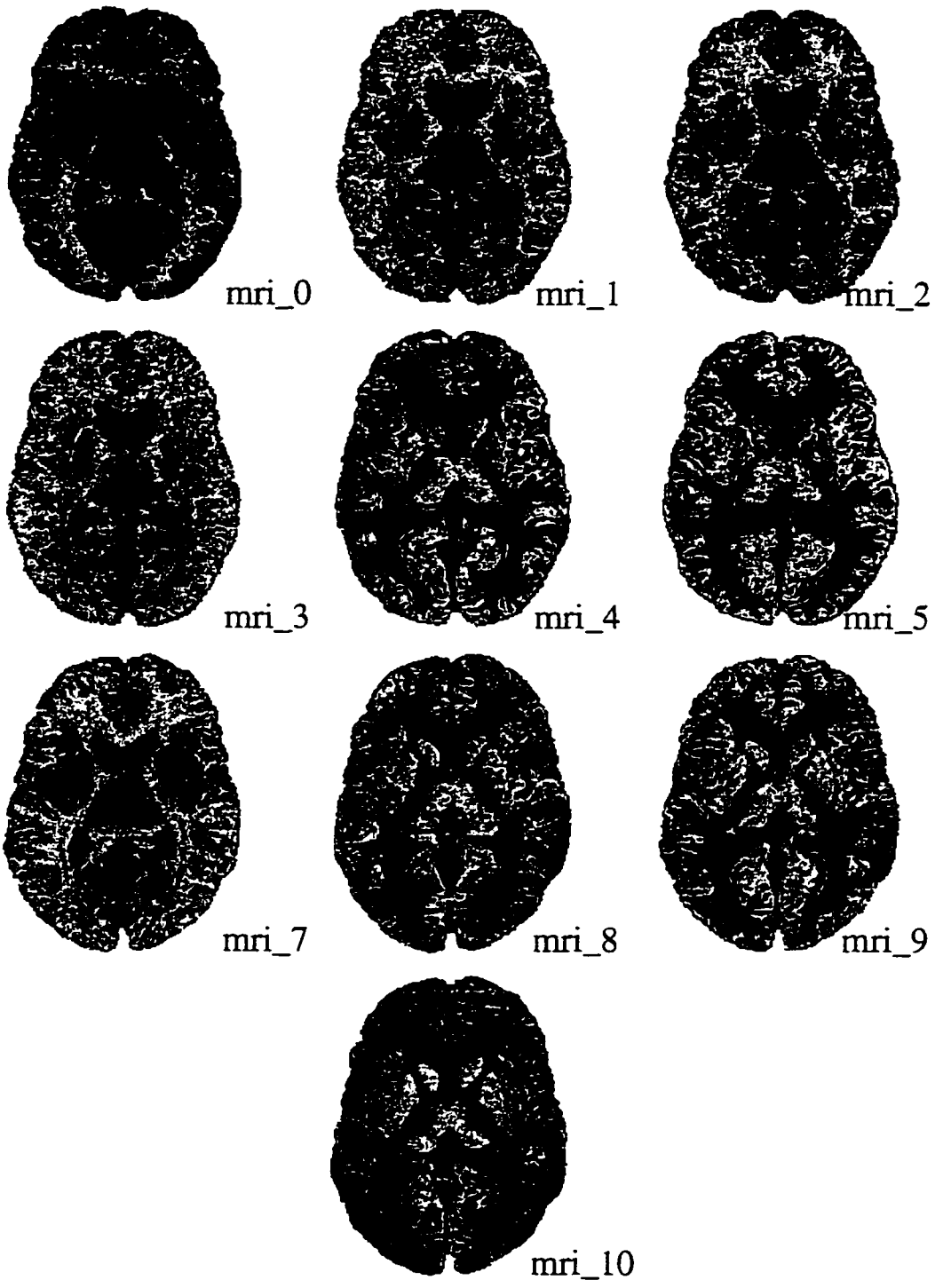


Figure 15. Ten MRI test images.

We now describe a fractal-transform based method (Barnsley, 1988) for zooming in and out on images. This fractal method outperforms current state-of-the-art *bi-linear* interpolation techniques. Although interpolation

Table 2. "Closest Neighbor" table for the 10 MRI scans, using Fractal metrics

Scans	Scheme I (Using the FTE that induces the convergent FTD algorithm)	Other 3 schemes
0	2	5 (or 2 or 1)
1	3	2 (or 4)
2	4 (or 3 or 1)	1 (or 3 or 4)
3	1	4 (or 1 or 2)
4	2 (or 3 or 1)	2 (or 3 or 1 or 5)
5	3 (or 2)	2 (or 0 or 1)
7	10	8 (or 9 or 10)
8	10	9 (or 7 or 10)
9	7 (or 8)	8 (or 10)
10	7 (or 8)	7 (or 9 or 8)

Table 3. "Closest Neighbor" table for the 10 MRI scans, using Wavelet metrics

Wavelet Transforms Of MRI Scans	"Closest" MRI	Distance	"Farthest" MRI	Distance
WTmri_0	2	2097	10	9534
WTmri_1	3	1191	8	6605
WTmri_2	3	1197	8	7435
WTmri_3	1	1191	10	6498
WTmri_4	5	1210	7	6437
WTmri_5	4	1210	7	6190
WTmri_7	9	2310	0	7463
WTmri_8	10	1229	0	9454
WTmri_9	10	858	0	8998
WTmri_10	9	858	0	9534

NOTE: For comparison $\|mri\_7 - mri\_3\| = 5850$, and mri_3 is the "closest" MRI to mri_7 according to the L2 measure.

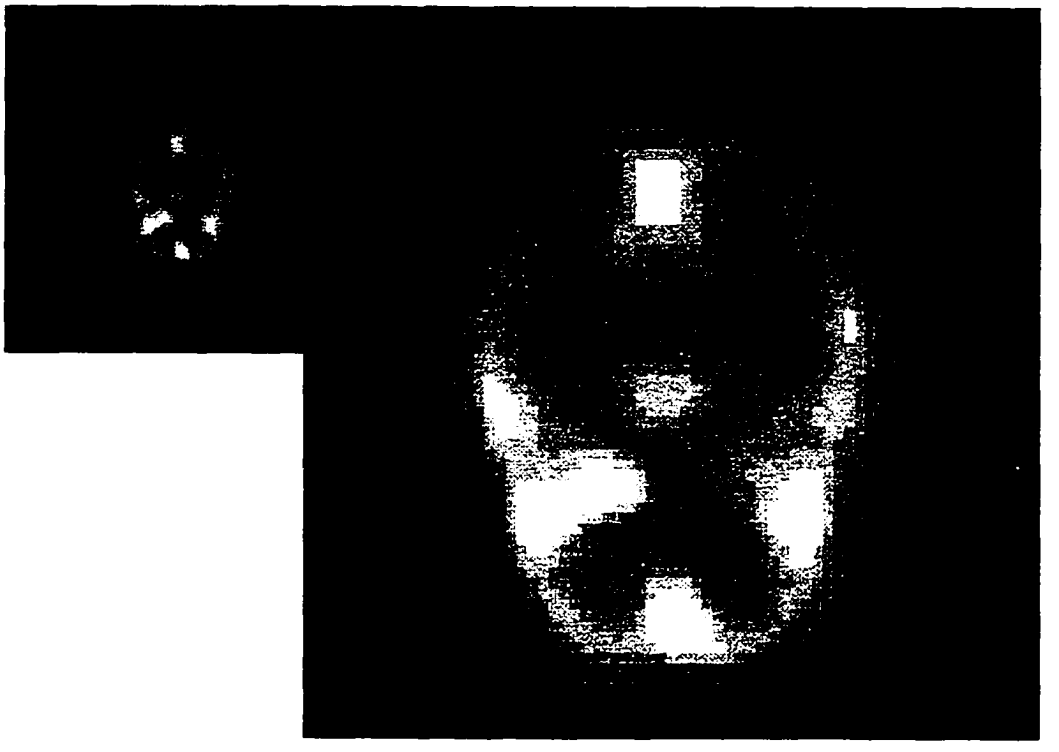


Figure 16. Blocking effects due to 128×128 to 512×512 magnification.

algorithms for resolution enhancement are of low complexity, relatively easy to implement and time efficient, they all carry the fundamental draw-back that they alter, blur and over-smooth the data. One example showing the advantages of the new fractal magnification algorithm is shown on Figures 17, 18 and 19. These images represent the lower portion of a 128×128 PET scan magnified to 2048×2048 again using these two methods. An interesting observation is that the fractal images have rougher (but very detailed) boundaries of the regions of (relatively) uniform densities. Also, they seem to exhibit some extra details that are not shown in the corresponding high-resolution interpolation images. It is true that some of these features could be artifacts of the (currently not perfect) DFT, however, some of them may not appear in the interpolation images due to smearing or over-smoothing effects.

Recall that the fractal images are "resolution independent", so we synthesized (recovered) the 2048×2048 fractal image on Figure 18 at lower (128×128) resolution and magnified it back to 2048×2048 using bi-linear interpolation. The result is shown in Figure 19. Not only does the interpolation smear and blur the data, it also "moves" details around (look at the high intensity spot in the middle - near the *central ventricle* - in Figure 18 and its re-positioning under the interpolation Figure 19 - right in the middle of the blob).

It is certainly worth pointing out that we can not expect any image zooming in (or out) technique to add up extra (new) detail to the pictures that is not already there. However, the fractal algorithm allows us to visualize and enhance features of the signals exploiting the self-similarities and self-affine symmetries of the data. It is more robust and accurate along edges and incorporates no smearing or over-smoothing effects which are fundamental for any interpolation technique.

5. Quantitative Warp Evaluation Schemes

5.1 Displacement Registration Fields (Warps)

In the field of medical imaging the identification and comparison of structures of interest between two images are fundamental for understanding and interpreting the data. In practice, researchers construct atlases on reference images (templates). These templates are studied thoroughly and used as models for particular types of data sets. The problem of data analysis through a template is that often the real images vary significantly from our models and hence our template atlases may not be of much help. The differences between the data and the template can be structural (size, orientation),

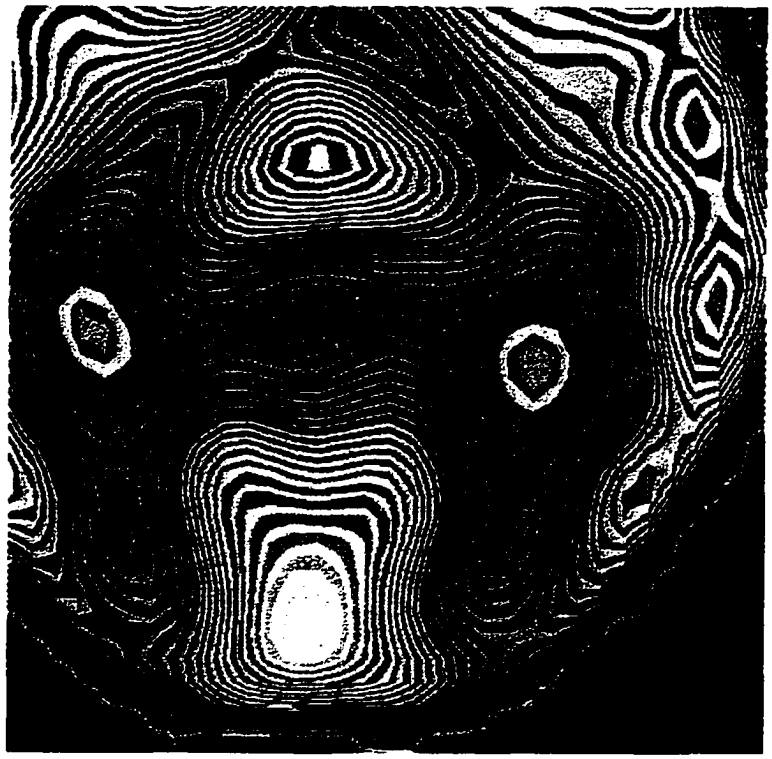


Figure 17. A portion of the interpolation blow-up, 128×128 to 2048×2048.

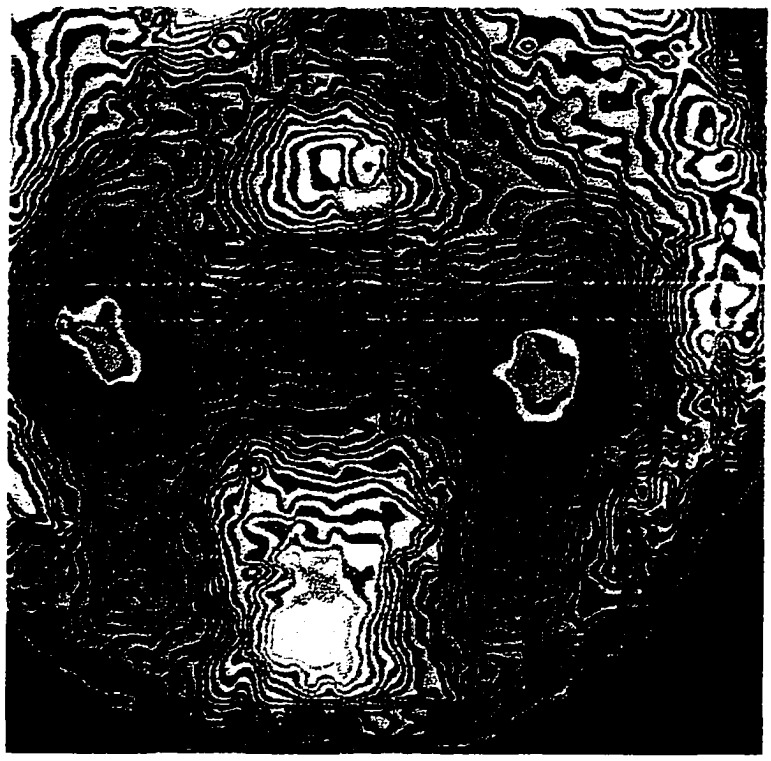


Figure 18. A portion of the Fractal blow-up, 128×128 to 2048×2048.

geometric-topological (features appearing in one image may not appear in

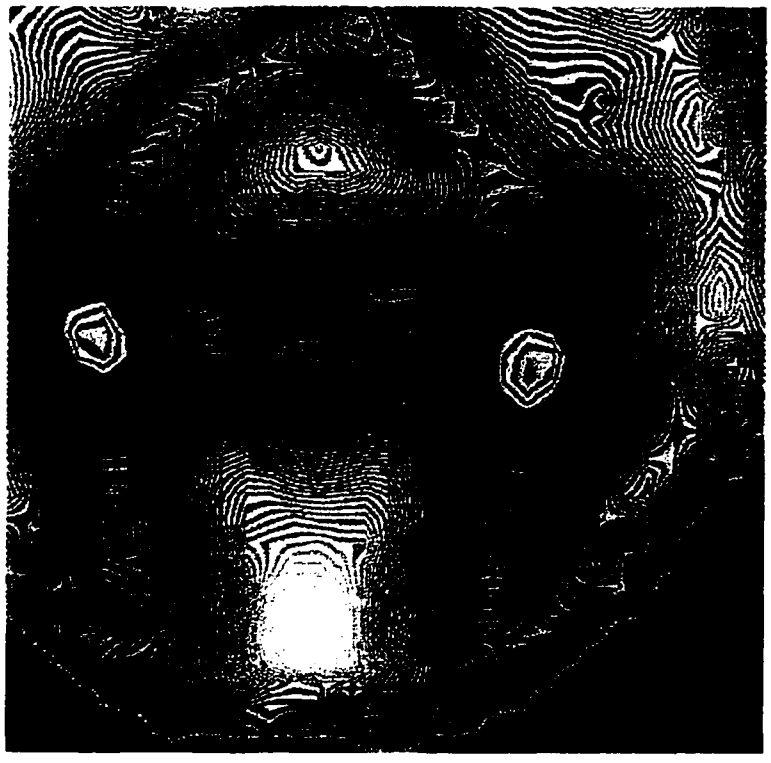


Figure 19. Smearing and over-smoothing effects of the interpolation algorithm.

other) or in the distribution of the image intensities (contrast, brightness). To overcome these problems registration techniques that align (warp) data onto the reference image (or vice-versa) are used. The aim of image warping is to simultaneously place all of the data images in a common anatomical (functional) reference frame.

Figure 20 contains a visual representation of the action of a 2D displacement field on an image. The reference data is shown in the upper-right and the target is in the upper-left corner. The deformed (warped) data is shown on the bottom-right. The warping field is also visualized as a grid deformation. One can clearly see the shifting, rotation and contraction components of the alignment (bottom-left).

There are two major warping approaches depending on the optimization procedure used to derive the deformation (warping) field that brings the data and the template in register. First are the density-based techniques

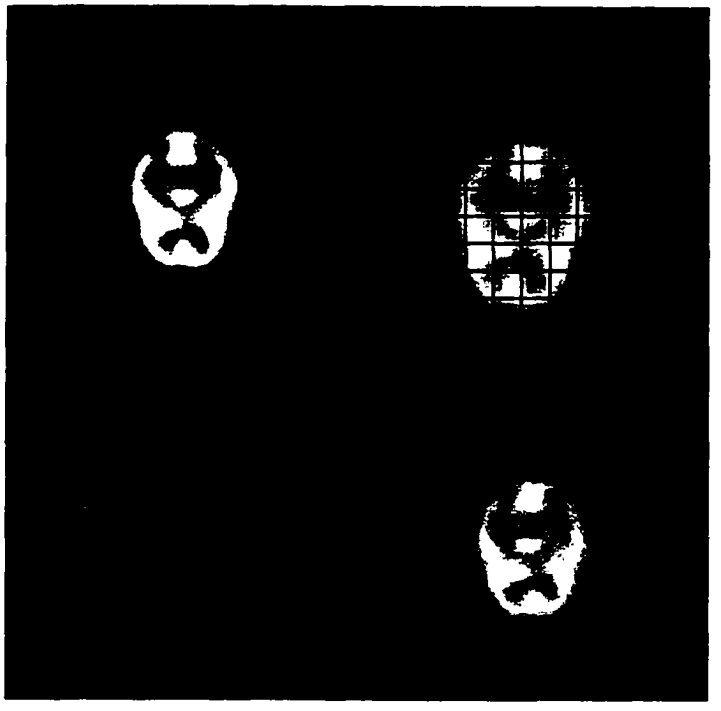


Figure 20. Example of an Original Data (top left), Template (top right), Displacement field action (bottom left) and Deformed warped data (bottom right).

which are purely intensity driven. They subdivide into *elastic warps* (Bajcsy [1989]), *viscous fluid warps* (Christensen *et al.* [1994]), and *polynomial warps* (Woods *et al.* [1992, 1993]). The second major warping approach is fiducially-based registration. In this case, a collection of landmarks (fiducial markers), like points (Bookstein, 1991), curves (Ayache & Faverjon, 1986) or surfaces (Thompson & Toga, 1996), are used to constrain the warping fields. The density-based warpings have the advantage that they do not require, in general, human (expert) intervention, have fast implementations and are applicable for a variety of data sets. On the other hand, the fiducial-based deformations are very accurate, robust, and in some cases allow for incorporating prior knowledge about the data into the model.

We begin by briefly outlining the ideas behind some commonly used affine and non-affine warping techniques. Let $A(v)$ and $B(v)$ be the template and

the data (subject) volumes indexed by the vector $v = (p, q, r) \in G$, consisting of discretely sampled and (8 bit unsigned integer) quantized voxel intensities a and b, respectively. The deformation field vector $\Phi^n_{i,j,k}$, at iteration n and at location (i, j, k) rooted at template voxel v, expresses a shift in coordinates so that the voxel intensities of $A(v)$ are mapped[1] to intensities in $B(v + \Phi^n_v)$. Depending upon the exact nature of this mapping (based on the acquisition protocols and the image scanning devices) we can use various voxel similarity measures such as least squares (LS), mutual information (MI), correlation coefficients etc. [Kjems *et al.*, 1997].

A volume of displacement vectors Φ placing the data over the template in a meaningful way may be computed in a recursive manner. Often, the vector field Φ is computed in a hierarchical manner, by first iterating the displacement field using a coarsely sampled vector grid on sub-sampled instances of the data and the template volumes. Then increasing the resolution whenever the similarity measure is below a certain threshold value.

Kjems *et al.* [1997] used a global cost function of the form

$$C(\Phi^n) = D(A(v), B(v + \Phi^n_v)) + R(\Phi^n),$$

where $v \in G$, $R(\Phi^n) = \frac{\alpha}{2} \sum_{i,j,k} \sum_{(i',j',k') \in N_{i,j,k}} ||\Phi^n_{i,j,k} - \Phi^n_{i',j',k'}||^2$ is a global regularization factor, α is a global parameter, $N_{i,j,k}$ is a set indexing the 6 nearest neighboring vectors of $\Phi^n_{i,j,k}$, and $D(a, b)$ is a function measuring the global similarity of a set of matched voxel intensities a and b. Iteratively the displacement field is updated with vectors $\Phi_{i,j,k}$. The vector field is smoothed in between iterations with a Gaussian spatial low-pass filter. The size of

[1] The notation Φ^n_v implicitly uses a tri-linear interpolation of the 8 immediate neighboring grid vectors $\Phi^n_{i,j,k}$ to find the displacement of the voxel at location v.

the Gaussian kernel, β, is carefully chosen along with the global regularizing parameter, α, in $R(\Phi)$.

Thompson and Toga [1996], on the other hand, modeled the images as viscous fluid solutions. Then the displacement field obeys the two-parameter Navier-Stokes equations

$$(\lambda + \mu)\nabla(\nabla \bullet \Phi(x)) + \mu\nabla^2\Phi(x) + F(x + \Phi(x)) = 0, \tag{13}$$

where the (elastic) parameters λ and μ determine the viscous properties of the solution and Φ is as usual the displacement field. $F(x)$ is a local internal force that drives the medium of the data into register with the target. The values of F are proportional to the gradient vector of a local image similarity function D. The displacement field is updated iteratively and the data is deformed until the external forces reach equilibrium with the internal restoring forces generated by the elasticity of the supporting material. The partial differential equations (13) are solved recursively on a finite grid and interpolated tri-linearly to obtain a continuous deformation field Φ [Thompson & Toga, 1996].

The affine polynomial warping fields are determined in a simpler manner than their non-affine counterparts. One typically minimizes a cost function

$$C(\Phi^n) = D(A(v), B(v + \Phi_v^n)),$$

where the image similarity measure D is either standard deviation of ratio images, or least squares, or least squares with intensity rescaling (AIR, Woods [1992, 1993]). The first cost function is image intensity independent, the second one assumes image intensities have been equalized first, but it allows for fast registration. The last measure tries to avoid the problems of least squares by adding an intensity scaling term to the model. Note that there is no need for a smoothing (regularization) factor in the cost function

because of the nice properties of affine/linear functionals. The displacement field Φ_v is again computed iteratively, however it has a particular (affine) form

$$\Phi_v = \Phi \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} f_1(x,y,z) \\ f_2(x,y,z) \\ f_3(x,y,z) \end{pmatrix},$$

where each $f_i(x,y,z) = a_{i,0} + a_{i,1}x + a_{i,2}y + a_{i,3}z : R^3 \longrightarrow R$ is a linear function. The (globally determined) 12 parameters $a_{i,j}, 1 \leq i \leq 3, 1 \leq j \leq 4$, minimizing the cost function $C(\Phi)$ define an affine spatial displacement field that brings the data and the target in (affine) register, modulo the choice of the measure of image similarity function.

5.2 Registration Evaluation in 2D

Based on the transformation approach we discussed in the previous sections we can develop a *goodness* of warp classification scheme. There seems to be no commonly accepted rigorous definition of what a "good" warp should be. Different criteria are used for different situations. We have assumed that a "good" warp should be one that produces an image approximating closely the target (with respect to some metric), yet deforming the initial (template) image the least.

For example, if α is a PET, an MRI or fMRI scan (image), and $\widehat{\alpha}$ is the result of applying a warp Φ to α, (with a target image β), one would expect that a "good" warp preserves the local symmetries of α (onto $\widehat{\alpha}$) and simultaneously yields an image ($\widehat{\alpha}$) having self-similarities close to these of the target β. Hence, a way to tell the "good" from the "bad" warps is to evaluate functionals like:

$$F(\alpha,\beta,\Phi,T) = \left(a+b-c \quad , \quad \left(\frac{c^4}{|c-a|} + \frac{c^4}{|c-b|} \right) \right),$$

where μ_1 is a metric on the transform space (S_2), $a = \mu_1(T(\alpha), T(\widehat{\alpha}))$, $b = \mu_1(T(\widehat{\alpha}), T(\beta))$ and $c = \mu_1(T(\alpha), T(\beta))$. The smaller the values of the components

of F, the "better" the warping transform (i.e. the warp preserves the most the self-similarities of the test image, and still produces an image "close" (in the transform-space metric) to the target).

As shown on Figures 21 and 23, various functionals like these may help simultaneously minimize $\mu_1(T(\alpha), T(\widehat{\alpha}))$ and $\mu_1(T(\widehat{\alpha}), T(\beta))$, penalizing warps which bring $T(\widehat{\alpha})$ too close to $T(\alpha)$ or $T(\beta)$, unless $T(\alpha) \approx T(\beta)$ in the transform metrics. Among these the classifying functional D is the most accurate fit based on our goodness of warp criterion placing the gold-standard warp halfway between the data and the target, in transform space.

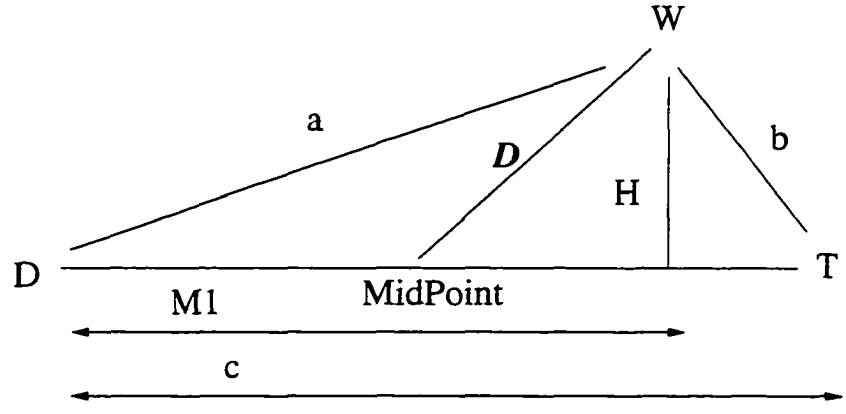
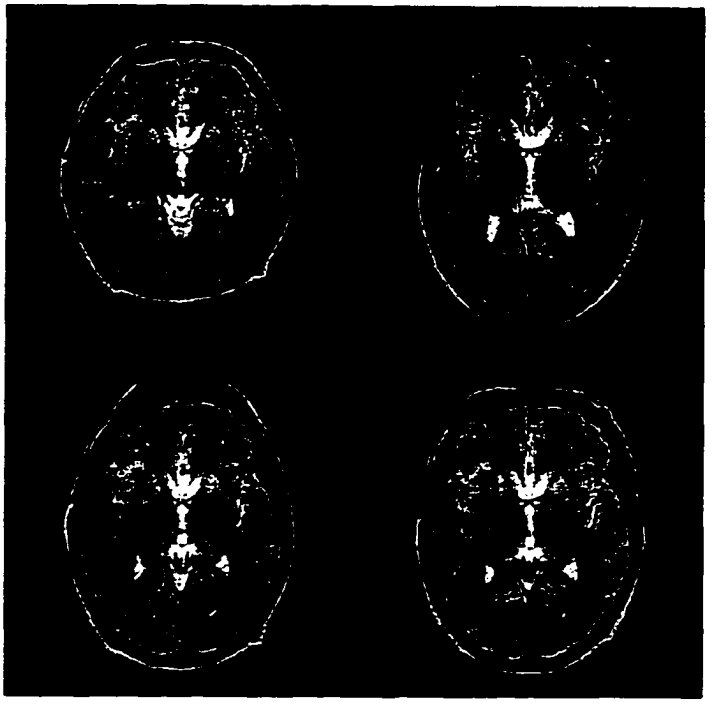


Figure 21. Various transform-based classifying functionals.

We apply these ideas to classify two warps transforming an axial MRI slice of an oriental subject (test image) to the corresponding axial slice of an occidental subject (target), Figure 22. The first warp is induced by a *linear* spatial deformation and the second one represents a *non-linear* displacement field. Both the Wavelet based and the Fractal based metrics yield smaller values of the two components of the classifying functional F for the second deformation, and therefore, give a uniform preference to the *non-linear* warp (Table on Figure 23).



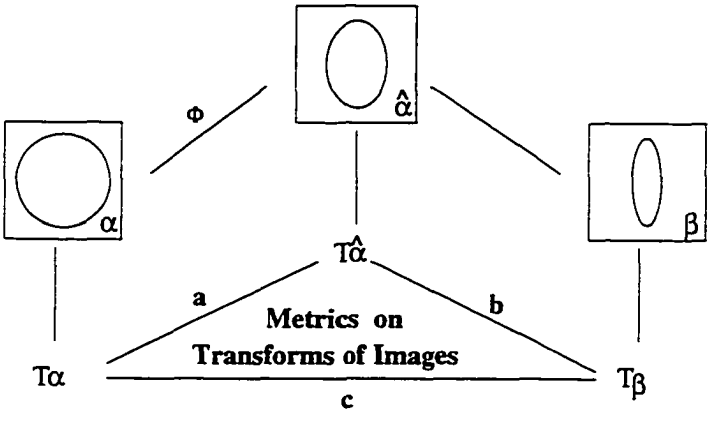
Upper-Left: Test Image
Upper-Right: Target Image
Lower-Left: Linear Warp of Test Image
Lower-Right: Non-Linear Warp

Figure 22. Test Images.

5.3 Applications of the Transform-based Technique to Stereotactic Warp Classification

5.3.1 Quantitative Evaluation of Polynomial Warps. In this section we numerically characterize and evaluate the performance of a number of (linear and non-linear) polynomial registration techniques. We used R. Woods' AIR 3.0 (Automatic Image Registration) package [Woods et al., 1992, 1993] containing warps with 6, 7, 9, 12, 30, 60, 105 and 168 parameters.

We show two examples. The first one characterizes 4 deformations based on registering a single MRI onto an "average" MRI representation volume. In this case we used the 9, 12, 30 and 168 parameter warps independently of each other (unsequentially), i.e. the results of each warp were not used as initialization (starting) point of the next (higher-order) deformation algo-



	Using DFT		Using DWT	
	F1	F2	F1	F2
LINEAR WARP	235	2.19	62	53.90
NON-LINEAR WARP	69	0.48	50	28.07

Warp classification functional: $F(\alpha, \beta, \Phi, T) = (F1, F2)$

$F1 = a+b-c;$ $\quad F2 = c^{4}\left(\frac{1}{|c-a|} + \frac{1}{|c-b|}\right)$

Figure 23. Comparing signals and warps of signals using their Transforms.

rithm.

The second example quantifies the performance of the 7, 12, 30 and 168 parameter warps, based on registering a single MRI volume to another single MRI. In this latter case, we did use the warping techniques sequentially. In other words, the output of each consecutive warping method was used as an input of the following (higher-degree) warp. The goal was to determine the validity and the robustness of this quantitative transform-based warp classification scheme.

We begin by looking at the "single-to-average" example. Different views

of the data and the target of the warps are shown on Figure 24. There is no prior registration done on the two volumes. The data represents an "average" MRI volume of 53 normal individuals (MNI study) and the target is a normal subject, not included in the 53 normal-average. Obviously, the differences of the data sets are not only spatial and structural but also similarity-like.

Figures 25 and 26 show sagittal, coronal and axial (transversal) slices of the warped (resliced) volumes, under the 4 polynomial fields.

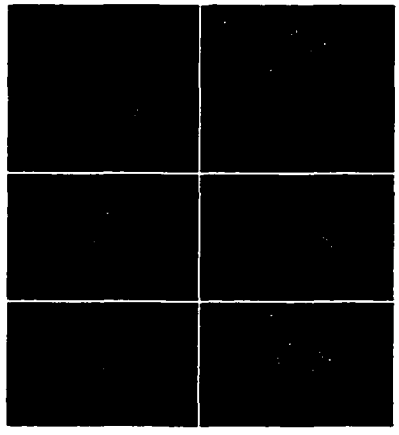


Figure 24. Sagittal, coronal and axial slices of the Target (left) and the Data (right).

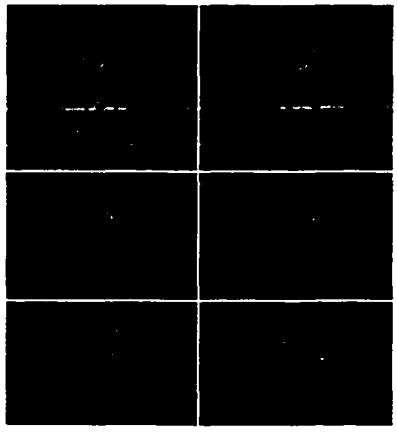


Figure 25. Linear 9 (left) and 12 (right) parameter polynomial warps.

Table 4 contains the final wavelet-based metric distances numerically characterizing the two linear and the two non-linear warps (see Section 5.2 for interpretation of a, b, c, $F_1, F_2, H, M1, D$). The distortion of the warps

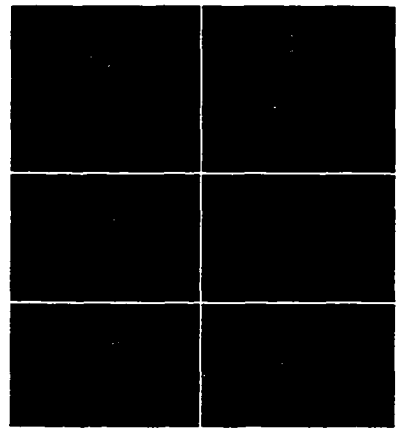


Figure 26. Non-linear 30 (left) and 168 (right) parameter polynomial warps.

whose resliced volumes lie on the "mid-line", Figure 27, is exactly right, since they produce a resliced volume having symmetries resembling the ones of the target the closest, modulo preserving most of the self-similarities of the template. Note that according to that classification the ultimate warp would produce a volume with a transform lying exactly on the intersection of the "symmetry" and the "mid-" lines, Figure 27, halfway between the reference image and the target.

Table 4. Transform-based polynomial warp classification table.

Classification functions / Warps	F1	F2	H	M1	D
Wp1 (9 par aff.)	7.30	1.7775	7.70	0.9384	14.31
Wp2 (12 par aff.)	7.17	1.6487	7.60	0.9376	14.24
Wp3 (30 par NL)	7.26	1.8968	7.63	0.9405	14.32
Wp4 (168 par NL)	19.44	0.4956	19.04	0.5216	19.03

Figure 27, illustrates a geometric representation of the quantitative in-

formation contained in Table 4. It is worth mentioning that our warp classification presented its geometric interpretation are template-target dependent. This is a consequence of the fact that the classifying functionals depend upon α (template), β (target), Φ (type of warp) and the discrete transform used in the model. In summary, this first example shows preference of the 12 parameter affine warp. It produced the closest resliced volume to the ultimate volume representing the intersection of the symmetry and the mid lines.

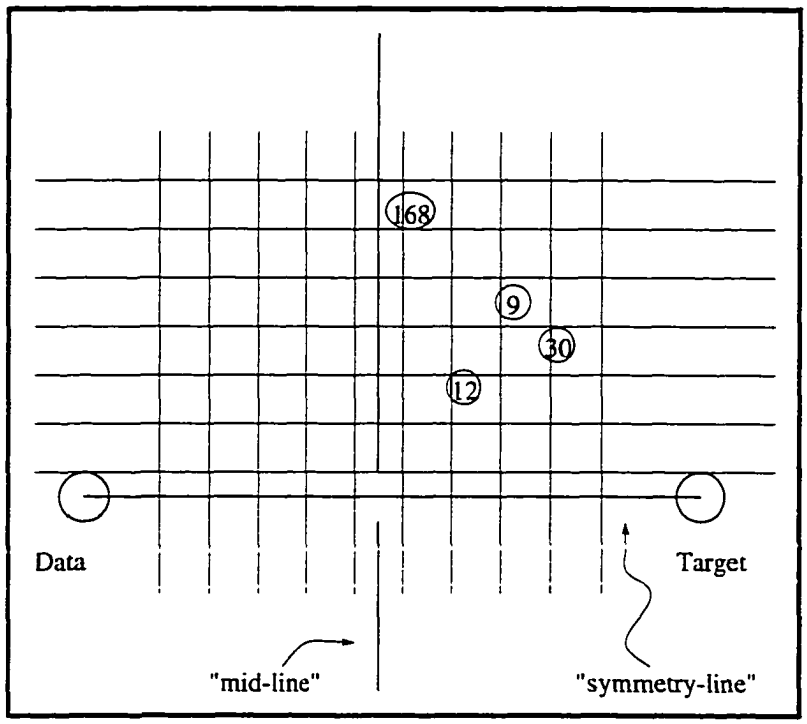


Figure 27. Pseudo-planar visualization of warp performance in symmetry (DWT) space.

We now proceed with our second ("single-to-single") registration and the corresponding warp (performance) evaluation. The template and the target are shown on Figure 28.

The resliced (warped) volumes are displayed on Figures 29 and 30. Again, registration and reslicing was done using Woods AIR3.0 package.

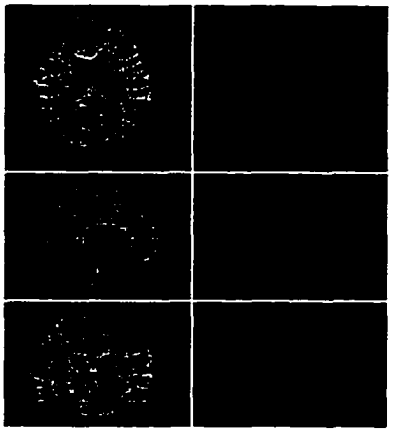


Figure 28. Axial, sagittal and coronal slices of the Target (left) and the Template (right), (ex. 2).

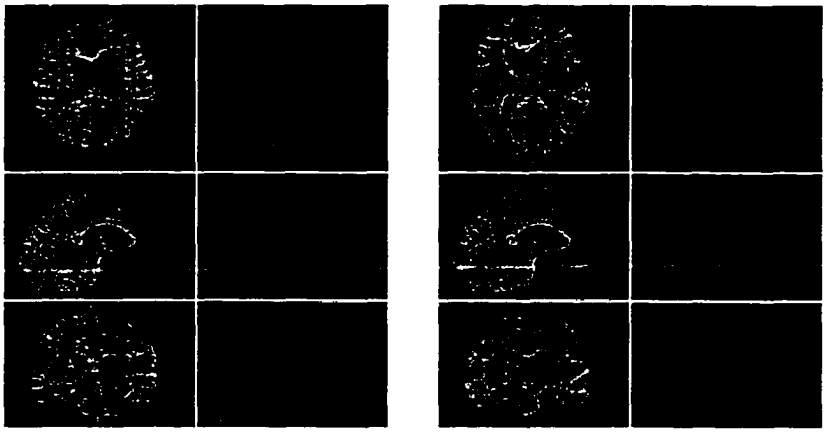


Figure 29. Target and Linear 7 parameter (left); and Target and Linear 12 parameter warp (right) (ex. 2).

The numeric transform-based estimates of the performance of the 4 polynomial warps are listed in Table 5. And the corresponding planar visualizations are depicted on Figure 31. The functional H, height, determines the overall global distortion of the warp in terms of the distance between the "symmetry-line" and the warped-resliced volume (in the transform space

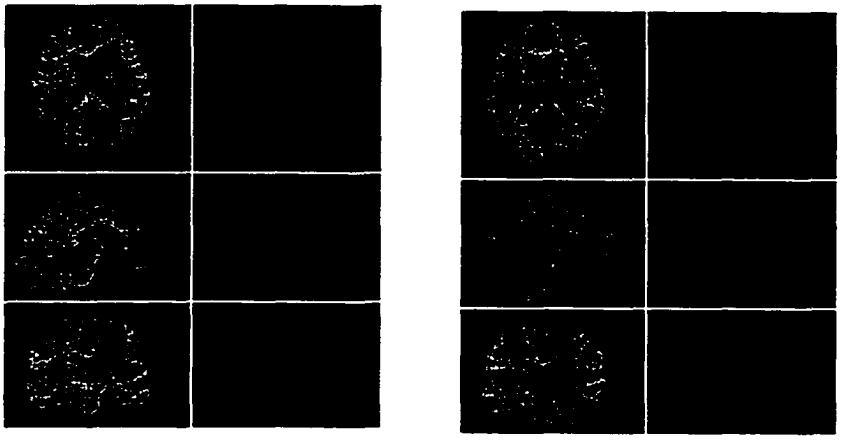


Figure 30. Target and NL 30 parameter (left); and Target and NL 168 parameter (right) resliced volumes (ex.2).

metric). In this case, the 168 parameter (non-linear polynomial) warp almost uniformly outperforms the other three warps (columns 4 and 5, Table 5).

Table 5. Numeric transform-based polynomial warp classification table (ex. 2).

Classification functions / Warps	F1	F2	H	M1	D
Wp1 (7 par aff.)	5.17	3.1677	5.44	0.0171	20.18
Wp2 (12 par aff.)	15.49	0.3362	17.28	0.1926	21.24
Wp3 (30 par NL)	16.03	0.2163	18.81	0.2947	20.53
Wp4 (168 par NL)	15.25	0.1812	18.67	0.3524	19.57

5.3.2 Quantitative Evaluation of Non-affine Warps. In this section two main problems are discussed. The first one is to quantitatively evaluate

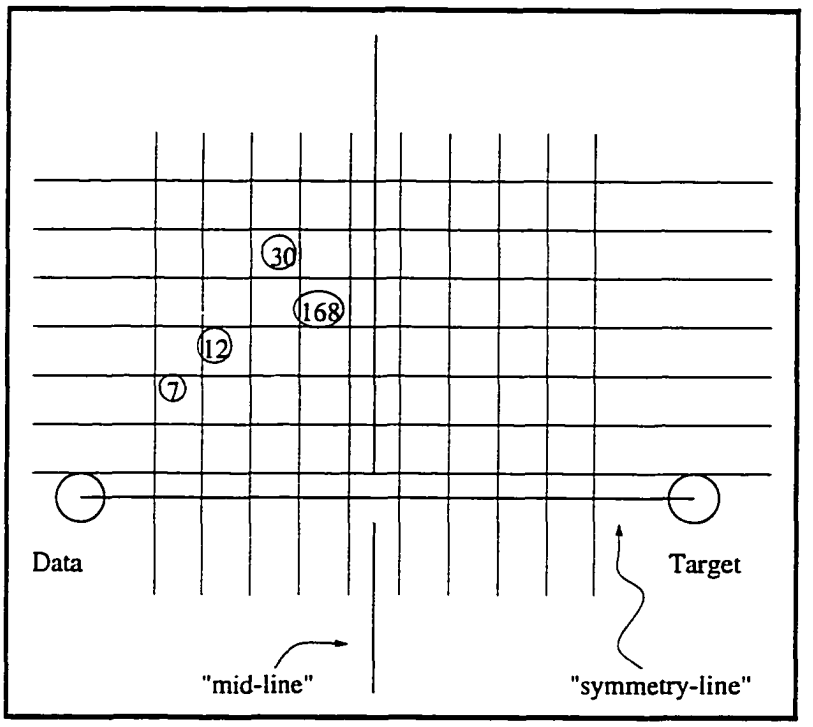


Figure 31. Pseudo-planar visualization of quantitative warp characterization (ex. 2).

and examine the performance of two new non-affine registration techniques (developed by Kjems et al., [1997]) and one affine polynomial warp (12 parameter, as implemented by Woods, et al., [1992]). Our second aim is to investigate and compare our performance estimates with the univariate warp evaluation (in inter-subject space) as proposed by Kjems *et al.* [1997].

Five right-handed subjects were scanned 8 times (randomized 2 baseline and 6 gradually increasing activation) in 2 scanning sessions. The subjects performed visually guided voluntary anti-saccades. This experiment involved fixating on a central LED until a random target LED appears. Then the task was to saccade to the LED contralateral to the lit target. The frequency of the target appearance varied from scan to scan (0.05, 0.1, 0.3, 0.5, 0.7, 0.9 Hz). The two baseline scans were acquired with the subjects fixating on the central LED. A whole body PET scanner (Advance, General Electric) was

used with an image spatial resolution of 5 mm in all directions. The volumes were reconstructed with 35 slices and voxel size $2.0 \times 2.0 \times 4.25mm^3$ covering a field of view of $25.6 \times 25.6 \times 15.2cm^3$. In addition T1-weighted MRI scans for the 5 subjects were obtained using a Siemens Magnetom Vision scanner in the Hvidovre Hospital, Copenhagen, Denmark. The voxel sizes for two of the subjects were $0.98 \times 0.98 \times 1.17mm^3$, and the remaining 3 subjects and the target (template) all had voxel sizes of $0.98 \times 0.98 \times 1.00mm^3$. A simulated PET image was registered to the template's MRI and was used as the template's functional volume. A common preprocessing step was done on all functional images. The 8 functional volumes for each subject were intra-subject aligned using a 6 parameter rigid body deformation (Woods [1992, 1993]). The structural MRI volumes were corrected for intensity inhomogeneity, predominant in the axial (z) direction, by computing the mean of the top 5 % intensity voxels for every transverse slice containing brain tissue. Then the intensities were normalized by fitting a 6-th degree polynomial as a function of the slice number. After that the MRI scans were carefully stripped from the skull and the dura using a manual interactive drawing tool.

The results of the CVA/SSM-based multivariate analysis of warp performance (of Kjems *et al.*) are summarized in Table 6. We see a direct decrease of the variance measures that occurs with the increase on the "non-linearity" of the field (first 5 rows of Table 6). According to this measure the non-affine warping using MI (Mutual Information) minimization functional is the best among the 3 warps. On the other hand, using CVA analysis Kjems *et al.* argue that the LS (Least Squares) driven non-affine field is the warp of choice, due to "over-warping" effects of the MI warp (CVA ranking, bottom 2 rows, Table 6).

Table 6. CVA/SSM-based multivariate analysis and quantitative warp evaluation.

Warping Method / Analysis	Warp_1 (LS)	Warp_2 (MI)	Warp_3 (Affine)
σ_T^2	1.83×10^{-2}	1.75×10^{-2}	2.06×10^{-2}
σ_e^2	1.02×10^{-2}	0.91×10^{-2}	1.32×10^{-2}
σ_a^2	1.10×10^{-2}	1.12×10^{-2}	1.09×10^{-2}
SSM: $\sum_{i=1}^{4} \lambda_i / I$	0.84×10^{-2}	0.75×10^{-2}	1.11×10^{-2}
Warp Ranking	II	I	III
CVA: $\lambda_1 / \sum_{i=2}^{5} \lambda_i$	7.65	7.31	5.51
Warp Ranking	I	II	III

As Kjems *et al.* point out the decrease of the total variance σ_T^2 is mainly due to a decrease of the inter-subject variance σ_e^2, the intra-subject variability is pretty stable across the three warping schemes (Table 6).The first 4 eigen-values of the SSM analysis absorb most of the inter-subject variance drop (Kjems *et al.* [1997]). These results indicate that the non-affine warp based on MI (mutual information) image similarity measure is superior to the other two methods. However, this is a direct analysis of variance measures influenced by the largest impact of the MI warp on the functional alignment. Such evaluation of image registration does not measure how well *functional* alignment is achieved, i.e. the magnitude of the signal-to-noise ratio of the observed (warped) signal. In the second part of their analysis Kjems *et al.* investigate how well the brain state is reflected in the images, assuming that the optimal registration extracts the brain state most clearly.

The CVA analysis gives information on the influence of the frequency

of the saccadic eye movement on the scans in high dimensional space. If one groups the images according to this frequency (7 groups) the canonical vectors form an orthogonal basis that maximally separates these groups. Assuming the relation between scan number and the frequency is linear then the variation of the 7 groups can be described by a linear model. Although the non-linear nature of the brain activation is well documented (eg. Morch *et al.* [1997]) one can still expect for this experiment the brain activation to be approximately proportional to the saccadic frequency. This in turn means that the functional activation should be captured by the first canonical component (λ_1). Under this assumption the ratio $\lambda_1 / \left(\sum_{i=2}^{5} \lambda_i \right)$ is an indicator of how well the brain activation signal is captured. In this sense, Kjems *et al.* conclude that in fact the LS based warp is the registration of choice, because of the "over-warping" effect of the MI non-affine alignment (last 2 rows, Table 6). In other words, this is evidence that very detailed (structural) registration can *degrade* the functional alignment of images.

As opposed to the SSM/CVA analysis of Kjems *et al.*, in our wavelet-based warp evaluation technique we used one classification functional (D), which does not consider changes of different variance measures. Thus, it is subject dependent. In the tests we have run so far, however, this functional exhibits small inter-subject variability. The induced warp ranking, according to the classifying functional D, is: (best to worst) MI, LS, Affine warp. This means that no "over-warping" effects are detected by D. Table 7 contains the values of the wavelet-based classification and the corresponding ranking of the three registration schemes. The performance of the 3 warping techniques (in wavelet space) are visualized on Figure 32, the best performing warp is on the bottom and the worst is on the top.

The columns on Figure 33 contain (left-to-right) the original MRI data,

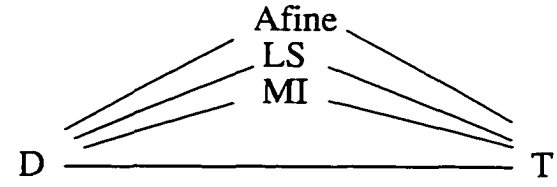


Figure 32. Overall visual representation of the wavelet-based registration evaluation on MRI data.

Table 7. Wavelet-based quantitative warp evaluation on the MRI data.

Warping Scheme / DATA	Warp_1 (LS)	Warp_2 (MI)	Warp_3 (Affine)	Warp Ranking		
				Wapr_1	Warp_2	Warp_3
MRI_1	15.71	15.53	15.92	II	I	III
MRI_2	16.44	16.29	16.38	III	I	II
MRI_3	14.19	13.89	14.28	II	I	III
MRI_4	15.11	15.01	14.94	III	II	I
MRI_5	13.91	13.87	14.01	II	I	III
Overall Warp Ranking (Across subj.)	II	I	III	Legend: I = Best		
				II = Medium		
				III= Worst		

the LS warp (registration using least squares as image similarity measure), the MI warp, the Affine (12 parameter polynomial) warp, and the target of the registration (template). All images were magnified (in frequency space) using the Fourier transform, and we are showing the 150^{th} axial (transverse) slice (out of 256) for each image. There are 5 rows for the five different subjects involved in this study.

Visualizations of the performance of the three warping techniques, according to Table 7, are shown in Figure 34. Depending on the prior assumption of goodness of warp one selects the best registration (for each data set

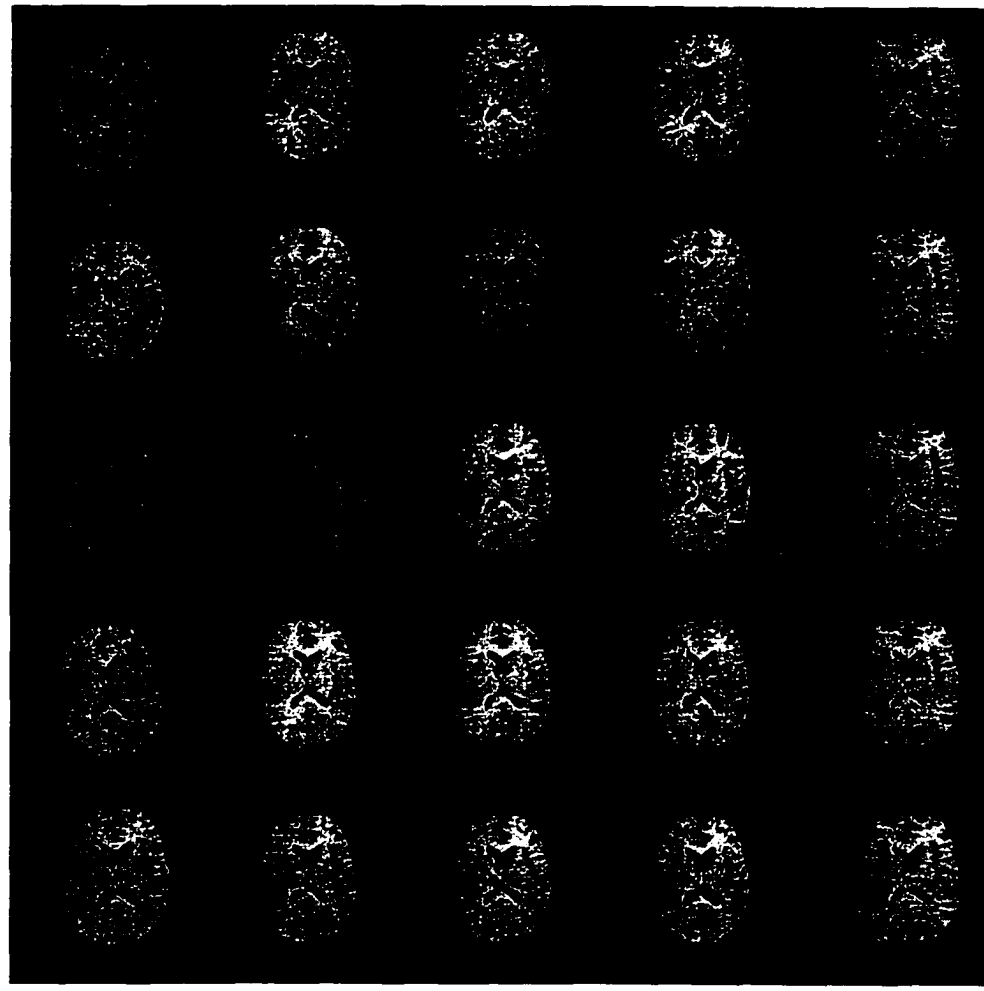


Figure 33. Original data, Warped (resliced) images and the Template. Each row contains (left-to-right): Original MRI data, LS warp, MI warp, Affine alignment, and Template.

separately), based on these diagrams. We explicitly prefer the alignments producing warped volumes closest to the intersection of the "symmetry-line" (horizontal line segment joining the WT of the Data and the WT of the Target) and the "mid-line" (vertical line through the midpoint of the symmetry-line) in transform space. Note that according to this prior goodness of warp hypothesis the ultimate ("best") warp lies at the intersection of these two lines, halfway between the data and the target.

We now present the analogous wavelet-based results for the same 3 registration techniques where the warp evaluation was done on the functional

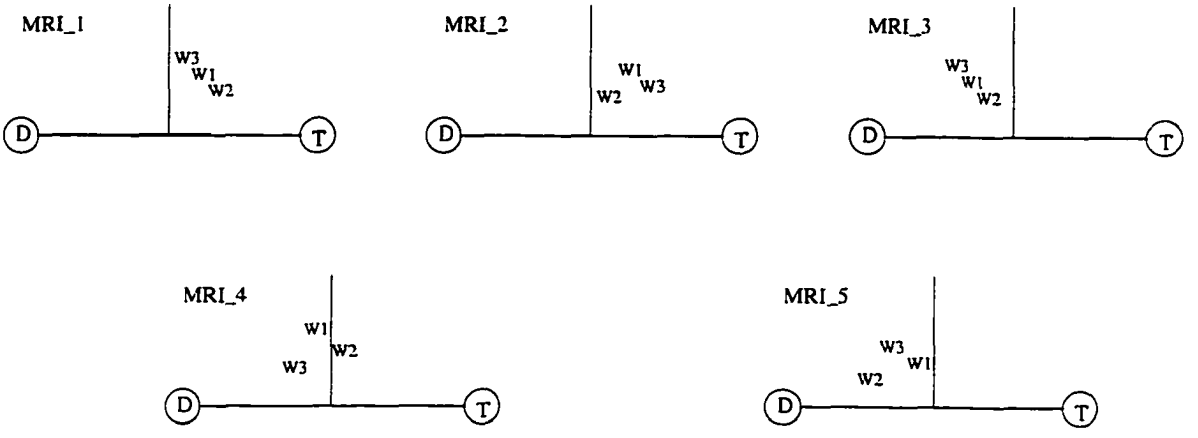


Figure 34. Planar representation of the quantitative warp evaluation on the MRI data in wavelet space.

data sets (PET). Each of the 5 subjects was scanned 8 times; twice under baseline and once under six different saccadic eye frequency (activation) paradigms. To analyze the warp performance we used the averages of the 8 (pre-registered) volumes for each subject. The MRI-derived displacement fields were applied to bring the corresponding functional PET images, Figure 35, in register with the template. Then the wavelet metrics between the data, the target and the warped volumes were computed, Table 8, and the visual interpretation of the results is shown on Figure 36. The functional D again selects MI as the best warp.

There are many natural questions arising in regard to the results listed in Tables 7 and 8, and the induced visual representations of warp performance shown on Figures 34 and 36. We will now attempt to address the following concerns: What are the differences and the similarities between the MRI-based and the PET-based wavelet registration classification schemes? Do we gain anything by studying the images in (reduced) wavelet space as opposed to the analogous study using the raw data sets (in image space)? How stable

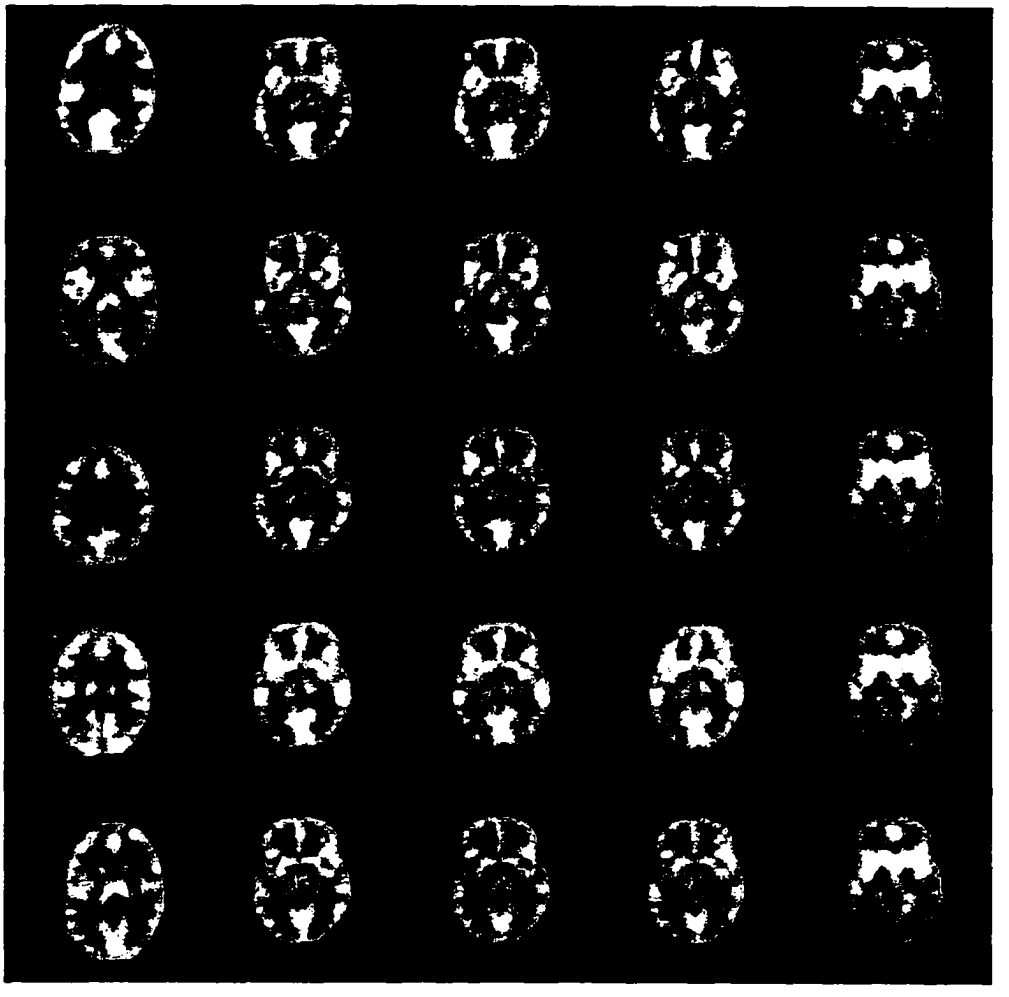


Figure 35. Original data, Warped (resliced) images and the Template. Each row contains (left-to-right): Original PET data, LS warp, MI warp, Affine alignment, and Template.

are the results for various wavelet bases?

To answer the first question we look at the overall ranking in Tables 7 and 8. The functional D measures the distance between the WT of the warped-resliced volume to the midpoint of the symmetry-line between the data and the target. The "gold standard", our best performing warp, is the one for which $D = 0$. In a group of warps, the ultimate warp is the one that minimizes D. In both tables D ranks *Warp2* (MI) as the best performing registration, across the 5 data sets. Again, the values in Tables 7, 8 and 10 represent the wavelet distances between the mid-point (halfway between the

Table 8. Wavelet-based quantitative warp evaluation on PET data.

Warping Scheme / DATA	Warp_1 (LS)	Warp_2 (MI)	Warp_3 (Affine)	Warp Ranking		
				Wapr_1	Warp_2	Warp_3
PET_1	19.29	19.49	19.21	II	III	I
PET_2	15.41	15.21	15.47	II	I	III
PET_3	19.89	19.76	20.03	II	I	III
PET_4	18.53	18.42	18.64	II	I	III
PET_5	19.77	19.23	18.93	III	II	I
Overall Warp Ranking (Across subj.)	--	I	--	Legend: I = Best		
				II = Medium		
				III= Worst		

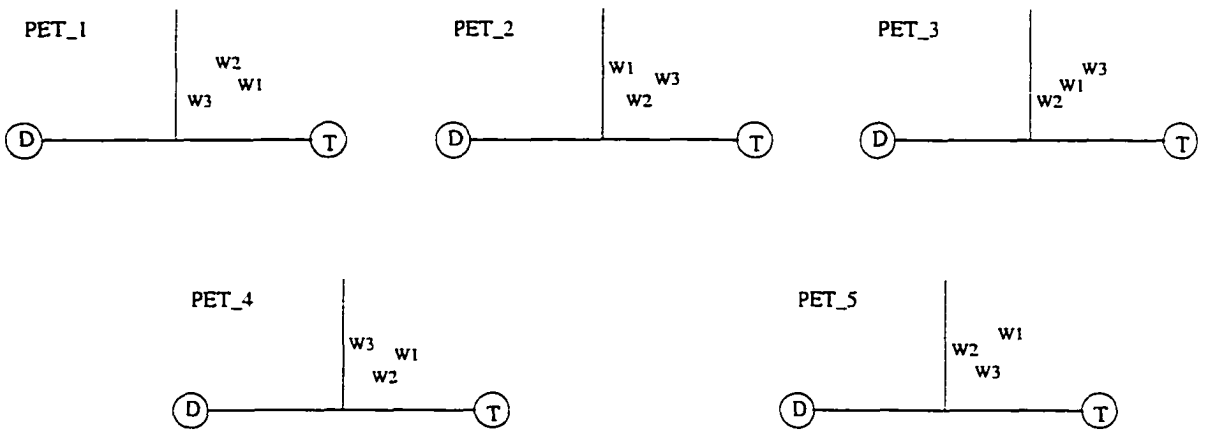


Figure 36. Planar representation of the quantitative warp evaluation on the PET data in wavelet space.

WT of the data and the WT of the target) and the WT of the warped image. These distances are computed using the L_2 norms on the "reduced" WT's (the top 5%, in absolute value, of the wavelet spectral coefficients). This makes sense because the "large" wavelet coefficients capture the essence of

Table 9. Quantitative warp evaluation on the structural MRI data in image space.

Warping Scheme / DATA	Warp_1 (LS)	Warp_2 (MI)	Warp_3 (Affine)	Warp Ranking		
				Wapr_1	Warp_2	Warp_3
MRI_1	27.72	27.49	27.32	III	II	I
MRI_2	29.24	28.82	28.96	III	I	II
MRI_3	24.55	24.22	24.42	III	I	II
MRI_4	25.93	25.81	25.20	III	II	I
MRI_5	24.02	23.92	23.78	III	II	I
Overall Warp Ranking (Across subj.)	III	II	I	Legend: I = Best		
				II = Medium		
				III= Worst		

the information content of the data [Mallat, 1989], see Section 2.2.2.

In regard to the second question, about the advantages of using wavelet analysis, we present the same study done on the raw (structural) MRI volumes in the time domain (image space), Table 9. The overall rankings of warp performance given in Tables 7 and 9 are significantly different. For example, the image space analysis indicates an overall preference to the Affine warp ($Warp3$), while the wavelet space study selects the MI ($Warp2$) warp as the best. Other major differences can be identified by examining the planar representations of these tables displayed on Figures 34 and 37.

Our wavelet-based warp evaluation appears to be independent of the choice of the wavelet representation basis. To show this we repeated the wavelet analysis on the structural MRI images replacing the $Daub20$ (used to obtain the results in Tables 7 and 8) by the $Daub4$ wavelet filter. Even though these two filter banks induce wavelet bases of the same (Daubechies wavelets) family the corresponding wavelet representations of the signals

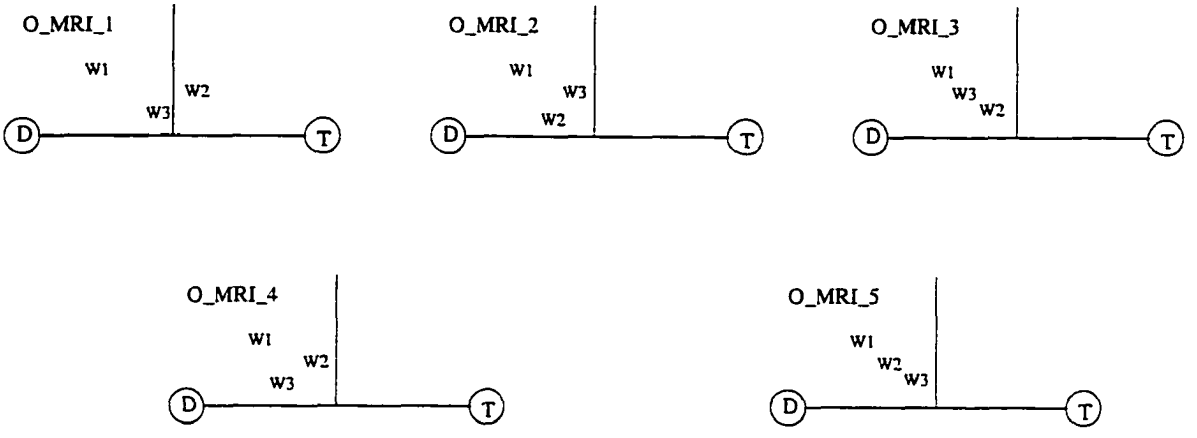


Figure 37. Planar representation of the quantitative warp evaluation on the raw data in image space.

are profoundly different (Daubechies [1988]). As Table 10 shows, the *Daub*4 wavelet analysis produced results in a very good agreement with the *Daub*20 study (Tables 7 and 8). The very small variations are probably due to computational errors, because the magnitudes of the wavelet coefficients of the *Daub*4 study are about 10 times these of the *Daub*20 representation. Also, some of the differences in the ranking can be explained by global uniform shifts of the positioning of the warped volumes from the left to the right side of the "mid-line", Figures 34 and 38.

6. Discussion

In this first chapter we described how discrete mathematical techniques for the analysis and synthesis of signals could be used for quantitative examination and comparison of medical images. First, we proposed a method for *quantitative* (numeric) estimation of image similarities. Our model transforms an "image matching" problem from the signal (physical) domain to

Table 10. Quantitative warp evaluation on the structural MRI data in wavelet space (*Daub4*).

Warping Scheme / DATA	Warp_1 (LS)	Warp_2 (MI)	Warp_3 (Affine)	Warp Ranking		
				Wapr_1	Warp_2	Warp_3
MRI_1	160.9	160.3	162.3	II	I	III
MRI_2	175.7	172.4	174.8	III	I	II
MRI_3	135.9	133.6	138.7	II	I	III
MRI_4	148.0	147.5	144.8	III	II	I
MRI_5	128.0	126.8	127.8	III	I	II
Overall Warp Ranking (Across subj.)	III	I	II	Legend: I = Best		
				II = Medium		
				III= Worst		

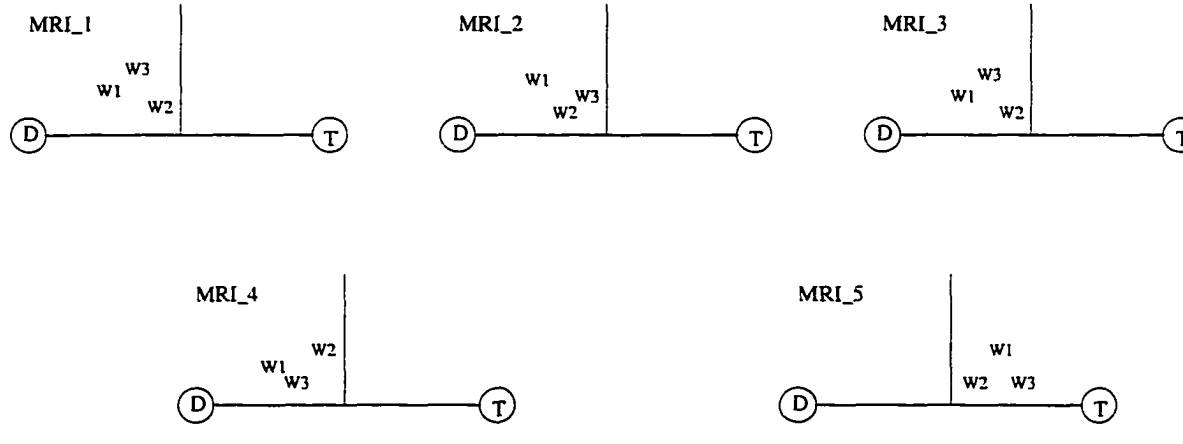


Figure 38. Planar representation of the quantitative warp evaluation on the MRI data in wavelet space (*Daub4*).

another (transform) domain, where we simplify and solve the problem, and pull the solution back into the initial signal domain.

Second, we use the DFT as a resolution enhancement and image mag-

nification device. We show by examples that the image-blow-up technique induced by the (resolution-independent) fractal transform produces better results than the most popular methods of *bi-linear* interpolation.

Third, we believe that the metrics we define on subsets of the transform space (S_2) are useful in comparing different warps and warping algorithms. We propose goodness of warp criteria that are fast, automated and do not require manual determination of anatomical landmarks and other fiducial points, curves or surfaces.

CHAPTER II

SUB-VOLUME THRESHOLDING TECHNIQUE FOR ANALYZING FUNCTIONAL HUMAN BRAIN DATA

In this chapter we propose a new method for determining an optimal threshold value, t_o, for human brain functional images (PET, fMRI) representing the difference between baseline and activation (stimulus) conditions. This **sub-image** technique is applicable for a single difference image as well as for multiple images. The implementation of this algorithm is straight-forward using the formulas we derive in Section 3 for estimating different variances. We have tested this method on human brain functional data (PET).

The main purpose of approximating the ultimate intensity threshold value is to be able to simultaneously denoise the data and determine which areas of the brain light up under a stimulation condition. Our test is more conservative than the commonly used $T - test$ (done on difference of averaged images), but less conservative than the Bonferroni's procedure test. An important advantage of our method is the low computational complexity. The performance of this test relative to the novel approach of Keith Worsley [1994], which uses the expectation of the Euler Characteristic on excursion sets, is not discussed in this chapter. Nor is the performance of our method compared to SPM (Statistical Parametric Mapping, Friston *et al.* [1991]), because we are mainly concerned with single subject (activation versus baseline) studies.

In functional imaging there are at least five major sources of error in the variance estimates [Friston *et al.*, 1990]. (Signal variance estimates are the

foundation of the analysis for determining the statistically significant changes of metabolic activity.) Morphological (brain positioning) spatial differences between the activation and the baseline scans; Second, using an inadequate statistical model; Third, differences in global activity in various regions of the brain; Fourth, inter-subject density variability effects (for multiple-subject studies); and fifth the poor resolution of the imaging equipment (Friston *et al.*, 1990).

In the sub-volume thresholding (SVT) technique the above potential errors are accounted for as follows: The probabilistically defined regions of interest (ROI) would control for the small local morphological differences of the activation and baseline images, after a rigid-body, affine or non-affine warping is performed. In the examples we present the more global morphological registration is achieved by a polynomial warping technique (Woods *et al.*, 1992, AIR). Modeling different ROI's as separate stationary random fields avoids the problem of non-uniform global activity within the brain. Of course, the choice of ROI's, defined as probabilistic (cloud-like) atlases, depends on the particular functional study. For single subject studies there is no across-subject variability. For multiple-subject studies, or for subject-to-group or group-to-group comparisons, one could invoke an approach similar to the block design of SPM's (Statistical Parametric Mapping), to account for the systematic inter-subject differences (Friston *et al.*, 1991).

The SVT method capitalizes on the fact that it requires no multiple scans of the same or different subject(s), and avoids the noise caused by inter-subject-variability. Moreover, it allows the incorporation of prior anatomical information within the process of determining appropriate threshold value(s) for sub regions of the brain. The correction factors for the necessary variance estimates are expressed in closed form as functions of the spatial auto-

correlation Gaussian model for the case of rectangular structural image partitioning. More general anatomical and probabilistic type structural image segmentations, and the stochastic approximations of the corresponding correction factors are also discussed .

In Section 1, we explain the main ideas underlining our "sub-image" thresholding method. Formulas for the required estimates of variances are derived in Section 1.2. A simple example involving a square-type partitioning of the domain of a 2D (PET) difference image is presented in Section 1.3. A family of useful admissible covariance models induced by a class of continuous functionals are presented in Section 2. Such covariograms appear naturally in our (spatial) density auto-correlation models, independent of the adopted density distribution (Z, T, F, χ^2 etc. fields).

In Section 2, we discuss classes of permissible covariograms studied by Christacos [1984], Matern [1986], Cressie [1991] and others. We prove that the class of continuous functions we use in our SVT model induce valid covariance functionals. Finally, in Section 3, we presents a number of examples illustrating the use of the SVT methodology.

1. The Sub-Volume Thresholding Technique

1.1 Foundations of the Statistical Analysis

Suppose X^{act} and X^{rest} represent the signals (functional images) of a subject(s) under baseline (rest) and stimulus (activation) conditions ($X^{act} = X^{act}_{(i,j,k)}$, $X^{rest} = X^{rest}_{(i,j,k)}$) on a 3D (or 2D) grid. Let $D = X^{act} - X^{rest}$, ($D = D_{(i,j,k)}$). Suppose also that the image D is partitioned according to some prior anatomically relevant basis. We think of D as being a disjoint union of images, $D = \bigcup_{m=1}^{M} D_m$, Figure 39. Observe that, depending upon the particular study,

the partitioning sub-images may be topologically connected or disconnected. The latter case could be applicable for studying regions only functionally connected. The reader is encouraged to think of the domains of the partitioning sub-images as being rectangles or parallelepipeds (in 2D and 3D, respectively). The first example we present in Section 1.3 involves one such (square) sub-image.

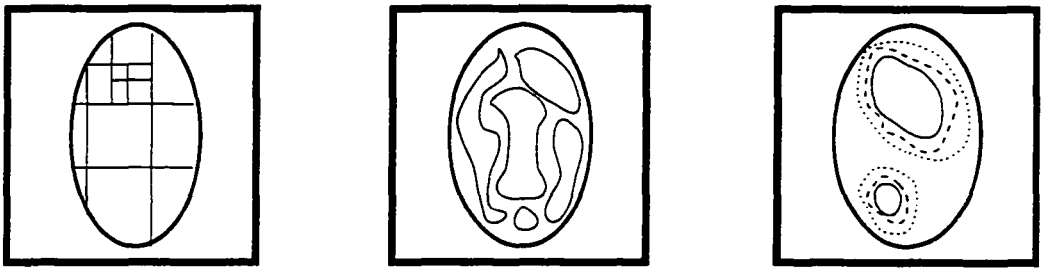


Figure 39. Geometric (left), Specific-anatomic or Anatomic-average (middle) and Probabilistic (right) partitioning schemes.

We will now describe a technique that determines whether a significant activation occurs within each sub-image D_m, and if so, locates the activation sites (voxels).

Let us concentrate on one sub-image, D_m, and think of it as a separate image. It is well-known that neighboring voxel intensities are highly correlated due to imperfect resolution of the imaging equipment, noise effects and the physiological nature of brain activation. We assume that the standard deviation of the Gaussian (covariance) smoothing kernel is known. Let ρ_I^2 be its variance, and $\rho = 2\sqrt{\rho_I^2}$ (thus two voxels farther apart than ρ are essentially uncorrelated).

A reasonable estimate[1] of the variance of the image D_m, $\sigma^2_{D_m}$, is the sub-sample variance of a random collection of voxels (I) within the domain of D_m

[1] The common "hat" notation, $\hat{}$, is used for estimated quantities.

that are far enough apart from each other.

$$\widehat{\sigma^2_{D_m}} = \frac{1}{|I|} \sum_{(i,j,k)\in I} (D_m(i,j,k) - \overline{D_m})^2,$$

where

$$\overline{D_m} = \frac{1}{|I|} \sum_{(i,j,k)\in I} D_m(i,j,k)$$

Under normal assumptions, once we have an estimate for $\sigma^2_{D_m}$, we do voxel-wise $Z - tests$

$$Z_{(i,j,k)} = \frac{D_m(i,j,k)}{\widehat{\sigma_{D_m}}}$$

to determine the location of the statistically significant sites of activation within D_m. Another measure of location activation (longitudinal, across-subjects) can be obtained by using multiple (registered) difference images $\{D^l_m\}^L_{l=1}$. Then for each voxel we do a T-test

$$T(i,j,k) = \frac{\overline{D'_m(i,j,k)}}{\sqrt{\frac{1}{L}\sum_{l=1}^{L} \frac{\left(D^l_m(i,j,k) - \overline{D'_m(i,j,k)}\right)^2}{L-1}}}$$

where $\overline{D'_m(i,j,k)} = \frac{1}{L}\sum_{l=1}^{L} D^l_m(i,j,k)$ is the across-subject average at $x = (i,j,k)$. One could also find meridianal (across-voxels, spatially) estimates of signal variances. Hybrid (longitudinal-meridianal) methods can also be employed.

Because of the large number of tests (number of voxels within a search region may be larger than 2^{18}) we will correct for the increasing false-positive test-error by testing at a significance level $\alpha_o = \frac{\alpha}{|I|}$, where $|I|$ is the approximate number of voxels (within the search region) that are uncorrelated with $(1-\alpha)100\%$ confidence. Typically, the initial significance level $\alpha = 0.05$. This hypothesis testing is less conservative by the well-known Bonferroni correction procedure.

Our first task, however, is to find out if there is a need to search for activation inside D_m. The partitioning of the original image into a disjoint union of sub-images will be done according to a scheme based on a priori anatomical (or even functional) information.

One of the main ideas of this method centers around identifying the sub-images D_m in which activation occurs with high confidence. These D_m's are going to be the only ones that we will subsequently search through voxel-by-voxel. We first begin by estimating the standard deviation of the sample average

$$\overline{D_m} = \frac{1}{n_{tot}} \sum_{(i,j,k) \in Dom(D_m)} D_m(i,j,k),$$

where n_{tot} is the total number of pixels in the domain of the sub-image D_m. Then using the standard error of $\overline{D_m}$ we will test the sub-image D_m, as a whole entity, for activation.

There are two fundamental assumptions we make in our model. The first one is that the neighboring sites (voxels) have Gaussian auto-correlation depending on the distance between them. The second implicit assumption is that the intensities at every voxel are normally distributed with mean zero and some unknown variance. Both of these hypotheses are reasonable and we now proceed to show this theoretically, using the physical properties of the imaging process, and empirically, using plots of real PET data.

If we place a single point-source of radioactive isotope in the center of a PET camera the image we obtain looks like a smeared blob, the result of a low-pass filter processing [Worsley, 1996]. Figure 40, displays the initial (real) image being scanned, on the left, and a side-view of the observed data, on the right. Indeed, the smoothing kernel has a bell-shape and can be modeled by a 2D normal distribution. In the Appendix, we describe

the fundamentals of the PET imaging technique, which we use to motivate these two reasonable assumptions. Briefly, there are two main reasons for observing (Gaussian) smooth PET images. The first one is the nature and the physiology of brain activation - blood flow changes occur smoothly and homogeneously. The second reason is the stochastic nature of the path of the positively charged β particles, (from their emission from the nucleus to their collusion with negatively charged electrons) and the attenuation effects causing close voxels to have highly positively correlated intensities. Coupling every detector, in the PET scanner, with several other detectors in a neighborhood of its 180°-opposite also introduces a distance dependent auto-correlation function similar to a Gaussian.

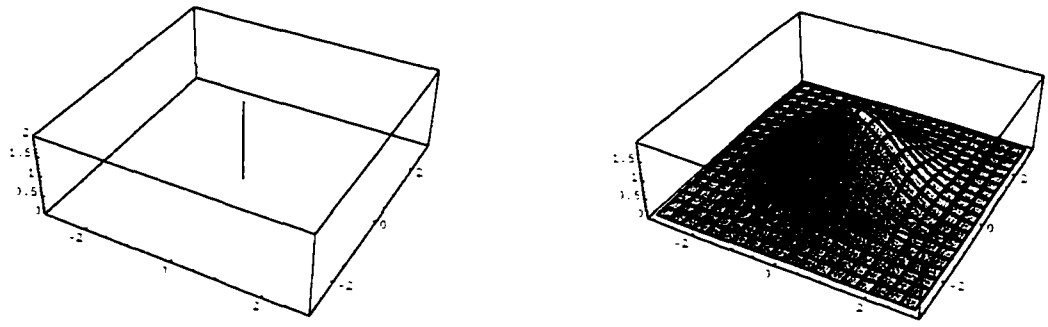


Figure 40. Gaussian voxel intensity correlation.
Point-source isotope data (left), observed image (right).

To explain the rationale behind the assumption for normal distribution of the voxel-intensities we again refer to the physics of the PET imaging technique, see the Appendix. A PET scan is constructed by detecting, comparing (times/places of arrival) and counting dual-photons emitted in the process of proton-electron annihilations. Photon strikes can be regarded as random arrivals, and modeled as a discrete Poisson process. Because of the large scale of this stochastic process its distribution can be approximated by a Gaussian (of mean zero, and some variance). Empirically, we demonstrate

the normal structure of the voxel intensities by taking 70 randomly selected intensities (that are far enough from each other and are not significantly correlated) of a difference image. Figure 41 shows the values of the differences, on the left, and the quantiles of a normal distribution (having the sample mean and variance of the difference data), on the right. The almost linear relation of the data and the normal quantiles yields that the sample was drawn from a (unknown) distribution closely related to normal.

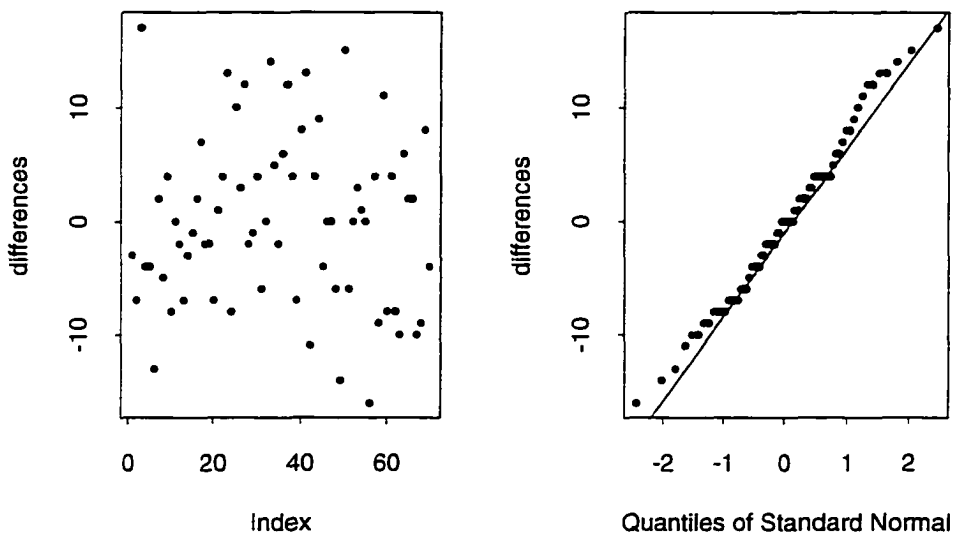


Figure 41. Normal nature of voxel-intensities.
70 randomly chosen differences (left), sample/normal quantiles plot (right).

1.2 Estimates of Variances

For simplicity of notation we will be suppressing the subindex m, and regard D_m as a whole new image, D. Assume that our data is a *stationary Gaussian* n-dimensional random field D_x [Adler, 1981]. Then D has constant (across-voxel) mean $E(D_x) = \mu = const$, for all $x \in R^n$, and and a spatial autocorrelation function of the form $Cov(D_{x_1}, D_{x_2}) = C(x_1, x_2) = c(x_1 - x_2)$, where $c : R^n \longrightarrow R$. Let $x_1 = (i_1, j_1, k_1)$, $x_2 = (i_2, j_2, k_2)$ and $d(x_1, x_2)$ be the l_1 distance on R^3. Then suppose $Cov(D_{x_1}, D_{x_2}) = \widehat{\sigma_D^2} \rho^{d(x_1, x_2)}$, where as before ρ is a measure

of the smoothing (noise) kernel. This *Covariogram* is in fact valid; that is, it is positive definite and underlined by a legitimate Gaussian probability (see Section 2). If the domain of D is a cube (square in 2D) of size n, then the total number of voxels in $Dom(D)$ is $n_{tot} = n^3$ ($n_{tot} = n^2$ in 2D) and

$$\widehat{\sigma^2_{\overline{D}}} = Var(\overline{D}) = Var\left(\frac{1}{n_{tot}} \sum_{(i,j,k)\in Dom(D)} D_{(i,j,k)}\right) =$$

$$= \frac{1}{n^2_{tot}} \sum_{x_1\in Dom(D)} \sum_{x_2\in Dom(D)} Cov(D_{x_1}, D_{x_2}) =$$

$$= \frac{1}{n^2_{tot}} \left(n_{tot}\widehat{\sigma^2_D} + \sum_{x_1,x_2\in Dom(D), x_1\neq x_2} \widehat{\sigma^2_D}\rho^{d(x_1,x_2)} \right)$$

Define

$$A = \sum_{x_1,x_2\in Dom(D), x_1\neq x_2} \widehat{\sigma^2_D}\rho^{d(x_1,x_2)}$$

We derive an explicit closed form for $Var(\overline{D})$ in 2D and state its extension to 3D.

Claim 2.1.

$$\sum_{i\neq j, 1\leq i,j\leq n} \rho^{|j-i|} = 2 \sum_{1\leq i<j\leq n} \rho^{|j-i|} = 2\left(\frac{\rho(n-1)}{1-\rho} - \frac{\rho^2}{(1-\rho)^2}\left(1-\rho^{(n-1)}\right)\right)$$

Proof:

$$\sum_{1\leq i<j\leq n} \rho^{|j-i|} = \sum_{k=1}^{n-1} \rho^k P_k$$

where P_k is the number of pixels $\{(i,j):\ 1\leq i<j\leq n,\ j-i=k\}$. So,

$$P_k = |\{i:\ 1\leq i<i+k\leq n\}| = (n-k)$$

$$\sum_{1\leq i<j\leq n} \rho^{|j-i|} = \sum_{k=1}^{n-1} \rho^k(n-k) = n\sum_{k=1}^{n-1}\rho^k - \sum_{k=1}^{n-1} k\rho^k =$$

$$= \rho n \sum_{k=0}^{n-2}\rho^k - \rho\sum_{k=1}^{n-1} k\rho^{k-1} =$$

$$= \rho n\frac{1-\rho^{n-1}}{1-\rho} - \rho\sum_{k=1}^{n-1} k\rho^{k-1} = \rho n\frac{1-\rho^{n-1}}{1-\rho} - \rho\frac{(n-1)\rho^n - n\rho^{n-1}+1}{(1-\rho)^2} =$$

$$= \frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2}.$$

Δ

Claim 2.2. *In 2D, if* $x_1 = (i_1, j_1)$ *and* $x_2 = (i_2, j_2)$ *then*

$$\sum_{i_1 \neq i_2, j_1 \neq j_2} \rho^{d(x_1,x_2)} = 4\left(\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2}\right)^2$$

Proof:

$$\sum_{\{i_1 \neq i_2, 1 \leq i_1, i_2 \leq n\}} \sum_{\{j_1 \neq j_2, 1 \leq j_1, j_2 \leq n\}} \rho^{d(x_1,x_2)} =$$

$$= \sum_{\{i_1 \neq i_2, 1 \leq i_1, i_2 \leq n\}} \sum_{\{j_1 \neq j_2, 1 \leq j_1, j_2 \leq n\}} \rho^{|i_1 - i_2| + |j_1 - j_2|} =$$

$$= \left(\sum_{i_1 \neq i_2, 1 \leq i_1, i_2 \leq n} \rho^{|i_1 - i_2|}\right)\left(\sum_{j_1 \neq j_2, 1 \leq j_1, j_2 \leq n} \rho^{|j_1 - j_2|}\right) =$$

$$= \left(2 \sum_{1 \leq i_1 < i_2 \leq n} \rho^{i_2 - i_1}\right)\left(2 \sum_{1 \leq j_1 < j_2 \leq n} \rho^{j_2 - j_1}\right) = 4\left(\sum_{1 \leq i_1 < i_2 \leq n} \rho^{i_2 - i_1}\right)^2 =$$

$$= 4\left[\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2}\right]^2.$$

The last equality follows from Claim 2.1.

Δ

Claim 2.3. *In 2D,* A_2 $(= A)$ *can be expressed as:*

$$A_2 = 4\widehat{\sigma_D^2}\left[\left(\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2}\right)^2 + n\left(\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2}\right)\right]$$

Proof: We use the following disjoint decomposition of the index set $\Omega = \{\{i_1 \neq i_2\} \bigcup \{j_1 \neq j_2\}\}$

$$\Omega = \Omega_1 \bigcup \Omega_2 \bigcup \Omega_3$$

where $\Omega_1 = \{\{i_1 \neq i_2\} \bigcap \{j_1 \neq j_2\}\}$, $\Omega_2 = \{\{i_1 \neq i_2\} \bigcap \{j_1 = j_2\}\}$, $\Omega_3 = \{\{i_1 = i_2\} \bigcap \{j_1 \neq j_2\}\}$, and

$$\Omega, \quad \Omega_i \subseteq \{(i,j);\ 1 \leq i,j \leq n\}$$

$$A_2 = \widehat{\sigma_D^2} \sum_{x_1 \neq x_2} \rho^{d(x_1,x_2)} = \widehat{\sigma_D^2} \sum_{\Omega} \rho^{d(x_1,x_2)} =$$

$$= \widehat{\sigma_D^2} \left[\sum_{\Omega_1} \rho^{d(x_1,x_2)} + \sum_{\Omega_2} \rho^{d(x_1,x_2)} + \sum_{\Omega_3} \rho^{d(x_1,x_2)} \right] =$$

$$= \widehat{\sigma_D^2} \left[4 \left(\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2} \right)^2 + n \sum_{i_1 \neq i_2} \rho^{|i_1 - i_2|} + n \sum_{j_1 \neq j_2} \rho^{|j_1 - j_2|} \right]$$

Rearranging the terms and using Claim 2.1, we get the above form for A_2.

Δ

Then the estimate of the variance of the average $\overline{D}$, $\widehat{\sigma_{\overline{D}}^2}$, can be expressed as a function of ρ:

$$\phi(\rho) = \widehat{\sigma_{\overline{D}}^2} = \frac{1}{n_{tot}^2} \left[n_{tot} \widehat{\sigma_D^2} + A_2 \right] =$$

$$= \frac{\widehat{\sigma_D^2}}{n_{tot}^2} \left[n_{tot} + 4 \left(\left(\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2} \right)^2 + n \right) \right] \tag{14}$$

Note: These formulas can be generalized to include rectangular partitionings ($D_{m \times n}$) and to 3D using the 7-term disjoint decomposition for

$$\Omega = \left\{ \{i_1 \neq i_2\} \bigcup \{j_1 \neq j_2\} \bigcup \{k_1 \neq k_2\} \right\} = \Omega_1 \bigcup \Omega_2 \bigcup \Omega_3 \bigcup \Omega_4 \bigcup \Omega_5 \bigcup \Omega_6 \bigcup \Omega_7$$

where

$$\Omega_1 = \{i_1 \neq i_2\} \bigcap \{j_1 \neq j_2\} \bigcap \{k_1 \neq k_2\} \quad \Omega_2 = \{i_1 \neq i_2\} \bigcap \{j_1 \neq j_2\} \bigcap \{k_1 = k_2\}$$

$$\Omega_3 = \{i_1 \neq i_2\} \bigcap \{j_1 = j_2\} \bigcap \{k_1 \neq k_2\} \quad \Omega_4 = \{i_1 = i_2\} \bigcap \{j_1 \neq j_2\} \bigcap \{k_1 \neq k_2\}$$

$$\Omega_5 = \{i_1 \neq i_2\} \bigcap \{j_1 = j_2\} \bigcap \{k_1 = k_2\} \quad \Omega_6 = \{i_1 = i_2\} \bigcap \{j_1 \neq j_2\} \bigcap \{k_1 = k_2\}$$

$$\Omega_7 = \{i_1 = i_2\} \bigcap \{j_1 = j_2\} \bigcap \{k_1 \neq k_2\}$$

The explicit formula (3D case, for a cubical search region) for $\widehat{\sigma^2_{\overline{D}}} = \phi(\rho)$ is:

$$\phi(\rho) = \widehat{\sigma^2_{\overline{D}}} = \frac{1}{n^2_{tot}} \left[n_{tot} \widehat{\sigma^2_D} + A_3 \right] \tag{15}$$

where,

$$A_3 = \widehat{\sigma^2_D} \left(2^3 \left[\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2} \right]^3 + 3n2^2 \left[\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2} \right]^2 + \right.$$

$$\left. + 3n^2 2 \left[\frac{\rho(n-1)}{1-\rho} - \rho^2 \frac{1-\rho^{n-1}}{(1-\rho)^2} \right] \right)$$

Because $D_x \sim N(0, \sigma^2_D)$, $\forall x$ and $\overline{D} \sim N(0, \sigma^2_{\overline{D}})$ we standardize $\overline{D}$ to determine, using a $Z - test$, whether activation occurs within the whole (sub-)image D. As a result only if the test statistic (under the null hypothesis, $H_o : \mu_D = 0$)

$$Z = \frac{\overline{D}}{\widehat{\sigma_{\overline{D}}}}$$

is large enough will we search through D voxel-by-voxel to determine the location(s) of the activation site(s). For this we use T or $Z - tests$ as we described previously.

The above technique for determining the significant regions of activation allows variable thresholding of functional data on different anatomical regions of the brain. In general, the activation sites found by this method may not be present on a simple global T statistic image, nor will all of the (uniform) T statistic voxels appear among the activation sites determined by the SIT (Sub-Image Thresholding) technique. In addition, if k multiple scans are available (no replicates) then an estimate of the variance of the mean of the data is obtained by $\widehat{\sigma^2} = \frac{1}{k-1} \sum_{i=1}^{k} (\mu_i - \mu)^2$, where $\mu = \frac{1}{k} \sum_{i=1}^{k} \mu_i$ is an estimate of

the average of the means μ_i. This, in general, yields tests with low degrees of freedom ($df \leq 15$) and often hybrid methods, where the variance is pooled spatially and across-subjects (meridianally and longitudinally), are employed to increase the degrees of freedom and the accuracy of the tests, Friston [1991].

For more complex regions such nice closed mathematical expressions of the variance estimates are not available. In which case one writes

$$\widehat{\sigma_{\overline{D}}}^2 = Var(\overline{D}) = Var\left(\frac{1}{n_{tot}} \sum_{(i,j,k)\in Dom(D)} D_{(i,j,k)}\right) =$$

$$= \frac{1}{n_{tot}^2} \sum_{x_1\in Dom(D)} \sum_{x_2\in Dom(D)} Cov(D_{x_1}, D_{x_2}) =$$

$$= \frac{1}{n_{tot}^2} \left(\sum_{x_1,x_2\in Dom(D)} \widehat{\sigma_D^2} \rho^{d(x_1,x_2)} \right) =$$

$$= \frac{1}{n_{tot}^2} \left(\sum_{k=0}^{diam(D)} \widehat{\sigma_D^2} \rho^k P_k \right) = \widehat{\sigma_D^2} \left(\frac{1}{n_{tot}^2} \sum_{k=0}^{diam(D)} \rho^k P_k \right)$$

where $P_k = |\{(x_1, x_2) : x_1, x_2 \in Dom(D),\ d(x_1, x_2) = k\}|$, and $diam(D) = \max\{d(x_1, x_2) : x_1, x_2 \in Dom(D)\}$ is the usual diameter of the sub-volume D. There seems to be no simple closed form for the factors P_k for an arbitrary region D. Also, for computational purposes it is not feasible to do an exhaustive search throughout the domain of D. In our tests we have used stochastic approximations of $P_k, \forall k$ (under l_1 distance) that yield in fact stable estimates. We define the expressions

$$CF = \frac{1}{n_{tot}^2} \sum_{k=0}^{diam(D)} \rho^k P_k$$

as Correction Factors. These are the scaling factors needed to estimate $\widehat{\sigma_{\overline{D}}}^2$, $\widehat{\sigma_{\overline{D}}} = \widehat{\sigma_D}\sqrt{CF}$. For the 3D probabilistic partitioning shown on Figure 44, the simulated correction factors for all ROI's seem to be uniformly bigger

than their exact counterparts, however, the errors are all within 1%. Table 11 contains the values of both types of estimates of the CF's for the 5 ROI's rescaled by a factor of 10^6. Here, the random search picked one out of 1,000 voxels. In addition, Table 11 contains (3^{rd} row) locally obtained exact estimates of CF. Because of the Gaussian auto-correlation model one may assume that local search could approximate well the exact correction factors. For each voxel v we used a search neighborhood $N(v, 20) = \{w : ||w - v||_{L_1} \leq 20\}$. For comparison, the magnitude of the Gaussian filter is about 6 voxels. Still the local estimates are about 2% lower than the exact values of CF, which leads to under-estimating the true variance and overestimating the Z test statistic and thus less conservative analysis.

Table 11. Stochastic estimates vs Exact values of the correction factors.

Probabilistic ROI / Methods for CF evaluation	Cerebellum	Frontal Lobe	Occipital Lobe	Parietal Lobe	Temporal Lobe
Stochastic Estim. of Corr. Factors (1/1000 pt)	665	274	773	446	533
Globally Obtained Exact values (Exhaustive Search)	661	272	766	443	529
Locally obtained Exact values (Exhaustive Search)	648	267	752	435	520

1.3 A simple 2D Example

Suppose we consider a 16×16 square sub-image of a 128×128 image. Then $n = 16$, $n_{tot} = 16^2$. Assume also the Gaussian smoothing kernel (related to the FWHM, Worsley [1994]) is determined by $\rho = \frac{\sqrt{2}}{2}$. Then an estimate of the standard deviation of the average, $\overline{D}$, is obtained by plugging in equation

(14)

$$\widehat{\sigma_{\overline{D}}} = \widehat{\sigma_D} \times 0.3357$$

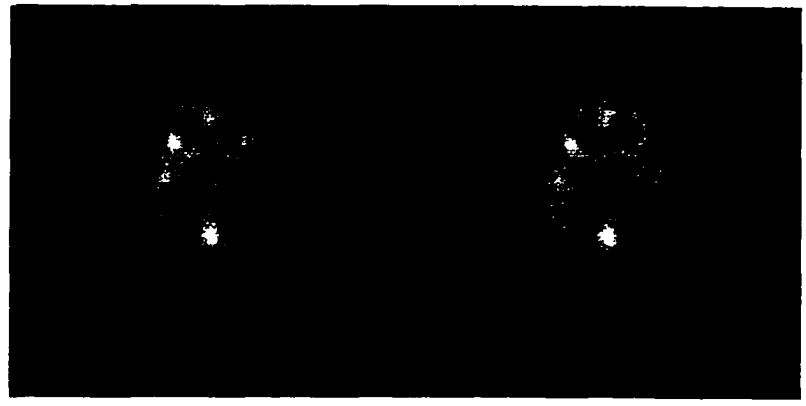


Figure 42. Original baseline (left) and stimulated (right) PET images.

Figure 42 depicts the original baseline and activated PET images and the left image in Figure 43 shows the difference image ($D = X^{act} - X^{rest}$). In this example we have concentrated our search on a part of the visual cortex (small boxed regions in Figure 43). This clearly illustrates that our sub-image testing technique is far more conservative than the uniform $T - test$ (middle, Figure 43).

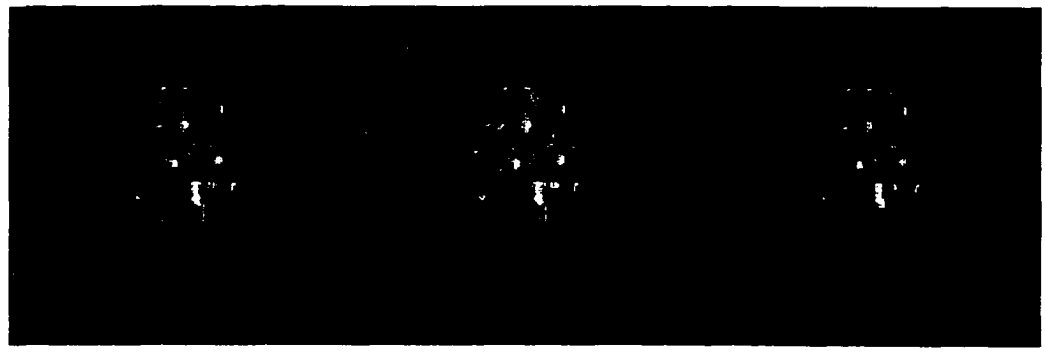


Figure 43. Difference image (left), Uniform $T - test$ image (middle), Sub-image test (right).

The main reason for the variation of the significance levels of the boxed

sub-images is the difference in the variance estimates. For the global T-test (middle, Figure 43) we have pooled the variance estimate over the whole image, while in the sub-image test (right, Figure 43) the local variance estimate was used to evaluate the T-statistic.

2. Validity of the Covariogram Model

We now proceed to show that the covariogram we adopted (Sections 1.1 and 1.2) and used is *permissible* (valid), that is, it is underlined by a legitimate probability model. In general, a continuous function $c(h) : R^n \longrightarrow R$ is an *admissible* covariance (covariogram) for a stationary random field D_x on R^n [Adler, 1981] if and only if $c(h)$ is non-negative definite[1], that is

$$\sum_{k=1}^{n}\sum_{l=1}^{n} \alpha_k \overline{\alpha_l} c(x_k - x_l) \geq 0$$

for all $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^t \in C^n$, where $c(x_k - x_l) = Cov(D_{x_k}, D_{x_l})$ and $E(D_x) = \mu$, $\forall x$ by stationarity.

Theorem 2.4. *A continuous function $c(h)$ is non-negative definite if and only if it can be expressed as the Fourier Transform of a non-negative bounded measure ϕ, that is*

$$c(h) = \int_{R^n} e^{2\pi i < w,h >} d\phi(w),$$

where $< w, h >= \sum_{k=1}^{n} w_k h_k$.

Proof: [Bochner, 1956]

Note: If both $c(h)$ and its Fourier transform $\widehat{c}(w)$ are in $L_1[R^n]$, i.e. both are measurable and have finite L_1 norms, then the above criterion is equivalent to saying that the Fourier transform of $c(h)$

$$\widehat{c}(w) = \int_{R^n} c(h) e^{-2\pi i < h,w >} dh$$

[1] The "bar" notation, $\bar{}$, indicates complex conjugation throughout this section.

is non-negative [Folland, 1984].

Proposition 2.5. *If $\|\bullet\|_1$ is the l_1 norm on R^3, ($\|h\|_1 = |h_1| + |h_2| + |h_3|$), then the function $c(h) = K\rho^{\|h\|_1}$, $0 < \rho < 1$, induces a valid covariance functional*

$$Cov(D_{x_1}, D_{x_2}) = C(x_1, x_2) = c(x_1 - x_2) = c(h) = K\rho^{\|h\|_1} = K\prod_{k=1}^{3}\rho^{|h_k|}$$

for any positive constant K (in our models we have used $K = \widehat{\sigma_D^2}$).

Proof: $FT(c) = \widehat{c}(w) = \int\int\int_{R^3} c(h)e^{-2\pi i<h,w>}dh =$

$$= K\left(\int_R \rho^{|h_1|}e^{-2\pi i h_1 w_1}dh_1\right)\left(\int_R \rho^{|h_2|}e^{-2\pi i h_2 w_2}dh_2\right)\left(\int_R \rho^{|h_3|}e^{-2\pi i h_3 w_3}dh_3\right) =$$

$$= K\prod_{k=1}^{3}\int_R \rho^{|h_k|}e^{-2\pi i h_k w_k}dh_k = K\prod_{k=1}^{3}\int_R e^{|h_k|\ln(\rho)}e^{-2\pi i h_k w_k}dh_k =$$

$$= 2K\prod_{k=1}^{3}\frac{a}{a^2 + b_k^2} \geq 0,$$

where $a = -\ln(\rho)$, and $b_k = 2\pi w_k$. Further, $\|\widehat{c}(w)\|_1 < \infty$. The last equality follows from the fact that an integral of an odd function on a symmetric interval is zero. Therefore, if a is a constant

$$\int_R e^{-|x|a}e^{-2\pi i x w}dx = \int_R e^{-|x|a}(\cos(2\pi x w) - i\sin(2\pi x w))dx =$$

$$= 2\int_0^{\infty} e^{-|x|a}\cos(2\pi x w)dx = 2\frac{a}{a^2 + (2\pi w)^2}$$

Δ

Observe, that similar approaches could be used to show that any l_p norm on R^n would induce an admissible covariogram model of the type $c(h) = K\rho^{\|h\|_p^p}$. In addition, having that $c(h)$ is valid on R^n implies that it is also valid on R^{n-k}, for $0 \leq k < n$, see Christacos, 1984.

Now we will prove a Lemma that is essentially the "if" part of the Bochner's theorem that would allow us to verify the validity of this covariance model $c(h) = K\rho^{\|h\|_p^p}$ in a slightly different (easier) way.

Lemma 2.6. *If the Fourier transform of an L_1 function, $c(h)$, is a non-negative L_1 function, $\widehat{c}(w)$, then $c(h)$ is non-negative definite.*

Proof: Since $c(h) \in L_1(R^n)$, the Fourier transform of c exists, $FT(c)(w) = \widehat{c}(w) = \int_{R^n} c(h)e^{-2\pi i<w,h>}dh$ and $c(h) = \int_{R^n} e^{2\pi i<w,h>}\widehat{c}(w)dw$. Therefore,

$$c(x_k - x_l) = \int_{R^n} e^{2\pi i(<w,x_k>-<w,x_l>)}\widehat{c}(w)dw$$

Expanding the quadratic form

$$\sum_{k=1}^{n}\sum_{l=1}^{n} a_k\overline{a_l}c(x_k - x_l) = \sum_{k=1}^{n}\sum_{l=1}^{n}\left(a_k\overline{a_l}\int_{R^n} e^{2\pi i(<w,x_k>-<w,x_l>)}\widehat{c}(w)dw\right) =$$

$$= \int_{R^n}\sum_{k=1}^{n}\sum_{l=1}^{n}(a_k e^{2\pi i<w,x_k>}\overline{a_l}e^{-2\pi i<w,x_l>}\widehat{c}(w))dw =$$

$$= \int_{R^n}\left|\sum_{l=1}^{n} a_l e^{2\pi i<w,x_l>}\right|^2 \widehat{c}(w)dw \geq 0$$

The last inequality follows from the assumption that the Fourier transform of $c(h)$ in non-negative.

Δ

Using Lemma 2.6, instead of the Bochner's theorem, and Proposition 2.5 we see that the covariogram induced by $c(h) = K\rho^{||h||_p^p}$ is permissible for any positive integer p.

3. Applications of the SVT Technique

We now discuss a few examples that use a probabilistic partitioning based on the space-filling atlases developed by Evans *et al.* [1996]. These stereotactic human brain atlases were produced using 53 normal (25-35 year-old) subjects and include nine segments (ROI's). We used the following 5: cerebellum, frontal lobe, occipital lobe, parietal lobe, and temporal lobe. Figure

44 shows sagittal, coronal and axial views of these color-coded structures. Every voxel (volume-element, equivalent to pixel in 2D) within an ROI has a probability map associated with it that gives the chance that this voxel lies within the ROI for an average "normal" brain. Brighter colors, within the ROI, indicate higher probabilities, and the colors dim toward the boundaries of the ROI's.

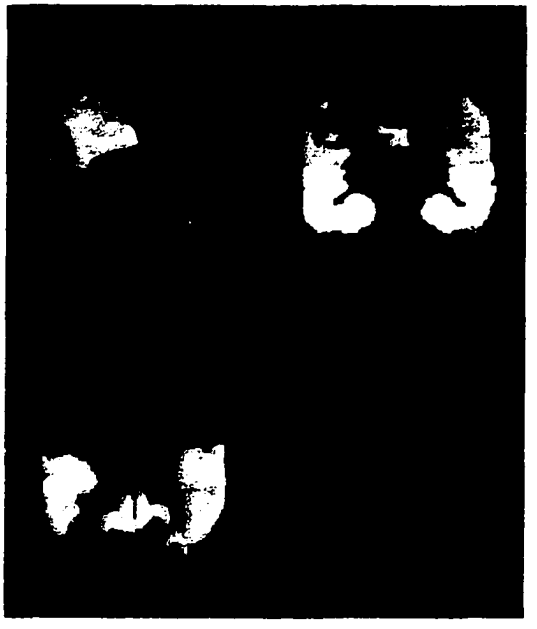


Figure 44. Five probabilistically defined ROI's.

The first example (ex. 1) represents a hypothetical situation. For a PET volume, Figure 45, the five ROI's were tested for statistical significance using hypothetical "prior" knowledge about the expected average activation. Throughout this section we use the short notations "c" - cerebellum; "f" - frontal; "o" - occipital; "p" - parietal; and "t" - temporal lobes. We have tested the following hypotheses:

$$H_o : \mu_c = 90 \quad H_o : \mu_f = 110 \quad H_o : \mu_o = 105 \quad H_o : \mu_p = 95 \quad H_o : \mu_t = 100$$

Modulo these prior averages, our study concludes, Table 12, that there is (global) statistically significant activation in all of the segments except the temporal lobe. The values of the Z statistic on the 5 different search regions are computed using the formula $\widehat{\sigma_{\overline{D}}} = \widehat{\sigma_D}\sqrt{CF}$, where CF are the "correction factors" for the ROI. The values of CF rescaled by a factor of 10^6 are listed in Table 11. Figures 46 and 47 illustrate the locations of the significant perfusion (brain activation significantly higher/lower than the mean for the ROI) within the globally statistically activated ROI's.

Table 12. Global sub-volume statistical analysis of a single PET volume (ex. 1).

Statistics & Tests / Probabilistic ROI	Mean		Standard Deviation	ROI Z statistic	Signif. of Atlas
	Null	Actual			
Cerebellum	90.00	87.21	40.80	2.18	Signif.
Flontal Lobe	110.00	113.11	46.61	2.80	Signif.
Occipital Lobe	105.00	123.34	49.21	15.37	Signif.
Parietal Lobe	95.00	109.20	47.36	5.82	Signif.
Temporal Lobe	100.00	99.15	52.44	0.50	Not Signif.

Our second example (ex. 2) involves pre- and post-treatment metabolic study. A subject was scanned twice, Figure 48. The first time under a drug (scopolamine) treatment and the second time without the drug treatment. The averages and the variances of the two volumes were then equalized. Normalization preprocessing steps were necessary because "blood-curves" for that study were not available. Bringing the volumes in the "same" image space is required by the fundamental assumption (frequently used in practice) that activation (metabolic activity) causes reallocation of CBF (cerebral

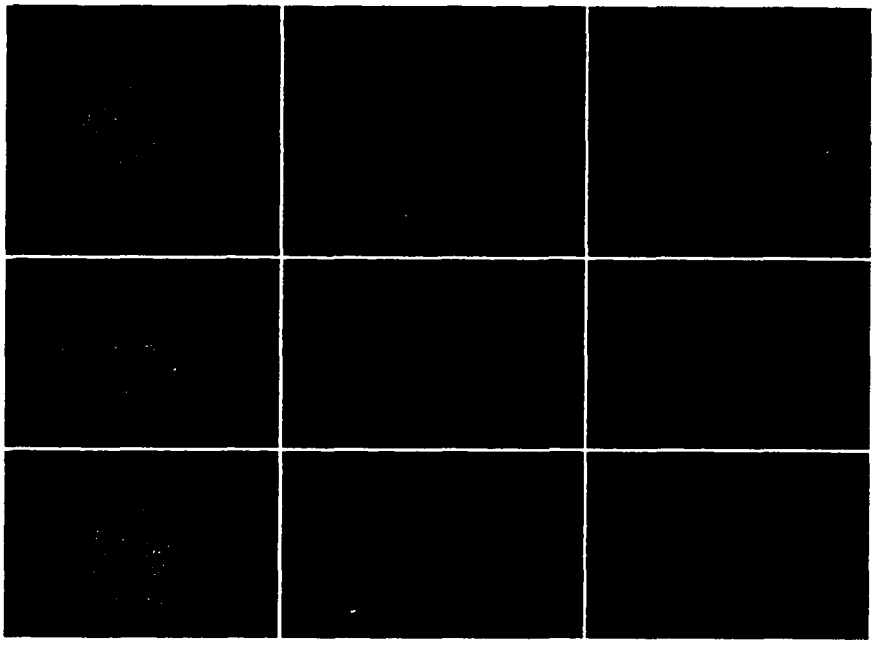


Figure 45. Registering the Functional data onto the average structural image (ex. 1).

blood flow), instead of increasing the global amount of CBF in the brain. The problem is to identify the regions of the brain that showed significant changes. Separate views of the positively and negatively statistically activated regions are shown on Figures 49 and 50. According to the SVT analysis, all ROI's exhibit significant metabolic perfusion except the occipital and parietal lobes, Table 13. Cerebellum and frontal lobe contain the most profound changes, which means that the drug targets these regions, Figures 49 and 50.

The last example (ex. 3) involves a motor-study. Its geared toward validating SVT testing for the well-known effects of motor studies. A subject is asked to trace a moving target once with his right hand and once with his left hand, Figure 51. Two volumetric PET data sets were obtained for the two paradigms and used to test the single-subject SVT for accuracy and

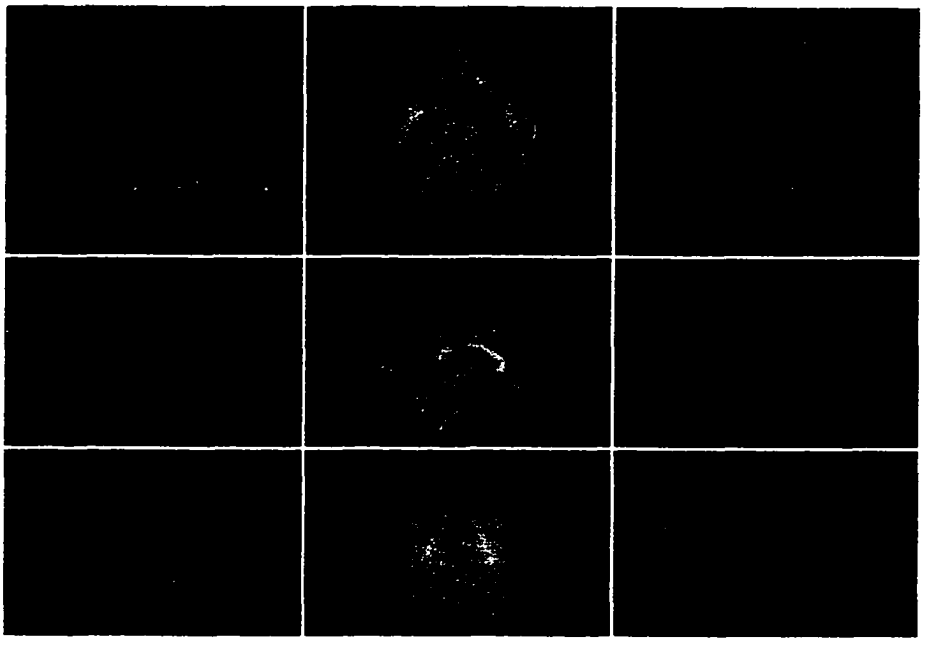


Figure 46. Statistically significant regions of activation, as determined by the sub-volume technique (ex. 1).

robustness. The density normalized volumes were also stereotactically (spatially) registered (AIR3.0, Woods [1992]) to the anatomical atlas associated with the Evans *et al.* [1996] probabilistic atlas, middle Figure 52.

It is well-known that motion stimuli activate neurons in the motor-cortex. Figures 53 and 54 display the statistically significant metabolic variations for the "right - left" and the "left - right" hand difference images, respectively. As expected, the left-hand study stimulated the right frontal and parietal lobes, and conversely the right-hand paradigm activated the sensory-motor cortex in the left frontal and parietal lobes. Moreover, our statistical analysis shows that frontal, occipital and parietal lobes are globally statistically significant and the temporal lobe is not (cerebellum was excluded from that study), Table 14.

For these single-subject studies we could not compare the SVT results

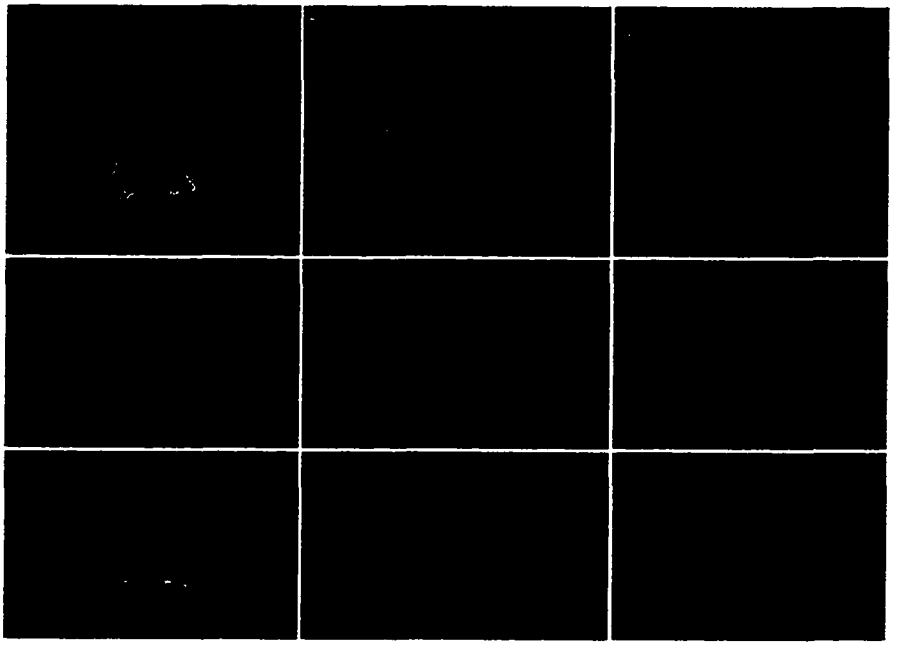


Figure 47. Statistically significant changes overlaid on the (Activation) functional volume (ex. 1).

Table 13. Sub-volume statistical tests on Activation vs Rest functional images (ex. 2).

Statistics & Tests / Probabilistic ROI	Mean (difference image)	Standard Deviation	ROI Z statistic	Significance of Atlas
Cerebellum	-6.43	22.23	9.23	Significant
Frontal Lobe	5.68	14.27	16.68	Significant
Occipital Lobe	0.66	18.36	1.47	Not Signif.
Parietal Lobe	0.20	14.60	0.27	Not Signif.
Temporal Lobe	7.35	15.44	14.68	Significant

with other multi-subject functional analysis tests (SPM, ANCOVA, Worsley's Euler characteristic etc.). However, we did evaluate the performance of

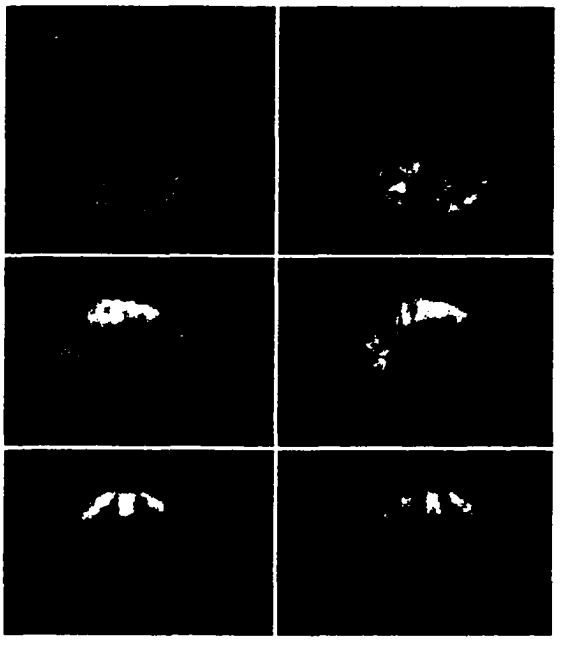


Figure 48. (Normalized) Activation (treatment) vs Rest functional images (ex. 2).

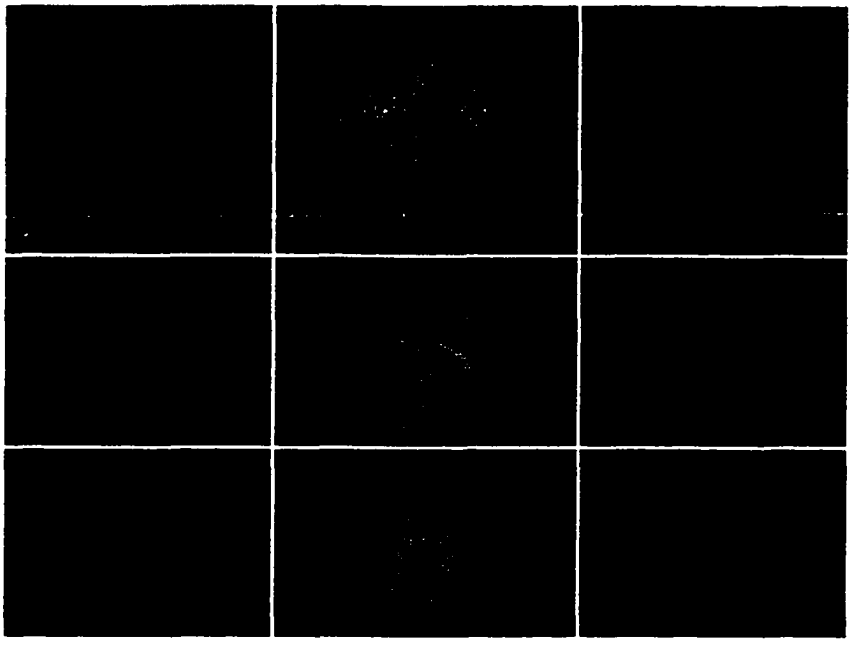


Figure 49. Positively activated significant metabolic changes (Active-Rest, ex. 2).

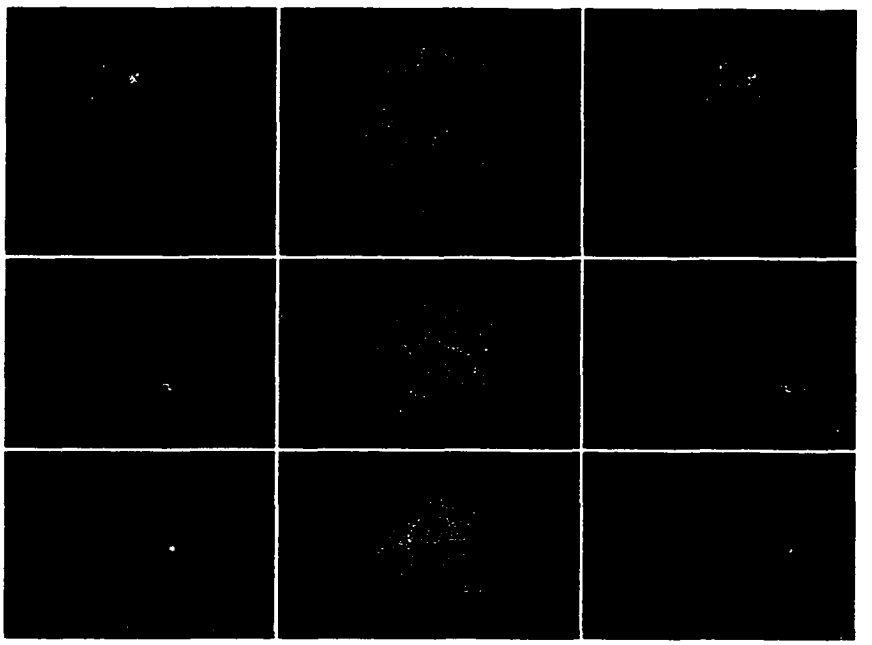


Figure 50. Negatively statistically significant (decrease) ROI's (Active-Rest, ex. 2).

the SVT test with a simple 97 percentile thresholding (SVT tests were done at 97% level, as well). Separate views of the positive and negative significant changes are shown on Figures 55 and 56. The two techniques are compared using the scopolamine treatment example (ex. 2). Interactive comparison of the data reveals some similarities and some differences between the two methods. One such significant mismatch is marked out on Figure 57. A relatively large region in the right temporal lobe appears only in the simple 97% thresholded image (middle column). The absence of this region of activation in the SVT image is the difference between the variance of the image over the temporal lobe and the variances of the other ROI's.

4. Discussion

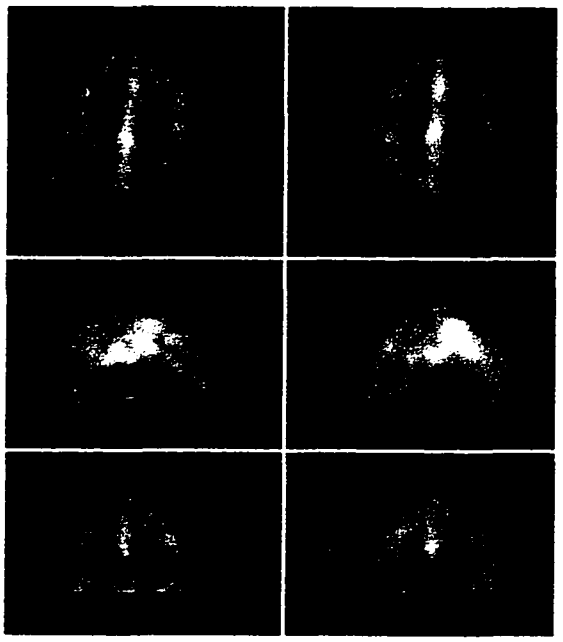


Figure 51. Motor study volumes: Right-hand stimulus (left), Left-hand stimulus (right). (ex. 3).

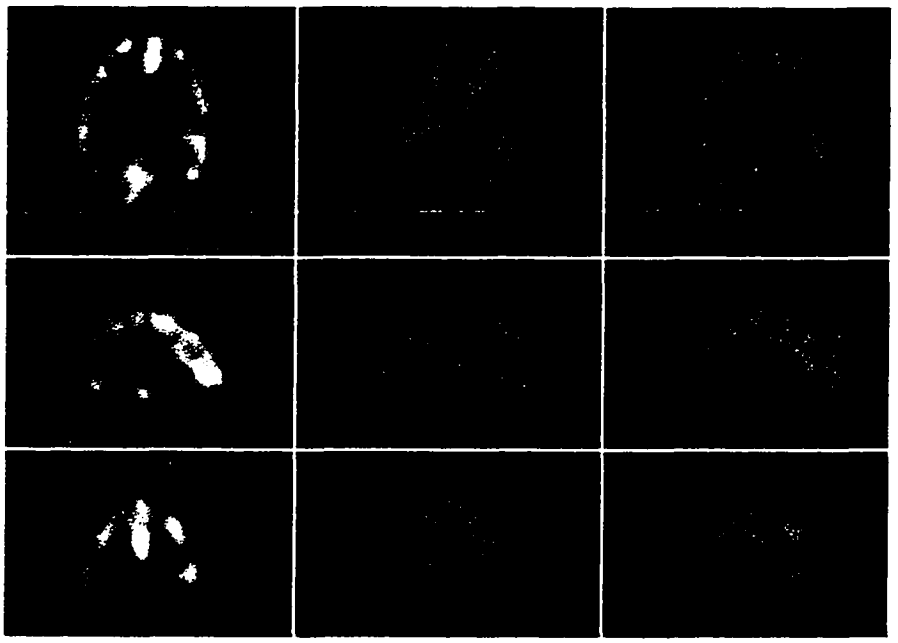


Figure 52. Registering the functional data to the anatomical atlas (ex. 3).

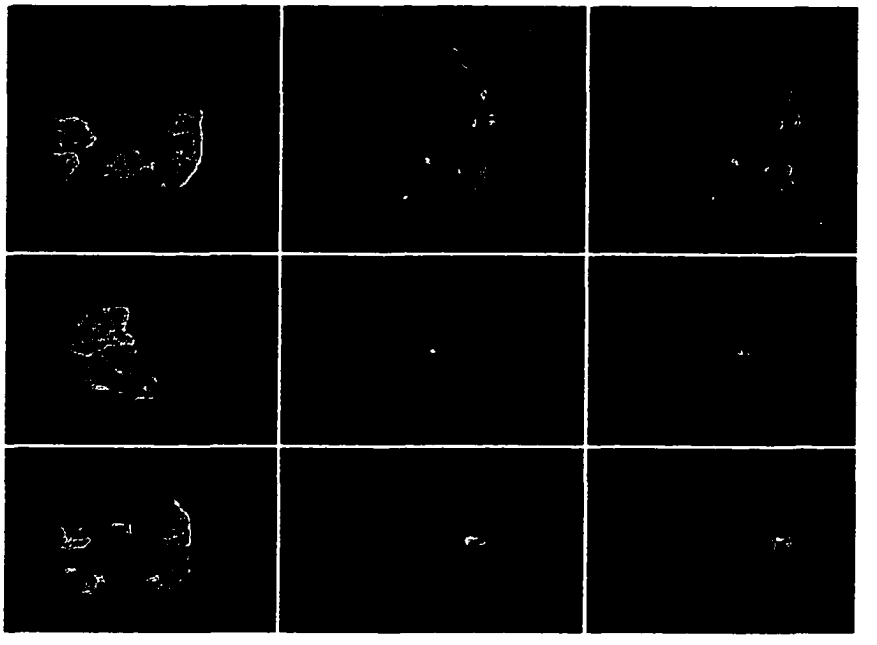


Figure 53. Positively statistically significant differences:
Right-Left hand motor study (ex. 3).

In this chapter we introduce and test a new sub-volume thresholding (SVT) technique for statistical analysis of single-subject functional data (PET, SPECT, fMRI). We spatially sub-divide the volumes into geometric, anatomical or probabilistic search regions based on different structural constraints on the data and various prior beliefs about the functional study.

Following this partitioning step, two types of statistical tests are applied. The first one, fundamentally dividing the SVT method from other techniques for statistical analysis of functional images, is aimed at determining the global significance of the functional data over each search region separately. Depending upon the topology of a sub-volume of interest we determine an estimate for the variance of the average of the signal (difference image). These estimates are then used to assess the global significance levels of the ROI's.

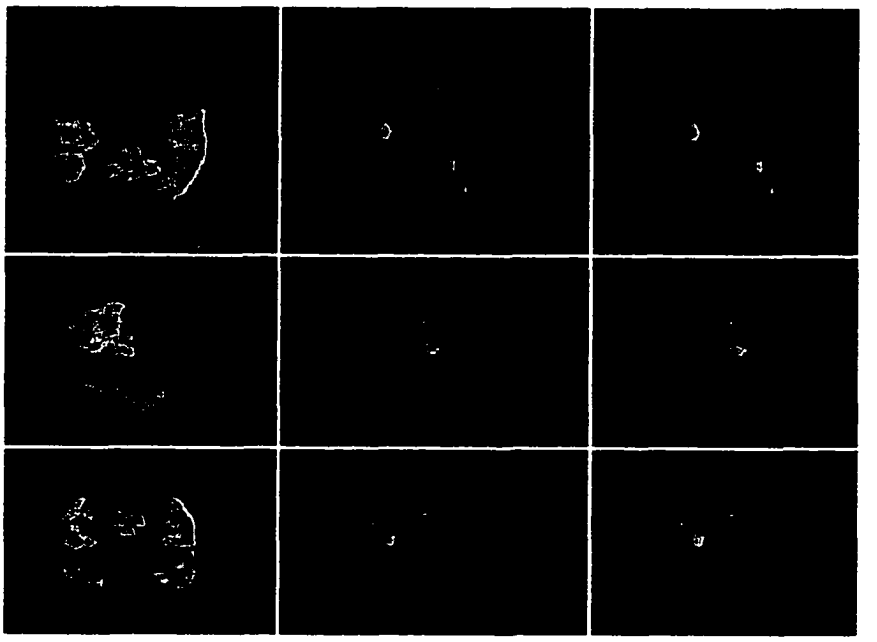


Figure 54. Negatively statistically significant differences:
Right-Left hand motor study (ex. 3).

Table 14. SVT statistical tests on Right-hand vs Left-hand motor study (ex. 3).

Statistics & Tests / Probabilistic ROI	Mean (difference image)	Standard Deviation	ROI Z statistic	Significance of Atlas
Cerebellum	--	--	--	--
Frontal Lobe	-0.60	11.74	3.07	Significant
Occipital Lobe	0.90	10.98	2.94	Significant
Parietal Lobe	-0.47	8.43	2.61	Significant
Temporal Lobe	0.21	11.47	0.81	Not Signif.

The second type of statistical testing is to determine the location (voxels) of statistically significant metabolic changes. This is a standard procedure in most techniques for functional image analysis. However, the SVT differs

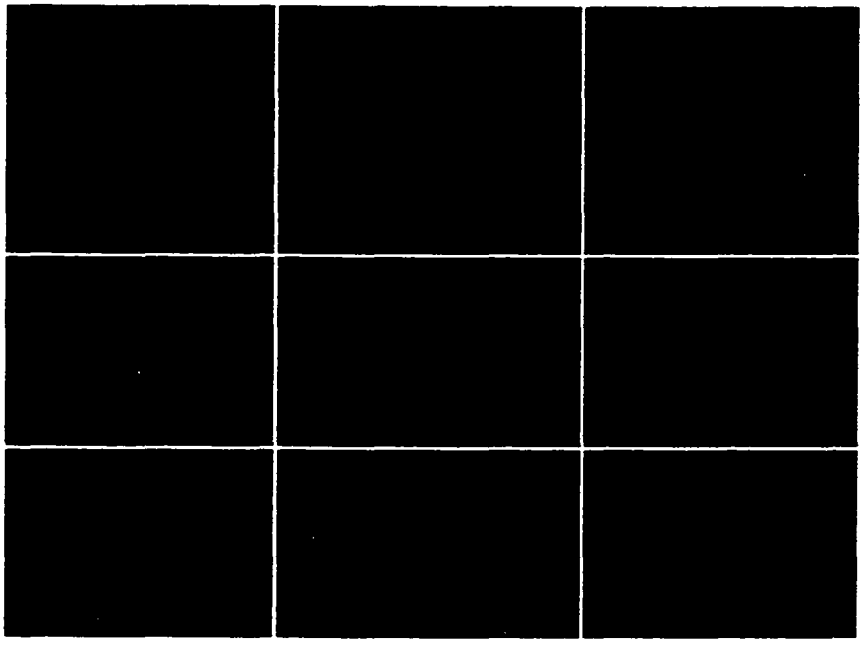


Figure 55. Positively activated significant (increase) changes, SVT left, Simple 97% thresholding, middle. Results are superimposed in the right-most column.

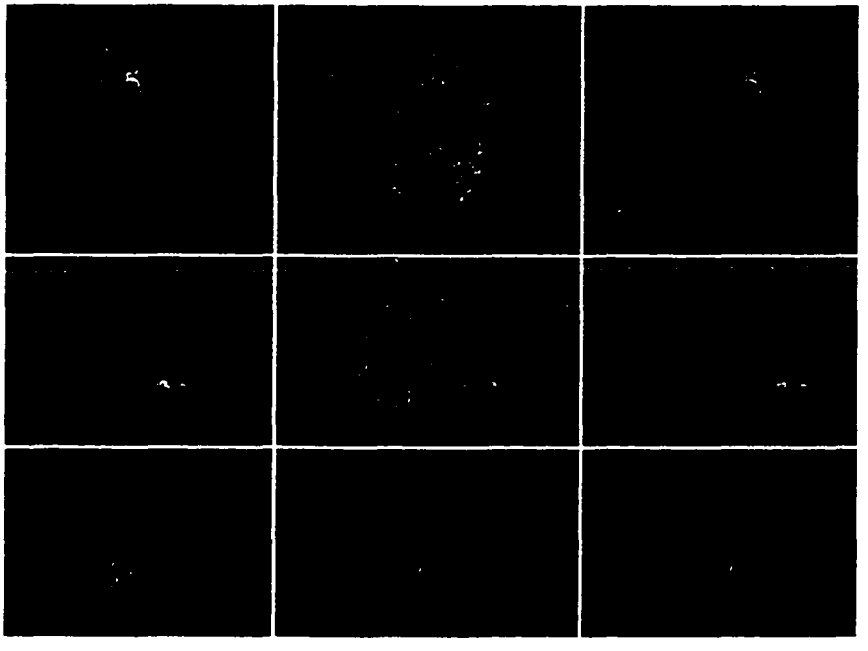


Figure 56. Negatively activated significant changes, SVT left, Simple 97% thresholding, middle.

in that location tests are run only over the search regions of high significance levels, according to the first tests.

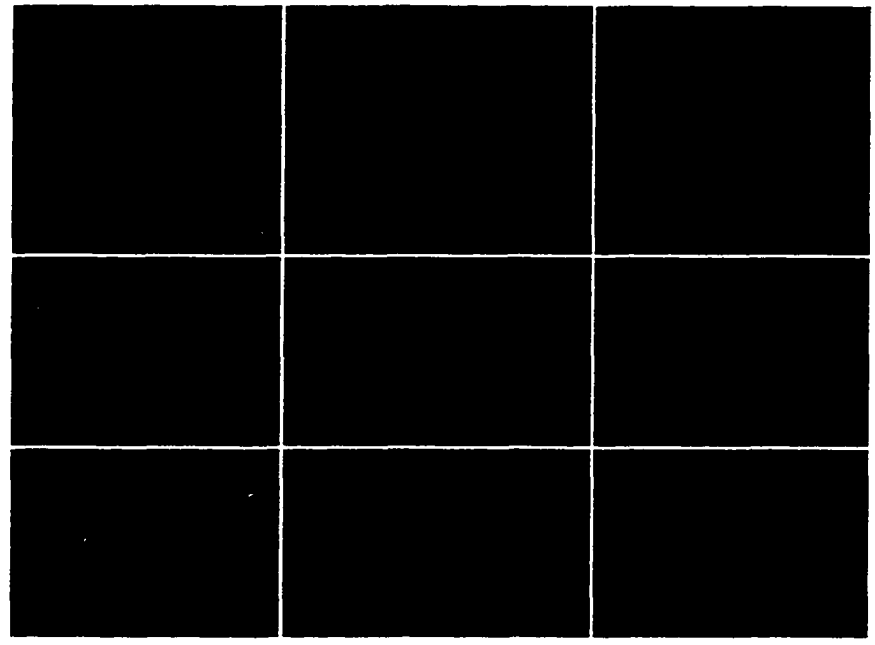


Figure 57. A major difference in the statistical analysis. between SVT (left) and Simple 97% thresholding (middle) methods. in the right Temporal lobe.

Various examples of motor studies and pre/post drug treatment are discussed and compared to the simple 97% thresholding in active-rest single-subject studies.

CHAPTER III

FREQUENCY ADAPTIVE WAVELET THRESHOLDING METHOD

In this chapter we will try to combine knowledge from the areas of Wavelet Analysis, Decision Theory and Parameter Estimation to address problems arising in Image Analysis. In particular, we are interested in developing a new "Cluster Group Classification" (CGC) method for quantitative evaluation of families of image-registration techniques applied to groups of volumetric data.

We first motivate the study by looking at a 1D signal and the effects of thresholding the wavelet coefficients. Following an approach of Donoho & Johnstone we use decision theoretical methods and least-squares estimates to propose a meaningful (in an "optimal estimator" sense) scheme for denoising, analysis and comparison of signals in wavelet space.

Along the way we will summarize the theory behind Multi-Resolution Analyses, wavelets, the discrete wavelet transform and their properties.

We will propose a "soft-thresholding" nonlinearity on the wavelet coefficients, and will exploit its optimality characteristics (regular and asymptotic). Finally, we will apply the theory to a collection of 3D PET data and evaluate the performance of three warps based on the CGC method.

1. Preliminaries

1.1. Motivation

Our main interest now is to develop a quantitative group ranking of various image registration techniques. To do this we will transform all warped data from the usual "spatial-domain" to a new "wavelet" space. The reason behind that is a two-fold: We can do image compression in wavelet space and thus have a way of extracting concisely the information content of the warped data, and secondly, we have a meaningful way to denoise the images (in sense of optimizing certain *Risk* functions) using their wavelet transformations. Figure 58, depicts this graphically.

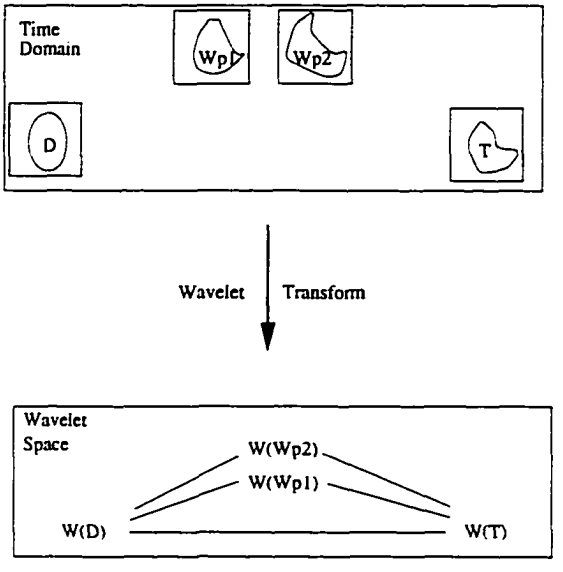


Figure 58. Image Analysis in Transform Space.

In Figure 59, we show the 5 original volumetric data sets (left-most column), their LS warps (second column), MI warps (third column), AF alignments (fourth column) and the target volume in the right-most column, see Section I.5.3.2. Our ultimate aim in this chapter is to quantify the group performance of these 3 registration algorithms.

1.2. General Problem

We now try to find a concise image representation that contains the

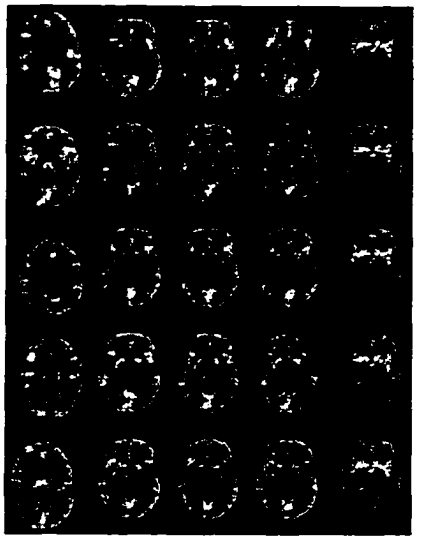


Figure 59. Five PET Data volumes, their LS, MI and Affine Warps, and the Target volume.

essential part of the signal. From an empirical point of view, thresholding the wavelet transform provides a way to extract the essence of the data content of a signal. In Figure 60, we have an image (dotted curve) and its wavelet transform (solid brighter curve). It is visible that very few of the wavelet coefficients, across frequency range and location, have magnitudes larger than 0.3. A natural question to ask is "If we set to zero all wavelet coefficients smaller than 0.3 and then recover the image (invert the WT) will we get a reasonable representation of the original function?" An example illustrating the answer to that question is shown in Figure 61, where we used only 0.5% of the wavelet coefficients to recover the "HeavySine" function. Even though the new signal is not a perfect approximation of the original, at 200 : 1 compression it does capture the main trend of the "HeavySine" function.

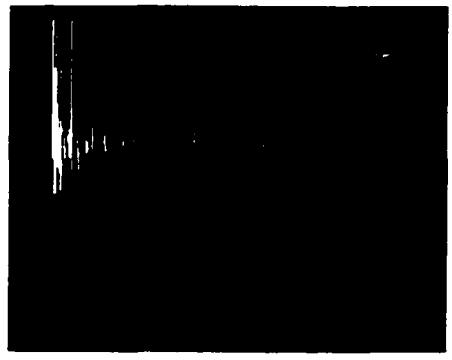


Figure 60. The "Heavy-Sine" Function and its WT.

Figure 61. The "Heavy-Sine" Function and the IWT of its WT thresholded at 0.5 percent (200:1 compression).

The above leads us to the empirical conclusion that the large in magnitude wavelet coefficients indeed determine the core of signals.

Question: How do we select a meaningful wavelet thresholding scheme?

Answer: Select a thresholding method that denoises the signal at the same time.

1.3. Decision Theory

Suppose we have data $\{Y_i\}$ and we propose the following

Model: $Y_i = f(t_i) + e_i,\ 0 \leq i \leq N-1, \quad e_i \sim N(0, \sigma^2) IID.$

Our goal is to recover the unknown function f from the data $Y = \{Y_i\}$. If

$\widehat{f} = \{\widehat{f}(t_i)\}$ is an estimate of the true function f, we measure its performance by the average quadratic loss at the sample points

$$R(\widehat{f}, f) = \frac{1}{N} E\left(||\widehat{f} - f||^2\right) = \frac{1}{N} E\left(\sum_{i=0}^{N-1} [\widehat{f}(t_i) - f(t_i)]^2\right)$$

Small values for the *Risk* functional yield good estimates.

Notation: We consider spatially adaptive estimators $\widehat{f}$ defined by reconstruction formulas

$$\widehat{f}(\bullet) = T(y, d(y))(\bullet),$$

where $d(y)$ is a data adaptive choice for a spatially smoothing parameter.

Examples: (a) Piece-wise Constant Reconstruction

$$T_{PC}(y, d(y))(t) = \sum_{l=1}^{L} Ave(y_i : \ t_i \in A_l) I_{A_l}(t),$$

where the intervals A_l form a partitioning of $[0, 1]$ and $A_1 = [0, d_1), A_2 = [d_1, d_1 + d_2), \cdots, A_L = [d_1 + \cdots + d_{L-1}, d_1 + \cdots + d_L]$, with $\sum_{l=1}^{L} d_l = 1$.

(b) Piece-wise Polynomial

$$T_{PP(D)}(y, d(y))(t) = \sum_{l=1}^{L} \widehat{p}_l(t) I_{A_l}(t),$$

where $\widehat{p}_l(t) = \sum_{k=0}^{D} a_k t^k$ are polynomials of degree D on the partitioning subinterval A_l.

Definition 3.1. *Ideal Adaptation is the risk performance achieved by our reconstruction method $T(y, d(y))$ for the "best" choice $(\Delta(f))$ of the smoothing parameter $d(y)$ for the underlying function f. That is:*

$$R(T(y, \Delta(f)), f) = \Re_{N,\sigma}(T, f) = \inf_d R(T(y, d), f)$$

is the Ideal Risk.

For example, if $f = \sum_{l=1}^{L} p_l(t) I_{A_l}(t)$ is a piece-wise polynomial of degree D, then an *ideal adaptation* smoothing parameter would supply us with the

information to reconstruct f separately over $A_1, A_2, \cdots, A_L$, instead of over another partitioning of $[0,1]$.

1.4. Least Squares Estimates

Suppose we work with a *Linear Model*

$$Y = X\beta + E,$$

where the design matrix $X_{N\times p}$ is full-rank, $E \sim N(0, \sigma^2 I_{N\times N})$ is random noise, and $\beta_{p\times 1}$ is a parameter vector. Then the LS (least squares) estimator of β is defined by

$$\min_{\beta} ||Y - X\beta|| = ||Y - X\widehat{\beta}_{LS}||.$$

In the linear case $\widehat{\beta}_{LS}$ also can be expressed in the form $\widehat{\beta}_{LS} = (X^T X)^{-1} X^T Y$. If $P_{N\times N}$ is the O.P. (orthogonal projection) matrix onto the $Rg(X)$ we get the *fitted values*

$$\widehat{Y} = X\widehat{\beta}_{LS} = PY = X(X^T X)^{-1} X^T Y$$

$$Var(\widehat{Y}) = Var(X\widehat{\beta}_{LS}) = Var(PY) = PVar(Y)P^T =$$

$$= P\sigma^2 I P^T = \sigma^2 PP^T = \sigma^2 rk(P) = \sigma^2 rk(X) =$$

$$= \sigma^2 \times (\#parameters(\beta))$$

Note: (a) $E(X\widehat{\beta}_{LS}) = X\beta$;

(b) $E\left((X\widehat{\beta}_{LS} - X\beta)^2\right) = Var(X\widehat{\beta}_{LS})$.

In our setting, for the model $Y_i = f(t_i) + e_i$, if $\widehat{f}(t_i)$ is any estimate of f, then the risk measure

$$R(\widehat{f}, f) = \frac{1}{N} E\left(\sum_{i=0}^{N-1} (\widehat{f}(t_i) - f(t_i))^2\right) = \frac{1}{N} E\left(||\widehat{f} - f||^2\right) \geq$$

$$\geq \frac{1}{N} E\left(||\widehat{f}_{LS} - f||^2\right) = \frac{1}{N} \times (NoiseLevel) \times (\#parameters),$$

since the LS estimator minimizes the square mean error and thus yields an ideal risk.

Example: If again $f(t) = \sum_{l=1}^{L} p_l(t) I_{A_l}(t)$ is a piece-wise polynomial of degree D, then the number of parameters in the system is $L \times (D+1)$, and

$$\Re_{N,\sigma}(\widehat{f}, f) = \frac{L(D+1)}{N}\sigma^2$$

Question: Can we approach this ideal performance of an estimator as measured by the risk functional?

2. Discrete Wavelet Transform - Review

2.1. Multi-Resolution Analyses

Definition 3.2. *If there exists a function $\phi(x)$ (father wavelet) satisfying these 3 properties then the collection of induced spaces $\{V_k\}$ is called a Multi-Resolution Analysis (MRA), and the function $\phi(x)$ is termed the scaling function of the MRA.*

(1) V_o is the $L^2(R)$ span of

$$\{\phi(x-s)| \quad s = 0, \pm 1, \pm 2, \pm 3, \cdots\},$$

$$V_k = \left\{ f \in L^2(R) | \quad f(x) = \sum_{s=-\infty}^{\infty} d_n \phi(2^k x - s) \right\}$$

and $\{2^{\frac{k}{2}}\phi(2^k x - s)| \quad s \in Z\}$ is an orthonormal basis of V_k;

(2) $\bigcap_{n=-\infty}^{\infty} V_n = \{0\}$;

(3) If $\overline{A}$ denotes the topological closure of the set A, $\overline{\bigcup_{n=-\infty}^{\infty} V_n} = L^2(R)$.

Note: Since $\phi \in V_o \subseteq V_1 = span\{\phi(2x-s)\}$ then there exist constants (wavelet filter coefficients), $\{d_s\}$,

$$\phi(x) = \sum_{s \in Z} d_s \phi(2x - s)$$

Definition 3.3. *Having the scaling coefficients* $\{d_s\}$ *we define the* Mother Wavelet *by*

$$\psi(x) = \sum_{s=-\infty}^{\infty} (-1)^s d_{s+1} \phi(2x+s).$$

Notes: (a) $\{2^{\frac{k}{2}}\psi(2^k x - s) \mid s,k \in Z\}$ is an O.N.B. (orthonormal basis) of $L_2(R)$;

(b) In practice only finitely many of the filter coefficients are non-trivial;

(c) Following Daubechies [1988] wavelet construction algorithm, if $d_s = 0$ for all $s < m_1$ and $s > M_1$, than the induced wavelets are compactly supported on intervals of length $S = 2K - 1 = M_1 - m_1$

$$supp(\phi) = [0, 2K-1] \qquad supp(\psi) = [-(K-1), K]$$

(d) If $M = K - 1$ than all wavelets $W_{j,k}(y) = c_1\psi(c_2 y + c_3)$ have vanishing moments up to order M.

$$\int_R \psi(y) y^l dy = 0, \quad 0 \le l \le M$$

(e) $\psi(y) \in C^M(R)$, and the number (M) of non-trivial wavelet filter coefficients affects the smoothness characteristics of the induced wavelets.

2.2. Discrete Signal Representations

Let $N = 2^{J+1}$ and $y_i = f(i/N)$, for $0 \le i \le N-1$ be a discretization of f on $[0,1]$. Given a father wavelet ϕ and the corresponding MRA there exists an orthogonal matrix $W = W_{N\times N}$ with entries the wavelet filter coefficients (up to a sign)

$$w = Wy \qquad y = W^T w$$

$w = \{w_i\}_{i=0}^{N-1} = \{w_{j,k}\}$, where $0 \le j \le J$ and $0 \le k \le 2^j - 1$. Also $y_i = \sum_{j,k} w_{j,k} W_{j,k}(i)$ where $W_{j,k}$ denotes the $n \leftrightarrow (j,k)$-th row of W.

Remarks: (a) $\sqrt{N}W_{j,k}(i) \approx 2^{\frac{j}{2}}\psi(2^j t - k)$, for $j \geq j_o$, where $t = \frac{i}{N}$, and ψ is the mother wavelet of the MRA.

(b) $\sum_{i=0}^{N-1} i^l W_{j,k}(i) = 0$, for $0 \leq l \leq M$, $j \geq j_o$ and $0 \leq k \leq 2^j - 1$.

(c) What is the support of $W_{j,k}$?

Recall that $supp(\psi) = [-(K-1), K]$, and $W_{j,k}(i) \approx \psi(2^j t - k)$. Thus, $t \in [(k-(K-1))2^{-j}, (k+K)2^{-j}]$, and $t = \frac{i}{N} = i2^{-(J+1)}$. Therefore, $supp(W_{j,k}) =$

$$= [2^{J+1-j}(k-(K-1)), 2^{J+1-j}(k+K)] =$$

$$= [2^{J-j}(2k-1-S), 2^{J-j}(2k+1+S)].$$

(d) To find out if there is a significant change of f near spatial location t, we only need to look at $w_{j,k}$, for $j \geq j_o$ and location indices k, such that $k2^j \approx t$, $t \in [0,1]$. Furthermore, large wavelet coefficients appear in areas of major functional spatial activity.

2.3. Selective Wavelet Reconstruction Schemes

Definition 3.4. *Selective Wavelet Reconstruction.*

Given a list d of pairs (j,k), if $w = Wy$, the *selective wavelet reconstruction* estimator is

$$T_{SW}(y,d) = \widehat{f}(i) = \sum_{(j,k)\in d} w_{j,k}W_{j,k}.$$

Why use such spatially adaptive reconstruction technique?

(a) Oftentimes, the important signal information is contained in a small subset d of all (j,k) pairs of empirical wavelet coefficients;

(b) Under the model $y_i = f(i) + e_i$, if $e_i \sim N(0,\sigma^2)$ IID, then $z_i = We_i \sim N(0,\sigma^2)$ IID, since W is orthogonal and $z = We \sim N(0, W\sigma^2 IW^T) = N(0,\sigma^2 I)$. Then the time domain model $y = f + e$ becomes $w = \theta + z$ in wavelet space with $w = Wy$, $\theta = Wf$ and $z = We$.

All of the empirical wavelet coefficients contribute noise with variance σ^2, but only a few of carry the essential information of the image f.

Definition 3.5. *Ideal Risk for the selective wavelet reconstruction scheme is*

$$\Re_{N,\sigma}(SW, f) = \inf_{d} R_{N,\sigma}(T_{SW}(y, d), f) = R_{N,\sigma}(T_{SW}(y, \Delta(f)), f),$$

where $\Delta(f)$ is the optimal spatial adaptation parameter selecting the "best" list of pairs (j, k).

Note: From now on f is a piece-wise (unknown) polynomial of degree D, $f(t) = \sum_{l=1}^{L} p_l(t) I_{A_l}(t)$, where $\{A_l\}$ is a partitioning of $[0, 1]$. We also use wavelet bases with "number of parameters" $M \geq D$.

What is an upper bound for the number of non-trivial empirical wavelet coefficients of this model?

If $\theta = Wf$, then $\theta_{j,k} \neq 0$ for

(a) $0 \leq j < j_o$, or

(b) for $j \geq j_o$ if the (support of $W_{j,k}$) interval associated with $\theta_{j,k}$, $[2^{-j}(k - (K-1)), 2^{-j}(k + K)]$, contains a break point of f. This is because, for $j \geq j_o$, $\left(\theta = Wf,\ f = \sum_{l=1}^{L} p_l I_{A_l} = \sum_{j,k} \theta_{j,k} W_{j,k}\right)$, by *orthonormality* of $\{W_{j,k}\}$ and because $D \leq M$ we have

$$\theta_{j,k} = < f, W_{j,k} > = \sum_{i=0}^{N-1} f(i) W_{j,k}(i) =$$

$$= \sum_{p=0}^{D} a_p \sum_{i=0}^{N-1} i^p W_{j,k}(i) = 0,$$

unless a break point occurs in this interval, since $W_{j,k} \approx \psi(2^j - k)$ and ψ has trivial moments up to order $M \geq D$.

Claim 3.6. *An upper bound for the non-vanishing wavelet coefficients is:*

$$\#\{(j, k) \in d :\ \theta_{j,k} \neq 0\} \leq 2^{j_o} + (J + 1 - j_o) SL.$$

Note that $(J+1-j_o)$ is the number of resolution levels beyond j_o and $L =$ is the maximum number of break points of f. Also, since $W_{j,k}$ is localized near $t = k2^{-j}$, we have

$$S = \left|\{k_1 : \ t = k_1 2^{-j} \in [2^{-j}(k-(K-1), 2^{-j}(k+K)]\}\right| =$$

$$= |\{k_1 : \ k_1 \in [k-(K-1), k+K]| = 2K-1 =$$

$$= Wavelet\_supp\_length.$$

Denote $d^* = \{(j,k) : \ \theta_{j,k} \neq 0\}$. Then,

$$|d^*| \leq 2^{j_o} + (J+1-j_o)SL.$$

Let $\widehat{Y}_{LS} = \sum_{j,k} w_{j,k} W_{j,k}$ be the LS estimator of f. Then $\widehat{f} = \sum_{(j,k)\in d^*} w_{j,k} W_{j,k}$ is also a LS estimator of $f = \sum_{j,k} \theta_{j,k} W_{j,k} = \sum_{(j,k)\in d^*} \theta_{j,k} W_{j,k}$, since

$$||\widehat{Y}_{LS} - f||^2 = \min_Y ||Y-f||^2 = \min_Y ||e||^2$$

But, $||Y-f||^2 = ||W(Y-f)||^2 = ||w-\theta||^2 =$

$$\sum_{j,k} |w_{j,k} - \theta_{j,k}|^2 = \sum_{(j,k)\in d^*} |w_{j,k} - \theta_{j,k}|^2 + \sum_{(j,k)\notin d^*} |w_{j,k}|^2.$$

The last expression is minimized for $w_{j,k} = 0$, for $(j,k) \notin d^*$.

Summarizing $R(T_{SW}(y, d^*), f) = \frac{1}{N}|d^*|\sigma^2 \leq (C_1 + C_2 J)\frac{\sigma^2}{N}$. Thus, the ideal risk for the selective wavelet estimators is

$$\Re_{N,\sigma}(SW, f) = O\left(\frac{\sigma^2 \ln(N)}{N}\right).$$

Observe that this upper bound is almost as good as the optimal bound for the "ideal" piece-wise polynomial reconstruction, $\Re_{N,\sigma}(PP, f) = L(D+1)\frac{\sigma^2}{N}$ when we have an access to an oracle providing us with the information about the break points of f. In general, however, it is unlikely that we would have

such information. Therefore, as we will see shortly, our adaptive shrinkage approach producing estimators within $\ln^2 N$ of the ideal oracle-supported adaptation is reasonably good.

2.4. Diagonal Linear Projection Reconstruction

For the model $Y_i = f_i + \sigma z_i$, $0 \le i \le N-1$, $z_i \sim N(0,1)$, and σ is the noise level, estimate performance is evaluated by the risk measure

$$R(\widehat{f}, f) = \frac{1}{N} E\left(||\widehat{f} - f||^2\right)$$

Consider a new, *Diagonal Linear Projection* (DP), reconstruction method:

$$T_{DP}(Y, d) = \{d_i y_i\},$$

where $d_i = 0$ or 1. This estimator either keeps or kills an observation.

Claim 3.7. *The "ideal" DP risk is attained for* $d_i = I_{|f_i|>\sigma}$.

Proof: $\Re(DP, f) = \inf_d R(T_{DP}(Y, d), f) = \frac{1}{N} \inf_d \sum_i E(|d_i Y_i - f_i|^2) =$

$$= \frac{1}{N} \left[\inf_d \sum_{|f_i|>\sigma} E(|d_i Y_i - f_i|^2) + \inf_d \sum_{|f_i|\le\sigma} E(|d_i Y_i - f_i|^2) \right].$$

Note that over the first index set ($|f_i| > \sigma$) $E(|d_i Y_i - f_i|^2) \ge \sigma^2$ with equality attained only if $d_i = 1$, because $E(|d_i Y_i - f_i|^2)$ is either equal to $|f_i|$ (if $d_i = 0$) or equal to $E|\sigma z_i|^2 = \sigma^2$ (if $d_i = 1$). Similarly, Over the second index set ($|f_i| \le \sigma$) $E(|d_i Y_i - f_i|^2) \le \sigma^2$ (if $d_i = 0$), and $E(|d_i Y_i - f_i|^2) = \sigma^2$ (if $d_i = 1$). Therefore, the infimum is obtained at $d_i = I_{|f_i|>\sigma}$.

In other words, $\Re(DP, f) =$

$$= \frac{1}{N} \left[\sum_{|f_i|>\sigma} \sigma^2 + \sum_{|f_i|\le\sigma} |f_i|^2 \right] = \frac{1}{N} \sum_i Min(|f_i|^2, \sigma^2)$$

Δ

Note: (1) This ideal risk may not be attained for any estimator, but how close to it can we get?

(2) $\Re_{N,\sigma}(SW, f) = \Re_{N,\sigma}(DP, f)$, because $T_{SW}(Y, d) = \sum_d w_{j,k} W_{j,k} = W^T T_{DP} W$, and $E(||T_{SW} - f||^2) = E(||W^T T_{DP} W(Y) - f||^2) = E(||T_{DP}(W(Y)) - W(f)||^2) = E(||T_{DP}(w) - \theta||^2)$. Taking infimum over d of the LHS (left-hand-side) gives $\Re_{N,\sigma}(SW, f)$, and infimum of the RHS gives $\Re_{N,\sigma}(DP, f)$, recall $T_{DP}(w, d) = \{d_i w_i\}_i$, where $d_i = 0$ or 1.

(3) If $\widehat{f}^* = W^T T_{DP} W(Y)$, we will show that $R(\widehat{f}^*, f) \leq (1 + \ln(N + 4)) \left(\frac{\sigma^2}{N} + \Re(DP, f)\right)$. But because, $\Re_{N,\sigma}(SW, f) = O\left(\frac{\sigma^2 \ln N}{N}\right)$,

$$R(\widehat{f}^*, f) \longrightarrow 0,$$

at the rate of $\frac{\ln^2 N}{N}$, as $N \to \infty$. However, if no threshold (NT) is applied the risk functional is constant in N

$$R(NT, f) = \frac{1}{N} E(||Y - f||^2) = \frac{1}{N} \sum_{i=0}^{N-1} E(||Y_i - f_i||^2) = \frac{1}{N} \sum_{i=0}^{N-1} \sigma^2 = \sigma^2.$$

This serves as a strong motivation for choosing the wavelet based to the spatial-domain image-registration analysis.

3. Spatially Adaptive Techniques

3.1. Upper bounds on *Risk* measures

Definition 3.8. *Let $w_i = \theta_i + \sigma z_i$, then we define a thresholding nonlinearity*

$$\eta_\lambda(x) = sgn(x) [|x| - \lambda]_+ ,$$

where $\lambda = \left\{ \begin{array}{lcl} \lambda^{DJ} & = & \sigma\sqrt{2 \ln N} \\ \lambda^{DS} & = & \sigma\sqrt{2 \ln (2n_j + 4)^{(1-\alpha)}} \end{array} \right\}$.

The spatially adaptive thresholding value λ^{DJ} was proposed by Donoho and Johnstone in 1994. In contrast, we define a frequency-adaptive non-uniform wavelet shrinkage based on the threshold λ^{DS}.

Theorem 3.9 [Donoho-Johnstone, 1994]. *Let* $\widehat{\theta}_i = \eta_{\lambda^{DJ}}(w_i)$, *then*

$$R(\widehat{\theta},\theta) = \frac{1}{N}E\left(||\widehat{\theta}-\theta||^2\right) \leq (1+2\ln(N))\left(\frac{\sigma^2}{N} + \frac{1}{N}\sum_i Min(\theta_i^2,\sigma^2)\right).$$

We now present the proof of an analogous result using our frequency-adaptive thresholding nonlinearity.

Proposition 3.10. *Let* $\widehat{\theta}_i = \eta_{\lambda^{DS}}(w_i)$, *then*

$$R(\widehat{\theta},\theta) = \frac{1}{N}E\left(||\widehat{\theta}-\theta||^2\right) \leq (1+2(1-\alpha)\ln(N+4)) \times$$

$$\times \left(\frac{\sigma^2}{N}\sum_j \frac{2^j}{(2^{j+1}+4)^{(1-\alpha)}} + \frac{1}{N}\sum_i Min(\theta_i^2,\sigma^2)\right).$$

Let $r = 1-\alpha$, then for $r \neq 1$

$$R(\widehat{\theta},\theta) \leq (1+2\ln(N+4)^r)\left(\frac{\sigma^2}{2-2^r}\left(\frac{1}{N^r}-\frac{1}{N}\right) + \Re_{N,\sigma}(DP,\theta)\right).$$

And for $r = 1$

$$R(\widehat{\theta},\theta) \leq (1+2\ln(N+4))\left(\frac{\sigma^2\ln N}{N} + \Re_{N,\sigma}(DP,\theta)\right).$$

Proof: The proof consists of 3 parts. We first show the result for a single observation with noise-level $\sigma = 1$. Then we extend this to an arbitrary noise-level σ. And at the end we generalize the result to any collection of N observations.

(1) Suppose $X \sim N(\mu,1)$, $\delta = (2n_j+4)^{-(1-\alpha)}$, $n_j = 2^j$, $t = \sqrt{2\ln(\delta^{-1})}$, and $\eta_t(x) = sgn(x)[|x|-t]_+$. We try to estimate $E((\eta_t(x)-\mu)^2)$. Observe that

$$|x| - |x| \wedge t = \left\{ \begin{array}{ccc} |x|-t & , & |x| \geq t \\ 0 & , & |x| < t \end{array} \right\} = [|x|-t]_+$$

Hence, $\eta_t(x) = sgn(x)[|x| - |x| \wedge t] = x - sgn(x)[|x| \wedge t]$. Taking the expectation

$$R = E((\eta_t(X)-\mu)^2) = E((X-\mu-sgn(X)[|X| \wedge t])^2) =$$

$$= E((X-\mu)^2) - 2E((X-\mu)sgn(X)[|X| \wedge t]) + E([|X| \wedge t]^2) =$$

$$= Var(X) - 2P(|X| < t) + E(X^2 \wedge t^2) =$$

$$= 1 - 2P(|X| < t) + E(X^2 \wedge t^2).$$

Here we used that $I = E((X-\mu)sgn(X)[|X| \wedge t]) = P(|X| < t)$, because

$$I = \frac{1}{\sqrt{2\pi}} \int_R (x-\mu)sgn(x)[|x| \wedge t] e^{-\frac{(x-\mu)^2}{2}} dx =$$

$$= -\frac{1}{\sqrt{2\pi}} \int_R sgn(x)[|x| \wedge t] de^{-\frac{(x-\mu)^2}{2}} = \frac{1}{\sqrt{2\pi}} \int_R e^{-\frac{(x-\mu)^2}{2}} dQ(x),$$

where $Q(x) = sgn(x)[|x| \wedge t] = \begin{cases} x & , \quad -t < x \leq 0 \\ -t & , \quad -\infty < x \leq t \\ x & , \quad 0 \leq x < t \\ t & , \quad t \leq x < \infty \end{cases}$.

And $dQ(x) = \begin{cases} dx & , \quad |x| < t \\ 0 & , \quad |x| \geq t \end{cases}$. Thus

$$I = \frac{1}{\sqrt{2\pi}} \int_{|x|<t} e^{-\frac{(x-\mu)^2}{2}} dx = P(|X| < t)$$

We now construct two different bounds on R and express the fact that R is less than the minimum of them.

For one thing, $X^2 \wedge t^2 \leq t^2$ and $R \leq 1 - 2P(|X| < t) + E(t^2) \leq 1 + t^2$.

Secondly, $X^2 \wedge t^2 \leq X^2$ and $R \leq 1 - 2P(|X| < t) + E(X^2) = 1 - 2P(|X| < t) + Var(X) + \mu^2 = 1 - 2P(|X| < t) + 1 + \mu^2 = 2(1 - P(|X| < t)) + \mu^2 = 2P(|X| \geq t) + \mu^2$.

Note that $P(|X| \geq t)$ is implicitly a function of the (unknown) parameter μ (because $X \sim N(\mu, 1)$), so we let $g(\mu) = 2P(|X| \geq t)$ and try to find an upper bound for g. Since g is infinitely smooth, for every μ we can expand it around the origin in a Taylor series

$$g(\mu) = g(0) + g'(0)\mu + g''(\xi)\frac{\mu^2}{2},$$

$$g(\mu) = 2P(|X| \geq t) = \frac{2}{\sqrt{2\pi}} \left[\int_{-\infty}^{-t} e^{-\frac{(x-\mu)^2}{2}} dx + \int_t^{\infty} e^{-\frac{(x-\mu)^2}{2}} dx \right].$$

$$g'(\mu) = \frac{2}{\sqrt{2\pi}} \left[\int_{-\infty}^{-t} (x-\mu) e^{-\frac{(x-\mu)^2}{2}} dx + \int_t^{\infty} (x-\mu) e^{-\frac{(x-\mu)^2}{2}} dx \right] =$$

$$= \frac{2}{\sqrt{2\pi}} \left[-e^{-\frac{(t+\mu)^2}{2}} + e^{-\frac{(t-\mu)^2}{2}} \right].$$

Thus, $g'(0) = 0$ and

$$g''(\mu) = \frac{2}{\sqrt{2\pi}} \left[(t-\mu)e^{-\frac{(t-\mu)^2}{2}} - (t+\mu)e^{-\frac{(t+\mu)^2}{2}} \right],$$

$$g(\mu) \leq g(0) + \max_{\mu} \frac{||g''(\mu)||}{2} \mu^2.$$

We first estimate

$$g(0) = \frac{2}{\sqrt{2\pi}} \left[\int_{-\infty}^{-t} e^{-\frac{x^2}{2}} dx + \int_{t}^{\infty} e^{-\frac{x^2}{2}} dx \right] = 4\Phi(-t),$$

where $\Phi(t) = \Phi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx$ is the cdf of $X \sim N(0,1)$.

$$4\Phi(-t) = 4P(X < -t) = 4P(X > t) = \frac{4}{\sqrt{2\pi}} \int_{t}^{\infty} e^{-\frac{x^2}{2}} dx \leq$$
$$\leq \frac{4}{\sqrt{2\pi}} \int_{t}^{\infty} \frac{x}{t} e^{-\frac{x^2}{2}} dx \leq -\frac{4}{t\sqrt{2\pi}} \int_{t}^{\infty} de^{-\frac{x^2}{2}} =$$
$$= \frac{4}{t\sqrt{2\pi}} e^{-\frac{t^2}{2}} \leq (t^2+1)e^{-\frac{t^2}{2}},$$

for $t \geq 1$.

We are now interested in determining an upper bound for the magnitude of $g''(\mu) = \frac{2}{\sqrt{2\pi}} \left[(t-\mu)e^{-\frac{(t-\mu)^2}{2}} - (t+\mu)e^{-\frac{(t+\mu)^2}{2}} \right]$.

$$\max_{\mu} ||g''(\mu)|| \leq$$
$$\leq 2 \left[\max_{\mu} || \frac{(t-\mu)e^{-\frac{(t-\mu)^2}{2}}}{\sqrt{2\pi}} || + \max_{\mu} || \frac{(t+\mu)e^{-\frac{(t+\mu)^2}{2}}}{\sqrt{2\pi}} || \right] =$$
$$= 4 \max_{x} || \frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}} || \leq \frac{4}{\sqrt{2\pi e}},$$

since $u(x) = x\phi(x) = \frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ has a maximum of $\frac{4}{\sqrt{2\pi e}}$ at $x = 1$.

Combining these two bounds for $g(0)$ and $g''(\mu)$ we obtain an upper bound for $g(\mu) \leq (t^2+1)e^{-\frac{t^2}{2}} + t^2\mu^2$. Going back to the performance estimate as measured by the risk factor

$$R = E((\eta_t(x) - \mu)^2) \leq \begin{cases} 1 + t^2 \\ e^{-\frac{t^2}{2}}(t^2+1) + t^2\mu^2 + \mu^2 \end{cases},$$

for all $t \geq 1$.

Thus, $R \leq \begin{cases} (1+t^2)(1+e^{-\frac{t^2}{2}}) \\ (t^2+1)(e^{-\frac{t^2}{2}}+\mu^2) \end{cases} \leq (1+t^2)(e^{-\frac{t^2}{2}}+\mu^2 \wedge 1)$, for all $t \geq 1$.

(2) How would this bound change if we apply it to $\widehat{\theta} \sim N(\theta, \sigma^2)$? $\left(\theta = \mu, \widehat{\theta} = \mu + \epsilon, \epsilon \sim N(0, \sigma^2)\right)$.

Let $X \sim N(\theta, 1)$. Denote $\delta = e^{-\frac{t^2}{2}}$. Then, we showed $R(X,t) = E((\eta_t(x) - \theta)^2) \leq (1+t^2)(\delta + \theta^2 \wedge 1)$. Define $\widehat{\theta} = \sigma X \sim N(\mu, \sigma^2)$, where $\mu = \sigma\theta$. Let

$$t = \sigma\sqrt{2\ln(2n_j+4)^{(1-\alpha)}}$$

$$R(\widehat{\theta}, t) = E((\eta_t(\widehat{\theta}) - \mu)^2) = E(((\widehat{\theta} - \mu) - sgn(\widehat{\theta})[|\widehat{\theta}| \wedge t])^2) =$$

$$= E(((\sigma X - \sigma\theta) - sgn(X)[\sigma|X| \wedge t])^2) =$$

$$= \sigma^2 E\left(\left((X - \theta) - sgn(X)[|X| \wedge \frac{t}{\sigma}]\right)^2\right) =$$

$$= \sigma^2 R\left(X, \frac{t}{\sigma}\right) = \sigma^2 R(X, t_1) \leq \sigma^2 (1+t_1^2)(\delta' + \theta^2 \wedge 1),$$

where $t_1 = \frac{t}{\sigma}$ and $\delta' = e^{-\frac{t^2}{2\sigma^2}} = e^{-\frac{t_1^2}{2}}$. Then,

$$R(\widehat{\theta}, t) \leq (1+t_1^2)(\sigma^2\delta' + (\sigma\theta)^2 \wedge (\sigma^2)) =$$

$$= (1+t_1^2)(\sigma^2\delta' + \mu^2 \wedge (\sigma^2)) =$$

$$= (1+2\ln(N+4)^{(1-\alpha)})\left(\sigma^2 \frac{1}{(2n_j+4)^{(1-\alpha)}} + \mu^2 \wedge (\sigma^2)\right)$$

(3) Finally, we consider the situation when we have N observations $\widehat{\theta} = \{\widehat{\theta}_i\}_i = \{\widehat{\theta}_{j,k}\}_{j,k}$, where $\widehat{\theta}_i \sim N(\theta_i, \sigma^2)$. In matrix notation $\widehat{\theta} \sim N(\Theta, \sigma^2 I)$.

$$R(\widehat{\theta}, t) = \frac{1}{N} E(||\eta_t(\widehat{\theta}) - \Theta||^2) = \frac{1}{N} \sum_{j,k} R(\widehat{\theta}_{j,k}, t).$$

Recall that we work with

$$t = \lambda^{DS} = \sigma\sqrt{2\ln\left[(2n_j+4)^{(1-\alpha)}\right]} = \sigma\sqrt{2\ln\left[\frac{1}{\delta}\right]}.$$

The restrictions on the parameter t is $t \geq 1$.

In summary,

$$R(\widehat{\theta},\theta) \leq$$

$$\leq \frac{1}{N}\sum_j \left[\left(1+2\ln(2n_j+4)^{(1-\alpha)}\right)\left(\sigma^2\frac{2^j}{(2n_j+4)^{(1-\alpha)}}+\sum_k \theta_{j,k}^2\wedge\sigma^2\right)\right].$$

Define $r=(1-\alpha)$. If $r=1$

$$R(\widehat{\theta},\theta) \leq (1+2\ln(N+4)^{(1-\alpha)})\left[\sigma^2\frac{J}{2N}+\frac{1}{N}\sum_{j,k}\theta_{j,k}^2\wedge\sigma^2\right] =$$

$$= (1+2\ln(N+4)^{(1-\alpha)})\left[\sigma^2\frac{J}{2N}+\Re_{N,\sigma}(SW,\theta)\right]$$

$$= (1+2\ln(N+4))\left[\sigma^2\frac{J}{2N}+\Re_{N,\sigma}(SW,\theta)\right].$$

If $r\neq 1$, note that

$$\sum_j \frac{2^j}{(2^{j+1}+4)^r} \leq \frac{1}{2^r}\frac{1-(2^{1-r})^{J+1}}{1-2^{1-r}} =$$

$$= \frac{1}{2^r-2}(1-N^{(1-r)}) = \frac{N^\alpha-1}{2-2^{(1-\alpha)}}$$

Therefore, in general,

$$R(\widehat{\theta},\theta) \leq (1+2r\ln(N+4))\left[\sigma^2\frac{1}{N}\frac{1}{2-2^{(1-\alpha)}}(N^\alpha-1)+\Re_{N,\sigma}(SW,\theta)\right].$$

△

What are the restriction on α? We have that

$$t=\lambda^{DS}=\sqrt{2\ln\left[(2n_j+4)^{(1-\alpha)}\right]} \geq 1$$

This restriction is satisfied provided $0\leq\alpha\leq\frac{1}{2}$ for all $n_j=2^j$. Typically, the threshold $\alpha=0.05$, mimicking statistical significance level of testing.

3.2. Optimality Properties of DJ and DS Estimates

Let $Y = f + \sigma e$, and $\widehat{f}$ be an estimate of f. Denote $w = WY$, $\theta = Wf$, $z = We$ and $\widehat{\theta} = W\widehat{f}$.

Note: (1) $||f - \widehat{f}|| = ||W(f - \widehat{f})|| = ||\theta - \widehat{\theta}||$.

(2) $T_{SW} = W^T T_{DP} W$. Hence, $R(T_{SW}(Y,d), f) = \frac{1}{N} E(||T_{SW}(Y,d) - f||^2) = \frac{1}{N} E(||W(T_{SW} - f)||^2) = \frac{1}{N} E(||T_{DP}(W(Y),d) - \theta||^2) = \frac{1}{N} E(||T_{DP}(w,d) - \theta||^2)$.

(3) Let $\widehat{\theta}_i = \eta_\lambda(w_i) = sgn(w_i)[|w_i| - \lambda]_+$, where

$$\lambda = \left\{ \begin{array}{lcl} \lambda^{DJ} & = & \sigma\sqrt{2\ln N} \\ \lambda^{DS} & = & \sigma\sqrt{2\ln(2n_j + 4)^{(1-\alpha)}} \end{array} \right\}.$$

Define $\widehat{f} = W^T \widehat{\theta} W(Y)$. Then, by note (1),

$$R(\widehat{f}, f) = \frac{1}{N} E(||\widehat{f} - f||^2) = \frac{1}{N} E(||\widehat{\theta} - \theta||^2) = R(\widehat{\theta}, \theta).$$

Therefore, by Proposition 3.10 for $0 < \alpha < 0.5$,

$$R(\widehat{f}_{DS}, f) = R(\widehat{\theta}, \theta) \leq (1 + 2(1-\alpha)\ln(N+4)) \left[\frac{\sigma^2}{N} \frac{N^\alpha - 1}{2 - 2^{1-\alpha}} + \Re_{N,\sigma}(SW, f) \right].$$

Whereas, by Theorem 3.9, the upper bound for the Donoho and Johnstone's estimator is

$$R(\widehat{f}_{DJ}, f) \leq (1 + 2\ln N) \left(\frac{\sigma^2}{N} + \Re_{N,\sigma}(SW, f) \right).$$

Question: Are there estimators $\widehat{f}$ that can make the first term of these upper bounds smaller?

Answer: Essentially, there are no such estimators!

Examining closely the proof of Proposition 3.10 we see that there is a dependence between the two terms involved in the upper bound of $R(\widehat{f}_{DS}, f)$. A decrease of one of them causes the other one to go up. Therefore, to study "optimal" estimators we fix, say, the second term in that product and try to find estimators that make the first term smaller. The following theorem, proven by Donoho and Johnstone in 1994, shows that the thresholding level $\lambda_{DJ} = \sigma\sqrt{2\ln N}$ yields an optimal estimator $\widehat{f}_{DJ} = sgn(w)[|w| - \lambda_{DJ}]_+$, in the sense that it achieves the smallest possible upper bound for the risk $R(\widehat{f}, f)$, having the second term $(\sigma^2/N + \Re_{N,\sigma}(f, \sigma^2))$ fixed.

Theorem 3.11 [Donoho & Johnstone]. *In the above setup,*

$$\inf_{\widehat{\theta}} \sup_{\theta \in R^N} \frac{\frac{1}{N} E(||\widehat{\theta} - \theta||^2)}{\frac{\sigma^2}{N} + \frac{1}{N} \sum_i min(\theta_i^2, \sigma^2)} \sim 2 \ln N.$$

We will now investigate optimality properties of our estimator $\widehat{f}_{DS}$ using the same approach.

Proposition 3.12. *For $\alpha = 0.0$ and $\lambda_{DS} = \sigma\sqrt{2\ln(2n_j + 4)}$*

$$\inf_{\widehat{\theta}} \sup_{\theta \in R^N} \frac{\frac{1}{N} E(||\widehat{\theta} - \theta||^2)}{\sigma^2 \frac{J}{2N} + \frac{1}{N} \sum_i min(\theta_i^2, \sigma^2)} \sim \frac{4}{3} \ln N.$$

Note: Compare to $R(\widehat{f}_{DS}, f) \leq (1 + 2\ln(N+4))(\sigma^2 \frac{J}{2N} + \frac{1}{N} \sum_i min(f_i^2, \sigma^2))$.

The difference between the multiple $\frac{4}{3}$ in the optimal estimator and the multiple of 2 in front of the $\ln N$ factor in our estimator is small. But can we make this difference even smaller? We like to get an upper bound for the risk function as close to the ideal as possible using an analogous "frequency-adaptive" thresholding approach.

Proposition 3.13. *In fact, if $\lambda'_{DS} = \sigma\sqrt{2\ln(\beta n_j + 4)}$, then*

$$\inf_{\widehat{\theta}} \sup_{\theta \in R^N} \frac{\frac{1}{N} E(||\widehat{\theta} - \theta||^2)}{\sigma^2 \frac{J}{\beta N} + \frac{1}{N} \sum_i min(\theta_i^2, \sigma^2)} \sim \frac{2\beta}{\beta + 1} \ln N.$$

Note: The proofs of Propositions 3.12 & 3.13 follow directly from the proof of Theorem 3.11 (Donoho & Johnstone), by replacing the factor $\frac{\sigma^2}{N}$ in the upper bound for the DJ estimator risk, by $\sigma^2 \frac{J}{\beta} \frac{1}{N}$. This is because,

$$\sum_{j=0}^{J} \frac{2^j}{(\beta n_j + 4)} \leq \sum_{j=0}^{J} \frac{2^j}{\beta 2^j} \leq \frac{J}{\beta},$$

and $R(\widehat{\theta}_{DS}, \theta) \leq (1 + 2\ln(\beta N + 4)) \left[\frac{\sigma^2}{N} \sum_{j=0}^{J} \frac{2^j}{(\beta n_j + 4)} + \Re_{N,\sigma}(SW, \theta) \right]$.

Corollary 3.14. *For $\beta << N$, but β large, our frequency-adaptive thresholding method $\lambda'_{DS} = \sigma\sqrt{2\ln(\beta N + 4)}$ induces an "almost-optimal" estimator $\widehat{\theta}'_{DS} = sgn(w)[|w| - \lambda'_{DS}]_+$ with*

$$R(\widehat{\theta}'_{DS}, \theta) \leq (1 + \ln\beta + 2\ln(N+4)) \left[\frac{\sigma^2 J}{\beta N} + \Re_{N,\sigma}(SW, \theta) \right].$$

This corollary is a direct consequence from Proposition 3.13. It simply says that for every $\epsilon > 0$ we can find β large such that the risk of the induced estimator $\widehat{f}$ is essentially bounded above by $2(\ln N)\Re_{N,\sigma}(SW, f)$, when the ideal risk has order $\frac{2\beta}{\beta+1}(\ln N)\Re_{N,\sigma}(SW, f) \geq (2-\epsilon)(\ln N)\Re_{N,\sigma}(SW, f)$.

4. Applications and Examples

4.1. Function Denoising

Example: The WT of the "Heavy-Sine" function was thresholded using 3 different schemes: Uniform Thresholding at 99% (Uniform); Donoho-Johnstone Optimality approach (DJ), and by our frequency-adaptive thresholding technique (DS).

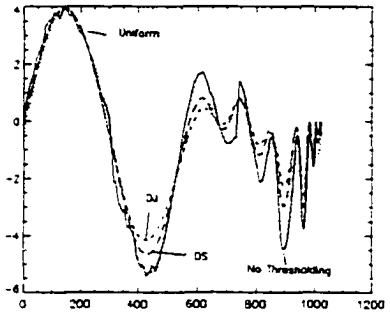


Figure 62. Differences between Uniform, DJ and DS thresholding of the wavelet coefficients.

This example shows that indeed there are some differences between the three function estimators using different wavelet thresholding approaches. From this picture, however, it is not possible to quantify estimator performance since only the ("noisy") observations (solid curve) are given and not the real data.

4.2. Classification of image-registration techniques

As we pointed earlier, the process of wavelet space thresholding can be viewed as a tool for robust and concise image representation. The particular scheme for wavelet shrinkage used in this process is constructed so that it induces "optimal" (in a minimal risk functional setting) estimators, and provides a meaningful procedure for signal denoising. We will now show how we use these "ideal" function estimators to perform image analysis in compressed wavelet space.

One of the main goals of image registration is to be able to compare groups of images and identify common regions of interest. Thus, to evaluate group performance of various warping techniques we used the "clustering group classification" (CGC) scheme proposed by DW Sumners. The idea is that we prefer alignment techniques that reduce the convex-hull of the warped data images, regardless of the target of the image-registration or the initial differences between the signals. There are two classifying functionals we used in our study; The diameter of the convex-hull of the group of warped data, and the "average-diameter" - as measured by the average pair-wise distances in reduced wavelet space - of the group of warped data. The smaller the values of the diameter/average-diameter the better the image-registration method. We now present the analysis and quantitative evaluation of three warping techniques LS (least squares), MI (mutual information) and AF (affine), see Section I.5.3.2, applied to a set of five PET stereotactic images.

The results in Table 15, and Figure 63 clearly indicate that the group-ranking of the three image-registration techniques is: MI, LS, AF (best-to-worst). This is in agreement with the Kjems *et al.* univariate warp analysis of these 3 alignment methods.

Do we gain anything (rank robustness, sensitivity increase) by using wavelets as opposed to performing the usual "time domain" analysis? To

Table 15. Results from the Cluster-Group-Classification of the LS, MI and AF warps of the PET data using three different wavelet thresholding schemes

Measure	Warp	Uniform 1%	DJ	DS
Diameter of Convex Hull	LS	10.29	9.58	9.76
	MI	9.92	9.23	9.41
	AF	11.67	10.89	11.11
Average Distance Between Pairs	LS	7.06	6.47	6.62
	MI	6.90	6.31	6.45
	AF	8.49	7.84	8.02

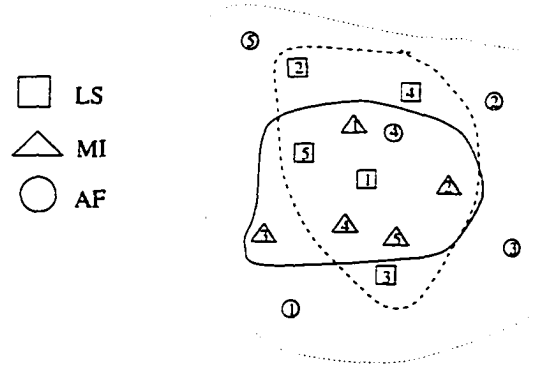


Figure 63. Planar representation of the wavelet analysis in Table 15.

answer this question we compare the image-space (spatial-domain) analysis with the compressed (under different thresholding schemes) wavelet-space results. In Tables 16 and 17, 100% refers to the the case of no thresholding which is equivalent to image-space analysis, since all wavelet coefficients (100%) are used to determine distances, and the wavelet transform is an orthogonal linear mapping that preserves distances. *DJ* and *DS* refer to "Donoho-Johnstone" and our thresholding schemes, respectively.

Even though there are no ranking differences between the spatial-domain and the wavelet-space analyses, there are small but robust sensitivity improvements in favor of the wavelet-space analysis. For example, using the

Table 16. Sensitivity increase in warp ranking of the wavelet-space analysis as opposed to the spatial domain analysis using the "diameter of the convex-hull" measure.

Warp	DJ	100%
LS	9.58	11.28
MI	9.23	11.00
AF	10.89	12.67

Warp	DS	100%
LS	9.76	11.28
MI	9.41	11.00
AF	11.11	12.67

Table 17. Sensitivity increase in warp ranking of the wavelet-space analysis as opposed to the spatial domain analysis using the "average-diameter" measure.

Warp	DJ	100%
LS	6.47	7.95
MI	6.31	7.84
AF	7.84	9.36

Warp	DS	100%
LS	6.62	7.95
MI	6.45	7.84
AF	8.02	9.36

"average-diameter" classification functional, the difference between the first (MI) and the second (LS) ranked warps in the DS wavelet study is 0.17 as opposed to the 0.11 difference between the same warps under the image-space analysis.

5. Discussion

The problem of finding good function estimators, based on the data alone, that have risk functionals close to the oracle-based ideal risks was first ap-

proached from the classical theory point of view using Fourier analysis (e.g., Efroimovich & Pinsker [1984]). Some of the inherited limitation, however, of the Fourier basis functions, as discussed in Section I.2.1, inclined Donoho and Johnstone to look at the problem of evaluating function estimators using wavelet analysis. They showed that on discontinuous functions their wavelet based spatial adaptation approach produces better results and achieves rates of convergence of the order of $\ln^2 N/N$.

Lately, Efroimovich (personal correspondence) proposed an exponentially increasing frequency-adaptive wavelet thresholding scheme which seems to perform, empirically, visually and in terms of compression, not worse (and in some cases even better) than the approach of Donoho and Johnstone.

We, on the other hand, employ a frequency-adaptive wavelet shrinkage that increases only as a power function with the increase of the frequency index of the wavelet coefficients. It is true that all of these function estimators attain risk functional of the order of $\ln^2 N/N$, but there are finer differences in the constants involved in their rates of convergence.

In contrast to the work of Donoho and Johnstone and Efroimovich we are not interested in a visual, or other, spatial domain inspection of the results of the denoising procedure. We perform our warp quality analysis in "compressed" wavelet space and the outcome of our study reports only the final image-registration ranking. Image denoising is only used as a motivation and a theoretical basis for the proposed transform space warp classification analytic approach.

CONCLUSIONS

In this work we develop, implement and test two methods for modeling and interpretation of human brain anatomical and functional data. Within the scope of our transform-based (mathematical) model we introduced a new discrete algorithm (Reverse Quadtree Partitioning) for determining of efficient fractal interpolations of signals. This method targets the best possible fractal encoding based on quadtree-type domain partitioning.

Extracting the self-similarities and the affine self-symmetries of images in fractal transform space proves to be very computationally expensive. To make the process feasible and deterministic we presented a new classification scheme that significantly reduces the computational complexity of the discrete fractal transform without limiting or affecting the essential fractal space of the data. This skewness-based classification method makes possible obtaining the DFT of 2D images within a few minutes.

The transform-based image analysis technique has wide range of applications. We engineered metrics on transforms of signals that help determining the similarities and the differences between two images and quantifying the performance of various image registration and alignment procedures. This model can also be adopted for segmentation, magnification and enhancement of images. The main advantage of our transform-based warp classification schemes is that they are completely automated and do not require human interaction in the process of quantifying image-registration.

Our second model (Sub-Volume Thresholding, SVT) is geared toward analyzing the statistical significance of regions of activation of human brain

metabolic studies. In general, human brain functional data comes with low signal-to-noise ratio. The SVT technique attempts to extract the regions of relevant (in statistical sense) brain activation in difference images, under various paradigms, for single or multi-subject studies. Prior information about the study can be incorporated into the model through selecting an appropriate topological partitioning of the search region.

Developing the SVT technique we naturally encountered a family of continuous functionals, which we prove induces a legitimate class of valid covariograms. In addition, we derived closed mathematical forms for the correction factors of the variance estimates for rectangular-type anatomical partitioning. For the cases of more complex domain partitions we introduced a scheme for stochastic approximation of these correction factors. Finally, the SVT model was verified by testing it on the well-known and understood motor-studies, and compared to a uniform simple 97% thresholding approach.

Even though the transform-based and the SVT techniques are tested only on human brain data they show promise in addressing problems associated with the analysis and interpretation of structural and functional medical data, in general.

Finally, in chapter III, we propose a new thresholding method that yields close to optimal function estimators. This scheme is different from the one introduced by Donoho and Johnstone [1994], because it uses "frequency-adaptive" thresholding nonlinearities. We showed that, under certain conditions, we can essentially produce estimators with risk measures approximately equal to the ideal risk. The model we developed is applied to quantify the group performance of a family of image registration techniques applied to collections of PET volumetric data.

APPENDIX

One of the long time goals of human kind is to be able to see what no one else has ever seen. In particular, people have been very interested in learning what is inside the human body. What is the internal structure of the body and how it functions. In our work we are primarily concerned with modeling anatomical and functional images of the human brain. But all of our models require digital 2D, 3D or 4D (time series) representations of the brain data.

There are two fundamental brain mapping techniques people are currently using: *Invasive* and *Non-invasive*. Post-mortem tissue cryosectioning and optical intrinsic signal (OIS) imaging are the main "invasive" brain imaging tools. Cryosectioning involves physical axial slicing and photography of the entire post-mortem brain. It has the highest (spatial) resolution one can hope for, however, it is not very practical because of the terminal nature of the process. In addition, one needs a fairly good algorithm for reconstructing the 3D structure of the brain from the available (unregistered) 2D snapshots. Also, physical tissue sectioning measures structural but not functional information from the sample.

OIS imaging is frequently done during neuro-surgical procedures when part of the brain can be physically exposed to white light. By photographic means one can again observe the cortical surface of the brain and track the distribution and spread of CBF (cerebral blood flow) over the 2D region of interest (ROI). If the subject's brain is sequentially stimulated under different activation paradigms one can obtain valuable functional information. The

main drawback of the OIS brain imaging is the fact that it produces 2D cortical images and it involves excision.

The "non-invasive" stereotactic brain mapping techniques include: CT (Computed (Axial) Tomography), PET (Positron Emission Tomography), SPECT (Single Photon Emission Computed Tomography), MRI (Magnetic Resonance Imaging), and fMRI (functional MRI). Generally speaking, they all have the advantage that they are reproducible and unlike the "invasive" methods do not require surgical procedures and have no long time harmful effects. On the other hand the non-invasive scanning techniques suffer from low spatial resolution and somewhat low sensitivity.

The main differences between the various brain imaging methods are in the length of the waves of the electro-magnetic signals they use, Figure 64.

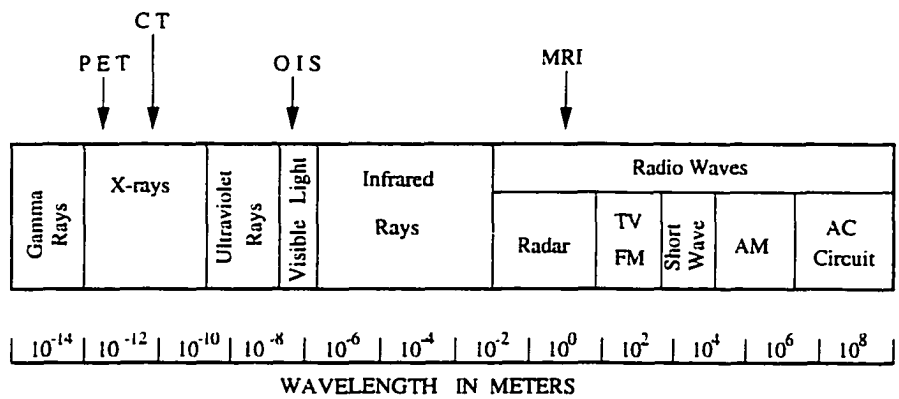


Figure 64. Electro-magnetic spectrum

For example, positron-electron annihilation (in PET/SPECT studies) produces γ-radiation (E, energy) in the 511-keV range. Using laws of physics, $\nu\ h = E$ and $c = \lambda\ \nu$, where E is the energy of 1 photon, ν is the frequency measured in $Hz = \frac{1}{\sec}$, $c = 2.9 \times 10^{8} m/\sec$ is the speed of light, λ is the wavelength and $h = 6 \times 10^{-34} J/\sec$ is the constant of Planck, we can calculate the frequency of the PET imaging, $\nu = \frac{511 \times 10^{3} eV}{2\pi \times 4.135 \times 10^{-15} eV\ \sec} = 1.97 \times 10^{19} Hz$. And therefore, the wavelength is $\lambda = \frac{c}{\nu} \approx 10^{-11} m$.

Similarly, the MRI imaging does not measure the wavelength of the radio frequencies (RF) but rather uses the RF pulses to alter the spin and the magnetic moment of the hydrogen nuclei.

We now describe the foundations of the most commonly used non-invasive brain imaging modalities: PET, MRI, fMRI.

1. Positron Emission Tomography Imaging

Positron Emission Tomography is an imaging modality which provides a unique insight into human and animal physiology and allows us to measure biochemical processes and interior organ functioning in living biological systems. PET can be used to monitor and record cerebral blood flow (CBF) and the rate at which glucose is utilized by the brain. In the mid-70's the PET imaging developed as a research tool, but it did not get widely used until the technology of the medical cyclotrons advanced dramatically in the 1980's [Saha *et al.*, 1992]. Its predecessors, the SPECT imaging instruments, were introduced in late 1950's and were applied first for studying the cerebral function [Lassen & Holm, 1992]. SPECT uses heavier radioactive isotopes (that do not require on-site production because of their long half-life) which emit a single γ-ray (photon) upon collision with an electron. By administering radioactive isotopes, like the ones listed in Table 18, to living systems, tracer spread and activity is observed, recorded and quantified.

Radiation is defined as the propagation of energy from one state to another. This is a process that naturally takes place everywhere. Nuclei that can spontaneously transform an atom of one element into another, with emission of radiation, are called *radioactive* nuclei. There are at least three forms of radiation emitted by unstable nuclei: (low energetic) α-particles, emitted by the nuclei of helium atoms; β-particles, (positively charged) nuclear

Table 18. Physical properties of commonly used positron emitting isotopes

ISOTOPES	Half-life (min)	Max Energy (keV)	Range (mm)
Carbon-11	20.4	960	0.69
Nitrogen-13	9.96	1190	0.91
Oxigen-15	2.07	1720	1.44
Fluorine-18	109.8	640	0.38

electrons; and (high energetic) γ-rays, photons. Smaller atoms are nuclearly stable if the number of protons (p^+) is approximately equal to the number of neutrons (n^o). Relative increase or decrease of either one makes the atom unstable and it naturally reconfigures its nuclear structure by emitting radiation. For example, if a heavy atom has too many protons in its nucleus one proton will eventually become a neutron. By laws of energy conservation a positively charged electron (positron) is released.

$$p^+ \longrightarrow n^o + e^+$$

Because of the short half-life of the radio-isotopes used in PET studies these tracers need to be produced on-site in the necessary amounts. A *Cyclotron* is a particle accelerator, composed of D-shaped electromagnets, that is used to produce high-energy ions by accelerating particles (like protons, H^+, or deuterons, D^{2+}) in a circular orbit. Once these high energy particles reach the necessary extraction energy (approx. 15 MeV) they are pulled out from their orbits (surrounding electrons can be stripped off by a thin carbon foil) and bombarded onto a stationary target. The collision yields the positron emitting isotope used in the PET study.

The PET camera does not actually detect the number of positrons in the sample, but rather counts the number of photons given off in the process of

positron-electron collision. Again by the law of conservation of energy the mutual annihilation of the beta-particle (positron) with a near-by electron results in the emission of two photons traveling in opposite directions from one another at nearly 180^o. These photons are detected, multiplied (using photo-multiplier tubes) and converted to electric impulses. Because each of the detectors in the tube surrounding the subject is coupled with several detectors on the opposite side, one can measure the approximate amount or the radioactive tracer in the particular axial slice. Detector D_o records a photon hit if and only if the dual photon is detected at the same time in one of the other detectors coupled to D_o. A typical PET scanner has 5-10 concentric circles of detectors 1-2 cm apart. Each circular band consists of 36 opposite pairs of detectors (5^o apart).

The PET cameras have two main identifying parameters: resolution (the ability to accurately locate positron-electron annihilation); and sensitivity (number of events registered per unit dose of isotope). Because image quality depends explicitly on the number of strikes detected, the signal-to-noise ration of the PET images is strongly influenced by the scanner sensitivity. As Table 18 showed there is a relationship between the mean-range-to-collision of the positrons and the half-life of the administered tracer. Better spatial resolution is achieved using slowly decaying positron-emitting isotopes, which may have longer effects on the subject.

A basic problem in determining the exact spatial position of the emitted pair of photons is accounting for the distance traveled by positrons before annihilation with an electron. In water, for example, the mean range of positron-travel after emission is $0.46mm$ and $1.8mm$ for Fluorine-18 and Oxygen-15, respectively. The distance traveled by the β-particles from emission to electron-positron annihilation can be modeled as a random variable

with a standard normal (Gaussian) distribution. This, along with the fact that brain physiology and activation do not exhibit sudden drastic (discontinuous) changes, justifies our spatial auto-correlation model we developed in Chapter II. Further, we can now explain the validity of out hypothesis that the voxel intensities of the difference image are normally distributed; We now show that the arrival times of simultaneous counts in the dual opposite detectors can be thought of as a large scale Poisson process and can be well approximated by a Gaussian distribution.

Binomial Distribution. Suppose the event A occurs with probability p at each trial, and it does not occur with probability $q = 1 - p$. We are interested in the number of times A happens in an $n-$trial experiment. Let X be the random variable representing the number of times the event A occurs in n trials. Then for $n = 1$ $P(X = 0) = q$ and $P(X = 1) = p$. Thus, the probability density function for X (Bernoulli trial experiment, $n = 1$) is $f_X(x) = p^x q^{1-x}$, for $x = 0, 1$. For $n-$trials we can extend this to the *pdf* (probability density function) of an $n-$trial Binomial distribution with probability of success p ($B(n, p)$). The probability of the event that $X = x$, $x = 0, 1, 2, \cdots, n$, is

$$f_X(x) = \binom{n}{x} p^x q^{n-x}.$$

Poisson Distribution. A Poisson process is a sequence of events randomly spaced in time (e.g., Geiger counter clicks). The rate μ of a Poisson process is the average number of events per unit time (over a long time). The probability of n arrivals, for one unit time interval, is $P(X = x) = f_X(x) = \frac{\mu^x}{x!} e^{-\mu}$. A basic property of the Poisson distribution is that the number of arrivals in two disjoint time intervals are independent of each other. This distribution is a convenient approximation of the binomial distribution in case of a large number of trials and small probability of success in a single trial. This is given by the following theorem [Kreyszig, 1970].

Theorem. *For a fixed x, if $n \longrightarrow \infty$ and $p \longrightarrow 0$ with $np \longrightarrow \mu < \infty$ then the pdf of $B(n,p)$ approaches $f_X(x) = \frac{\mu^x}{x!}e^{-\mu}$, as $n \to \infty$.*

Normal Distribution. A random variable X with mean μ and variance σ^2 is said to be normally distributed, $X \sim N(\mu, \sigma^2)$, if its pdf has the form

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The normal distribution can also be shown to be a useful approximation of the binomial distribution when n, the number of trials, is large [Kreyszig, 1970].

Theorem. *Let $0 < p < 1$ be the probability of success in a single (Bernoulli) trial. For large number n the pdf of the binomial distribution $B(n,p)$ can be approximated by the normal distribution with mean $\mu = np$ and variance $\sigma^2 = npq$. That is, for $x = 0, 1, 2, \cdots, n$,*

$$\binom{n}{x} p^x q^{n-x} \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(x-np)^2}{2npq}}.$$

The fact that the Poisson and the normal distributions both approximate the large scale (large number of trials, when probability of success goes to zero) binomial distribution yields that for large n the pdf's of $N(np, npq)$ and $Poisson(np)$ are close to each other.

Another inherent property of photon emission is a phenomenon known as *attenuation*. If two dual photons travel different distances and/or navigate through tissue of different (density) type, then they are unlikely to strike two dual (opposite) detectors at the same time due to attenuation. This introduces another smoothing of the image (low-pass Gaussian filtering) that decreases the spatial resolution of the PET images.

The above mentioned problems of reconstruction and sensitivity of the PET scans are outweighed by its overall usefulness as a tool providing quantitative information in biological units (grams of glucose consumed per 100 grams of tissue per minute) about human brain function and metabolism.

2. Magnetic Resonance Imaging

The Magnetic Resonance phenomenon, also known as Nuclear Magnetic Resonance (NMR), was first introduced in 1946 by two groups of researchers Bloch, Hansen and Packard [Bloch *et al.*, 1946]; and Purcell, Torrey and Pound [Purcell *et al.*, 1946]. The first MRI images of living systems, however, were not due until 1973-74 [Lauterbur, 1973, Mansfield & Grannell, 1973].

MRI is a powerful, high spatial resolution, non-invasive imaging technique, that, unlike the computed tomography (CT) and PET scanning, requires no ionizing radiation. There are many tissue parameters that affect the MR signals. The two most significant parameters, the T_1 and T_2 relaxation times, cover a wide range of values for various types of tissue. Signal acquisition parameters can be manipulated by the investigator in a variety of ways allowing control over the contrast characteristics of the MRI image. Besides the many advantages the MRI technology offers there is one main limitation; poor sensitivity. Because of the physical laws that govern the equilibrium of the magnetic moments, the signal level in MR data is low and depends on the strength of the background magnetic field. For human studies field strength of about 1.5 - 7.0 T (Tesla) are used. For comparison, the Earth's magnetic field is approximately 0.5 G, and 1T = 10,000 G.

According to the theory of quantum mechanics of atomic structures most nuclei posses a property called *spin angular momentum*, which is the basis of nuclear magnetism. Because some atomic nuclei are positively charged, the spinning motion causes a magnetic moment which is co-linear with the spin axis. The strength of this magnetic moment is a property of the type of nucleus and thus determines the sensitivity of the MRI image. In the absence of external magnetic field the nuclear magnetic moments are ran-

domly oriented in space. However, if we apply an external magnetic field, F_o,the nuclear magnetic moments align (almost) parallel (lower energy state) or anti-parallel (higher energy state) to F_o. Since the energy difference between the two states is small, ambient thermal energy causes the two states to be approximately equally populated (at about 75 F^o the population ratio is about 100,000 to 100,006). It is only the net nuclear magnetization, arising from this small population difference, that accounts for the signal detected by the MRI device. Researchers in biology, physics and chemistry frequently acquire microscopic MRI images to determine the structure of DNA molecules, various crystallized proteins and viruses. How do they obtain such highly sensitive NMR representations in microscopic detail? The answer comes from the fact that at temperatures close to the absolute zero (0 K^o) all of the nuclear spin magnetic moments of the atoms will align parallel to the field (low-energy state). Because of that performing NMR to crystallized samples at low temperatures produces strong (parallel versus anti-parallel) signal and high resolution images. This, of course, is not applicable for human and other living system studies, which imposes a limitation on the sensitivity characteristics of the MRI technology.

The spin magnetic moments of the nuclei precess around the external field F_o, because their spin-axes are not exactly parallel or anti-parallel to F_o, but rather at a small angle, Figure 65.

A spinning top on the ground can be thought of as a model for this precessive motion, where the external field is the Earth's gravity. The top would almost never stand exactly vertical, but will precess about the vertical axis of the gravity force.

The precessional frequencies of the individual spins around the magnetic

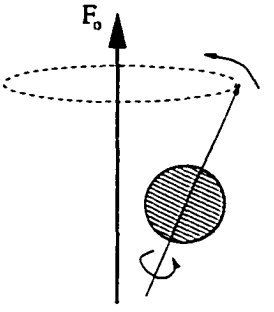


Figure 65. Precession of the nuclei spin-axis around the direction of the external field.

field are given by the *Larmor* equation

$$\gamma F_o = \nu,$$

where ν is the precessional frequency, F_o is the strength of the external magnetic field and γ is a constant determined by the type of nucleus and its magnetic moment. For instance, hydrogen nuclei have $\gamma_H = 4257 Hz/Gauss$. Therefore, in a 1.5 T magnetic field the frequency $\nu = 15,000 Gauss \times 4257 Hz/Gauss = 64 MHz$. Indeed, a very fast precession rate.

To detect a signal a resonance condition (alternating absorption and dissipation of energy) is established by a radio-frequency (RF) perturbation. Applying a RF impulses at the Larmor frequency ν yields transition of spins between the two (high and low) energy states. This effects only the nuclei having precession frequency equal to ν. The RF radiation can be regarded as another external magnetic field F_1 perpendicular to F_o, along a given axis in the plane transverse to F_o. If the RF is on for a short period of time this would rotate the net magnetization (of the particular nuclei having precessional frequency ν) by a certain angle away from the axis of the field F_o. This angle is called *flip angle* and it is proportional to the duration of the RF and its amplitude. If the net magnetization is tilted away from F_o its precession about F_o would induce a small alternating current (AC) that

can be detected. (Recall that regular car alternators generate electricity by rotating a "rotor" inside magnetic field.) As time goes on (in milliseconds) the net magnetization tends to align back in the direction of F_o and the AC current decreases. This process, known as *relaxation*, can be modeled as an exponential decay. If M_o is the initial transverse magnetization (immediately after the RF), then the magnetization at time t is

$$M_t = M_o e^{-\frac{t}{T_2}},$$

where T_2 is a (spin-spin) relaxation time that characterizes the decay of the net (transverse) magnetization to F_o. Therefore, at time T_2, $M_{T_2} = M_o e^{-1} \approx 0.37 \times M_o$, the net transverse magnetization has decayed to 37% of its initial value.

The reason of the observed decay of transverse magnetization is the fact that different components of the magnetization may precess at slightly different rates. This is known as transverse plane *dephasing*. The main (but not the only) cause for dephasing is the inhomogeneity of the background field F_o. Spins at different locations are not exposed to exactly the same magnetic field F_o. This in turn yields a range of Larmor frequencies. The results of various precessional frequency components of the net transverse magnetization, M_o, is a spread over time, of its components, which leads to loss of phase coherence and self cancelation of the signal.

Simultaneously to the spin-spin relaxation another process restores the "longitudinal" net magnetization after the RF pulse. If M'_t is the longitudinal magnetization at time t, $M'_o = 0$ since the net magnetization is in transverse plane immediately following the RF. It gradually increases with time to its initial value M_o, before the RF went on. We can calculate M'_t using the exponential model

$$M'_t = M'_o \left(1 - e^{-\frac{t}{T_1}}\right),$$

where T_1 is the *spin-lattice* relaxation time required to decrease the difference between the current value of M_t' and the equilibrium value M_o by a factor of $(1 - \frac{1}{e}) = 63\%$.

The percent concentration of water molecules in a given tissue type would effect the MRI signal, since smaller amounts of hydrogen would show shorter relaxation times than anatomy with higher hydrogen concentrations. Efficiency of relaxation is also influenced by the strength of the background magnetic field, F_o, because the relation between the Larmor frequency ν and the difference between the two energy states $\Delta E = E_2 - E_1$. The cumulative energy parallel and anti-parallel spins are denoted by E_1 and E_2, respectively. The higher the magnetic field, F_o, the higher the Larmor frequency ν and in turn larger ΔE is supplied to the system, which yields increased sensitivity and stronger signal.

Finally, complex 2D and 3D image reconstruction techniques and Fourier analysis are employed to find the spatial location of the AC signals and reconstruct the stereotactic (volumetric) image [Horowitz, 1995, Toga 1996].

3. Functional Magnetic Resonance Imaging

The first fMRI (functional magnetic resonance imaging) scan of brain metabolism was done in 1991 by Belliveau and his colleagues [Belliveau *et al.*, 1991]. They initially used echo-planar techniques to measure the amount of exogenous contrast enhancement injected into the subject. Later, the same group used gradient-echo and spin-echo inversion recovery fMRI to examine the hemoglobin deoxigenation and blood flow rate [Kwong *et al.*, 1992].

The fundamental principle of fMRI imaging is to quickly acquire (40 msec) a series of 2D slices of the sample, with inner-plane spatial resolution of about 1 mm. It takes a few seconds to produce a 3D volumetric data set

of say $128 \times 128 \times 120$ voxels. Typically, every fMRI scanning session involves several sub-sessions, each being a temporal sequence of approximately 12-15 alternating stimulus and rest paradigms. Thus, the intensity at a single voxel may vary as shown on Figure 66, where the dashed lines separate the times of the two activation conditions.

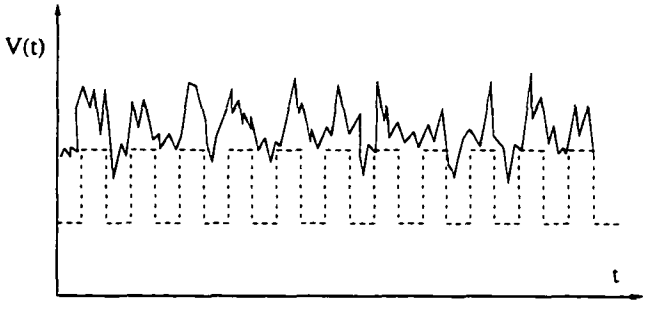


Figure 66. Time series representation of the change of intensities at a single voxel location due to alternating stimulus/rest conditions in fMRI.

Neural activity causes an increase in regional Cerebral Blood Flow (rCBF) to compensate for the increase in metabolic activity. The body actually over-compensates for the increased metabolic activity by providing in excess oxygenated hemoglobin in active brain tissue.

The deoxygenated hemoglobin is paramagnetic while oxyhemoglobin is diamagnetic. Thus, the MRI technology can discriminated between the two types of blood. In a region of activation we have more oxygenated blood than we did before the activation started. This results in a net decrease of paramagnetic material in the active cerebral tissue. Therefore, we get a net increase in the signal for the activated areas due to less dephasing of the signal.

REFERENCES

Ayache, N., Faverjon, B., "Efficient registration of stereo images by matching graph descriptions of edge segments", INRIA, inv. no. 119103, 1986.

Adler, R.J., "The Geometry of Random Fields", Wiley, 1981

Bajcsy, R. Kovacic, S., "Multiresolution Elastic Matching," Computer Vision, Graphics, and Image Processing, 46(1989), 1-21.

Barnsley, M., "Fractals Everywhere", Academic Press, San Diego, 1988.

Barnsley, M., Hurd, L., "Fractal Image Compression", 1993, AK Peters, Ltd.,

Belliveau, J.W., Kennedy, D.N.Jr., McKinstry, R.C., Buchbinder, R.R.,

Weisskopf, R.M., Cohen, M.S., *et al.*, "Functional Mapping of the Human Visual Cortex by Magnetic Resonance Imaging", Science, 254, 716-719, 1991.

Billingsley, P., "Probability and Measure", Wiley and Sons, 1986.

Bloch, F., Hansen, W.W., Packard, M., "Nuclear Induction", Phys. Rev., 69, 127, 1946.

Bookstein, F., "Morphometric Tools for Landmark Data", Cambridge University Press, New York, 1991.

Christacos, G., "On the Problem of Permissible Covariance and Variogram Models", Water Resources Research, vol. 20, No. 2, Feb. 1984.

Christensen, G., Rabbitt, R., Miller, M., "A deformable neuroanatomy textbook based on viscous fluid mechanics", Invited paper. In Prince and Runolfsson, editors, Proceedings of the 1993 Conference on Information Sciences and Systems, pp. 211-216, Johns Hopkins University, March 24-26, 1993.

Cohen, M.S., Bookheimer, S.Y., "Localization of Brain Function using Magnetic Resonance Imaging", Techniques in Neuroscience, 17(7), 268-277, 1994.

Cressie, N., "Statistics for Spatial Data", Wiley, 1991.

Daubechies, I., Communications of Pure and Applied Mathematics, vol. 41, pp. 909-996, 1988.

Dinov, I.D., Sumners, DW, " A Fractal Transform Approach for Studying the Human Brain", Proc. of the ImageTech Conf., GaTech, Atl., GA, March,

1996.

Donoho, D.L. & Johnstone, I.M., "Ideal Spatial Adaptation by Wavelet Shrinkage", Biometrika, 81, 1994.

Duncan, R.C., Knapp, R.G., Miller, M.C., "Introductory Biostatistics for the Health Sciences", Wiley and Sons, 1977.

Efroimovich, S.Y. & Pinsker, M.S., "A learning Algorithm for non-parametric filtering" (in Russian), Automat. i Telemeh., 11, pp. 58-65, 1984.

Evans, A.C., Collins, D.L., Holmes, C.J., "Automatic 3D Regional MRI Segmentation and Statistical Probabilistic Anatomical Maps", in Quantification of Brain Function Using PET, Myers *et al.* (eds.), Chapter 25, pp. 123-130, Academic Press, 1996.

Falconer, K., "Fractal Geometry - Mathematical Foundations and Applications", John Wiley & Co., 1990.

Federer, H., "Geometric Measure Theory", Springer-Verlag New York Inc., 1969.

Fisher, Y., "Fractal Image Compression", Springer-Verlag New York Inc., 1995.

Folland, G., "Real Analysis: Modern Techniques and their Applications", John Wiley, 1984

Friston, K.J., Frith, C.D., Liddle, P.F., Dolan, R.J., Lammertsma, A.A.,

Frackowiak, R.S.J., "The relationship between the global and local changes in PET scans", J. of Cereb. Blood Flow and Metabolism, 1990.

Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S.J., "Comparing functional (PET) images: The assessment of significant change", J. of Cereb. Blood Flow and Metabolism, 1991.

Gallian, J.A., "Contemporary Abstract Algebra", D.C. Health & Co., 1986.

Horowitz, A.L., "MRI Physics for Radiologists: A Visual Approach", (3-rd ed.), New York, Springer-Verlag, 1995.

Hutchinson, J.E., "Fractals and self similarity", Indiana Univ. Mathematics Journal, Vol. 30, 1981 (pp. 713 - 747).

Ida, T., Sambonsugi, Y., "Image Segmentation Using Fractal Code", IEEE Trans. on Circuits and Systems for Video Tech., Dec., 1995

Kjems, U., Strother, S.C., Anderson, J., Law, I., Hansen, L.K., "A New 3D

Non-Linear Brain MRI Registration Algorithm Improving Functional [^{15}O] Water PET Registration", to appear in IEEE Trans. on Medical Imaging.

Kreyszig, E., "Introductory Mathematical Statistics", John Wiley, 1970.

Kwong, K.K, Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskopf, R., Poncelet, B.P., *et al.*, "Dynamic Magnetic Resonance Imaging of Human Brain Activity During Primary Sensory Stimulation", Proc. Nat. Acad. Sci., USA, 89, 5675-5679, 1992.

Lauterbur, P.C., "Image Formation By Reduced Local Interactions: Examples Employing Nuclear Magnetic Resonance", Nature, 242, 190, 1973.

Mallat, S., "A Theory of Multiresolution Signal Decomposition: The Wavelet Representation", IEEE, Trans. on Patt. Anal. & Machine Intelligence, vol. 11, No. 7, July 1989.

Mandelbrot, B.B., "The Fractal Geometry of Nature", W.H. Freeman & Co., N.Y., 1982

Mansfield, P., Grannell, P.K., "NMR 'Diffraction' in Solids?", J. Phys. C.: Solid State Phy., 6, L422, 1973.

Matern, B., "Lecture Notes in Statistics", Springer, No. 36, 1986.

Miller, L.H., "C Programming Language", Wiley and Sons, 1987.

Milnor, J., "Morse Theory", Princeton Univ. Press, 1963.

Peitgen, H.O. et al., "Chaos and Fractals", Springer-Verlag New York Inc., 1992

Press, W.H. et al., "Numerical Recipes in C", Cambridge Univ. Press, 1992.

Purcell, E.M., Torrey, H.C., Pound, R.V., "Resonance Absorption by Nuclear Magnetic Moments in a Solid", Phys. Rev. 69, 37, 1946.

Rottenberg, D.A., et al., "Abnormal Cerebral Glucose Metabolism in HIV-1 Seropositives with and without Dementia", J.Nucl. Med., 1996.

Santalo, L.A., "Integral geometry and Geometric Probability", Adison-Wesley, 1956.

Schildt, H., "Using Turbo C", McGraw-Hill, 1988.

Schneider, D.M., "Linear Algebra", Macmillan, Inc., 1987.

Stevens, R., "Fractal Programming in C", M & T Publishing Inc., 1989.

Thompson, P., Toga, A.., "A Surface based Technique for Warping Three-Dimensional Images of the Brain", IEEE Trans. on Medical Imaging, Vol.

15, No. 4, Aug. 1996.

Toga, A., (ed.)., "Brain Mapping: The Methods", Academic Press, Inc., 1996.

Woods, R.P., Cherry, S.R., Mazziotta, J.C., "Rapid Automated Algorithm for Aligning and Reslicing PET Images", J. of Comp. Asst. Tomography, 16, 1992.

Woods, R.P., Mazziotta, J.C., Cherry, S.R., "MRI-PET registration with automated algorithm", J. of Comp. Asst. Tomography, 17, 1993.

Worsley, K., "A Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain", J. Cerebral Blood Flow and Metabolism, vol. 12, No. 6, 1992.

Worsley, K., "Local Maxima and the Expected Euler Characteristic of Excursion Sets of χ^2, F and t Fields", Adv. Appl. Prob., 26, 1994.

Worsley, K., "Quadratic Tests for Local Changes in Random Fields with Applications to Medical Images", Technical Report, Dept. of Math & Stat, McGill University, Canada, Aug. 1994.

Worsley, K., "The Geometry of Random Fields", Chance, vol. 9, No. 1, 1996.

Zeider, E., "Nonlinear Functional Analysis and its Applications", Springer-Verlag, 1985.

BIOGRAPHICAL SKETCH

Ivaylo (Ivo) Dinov was born on Sept. 19, 1968 in Sofia, Bulgaria. Between 1987 and 1991, as a mathematics/computer science undergraduate student, he attended Sofia University, Sofia. During these four years he was a fellow at the Institute for Micro-Processing Techniques and Technologies in Sofia.

Ivo received a Master's of Science degree in mathematics from Michigan Technological University, Houghton, MI, USA, where he was a graduate student and a teaching assistant between 1991 and 1993. He wrote a Master's project on Bochner Integrals and Vector Measures under the guidance of Prof. Kenneth Kuttler.

For the past five years (1993-1998) Ivo was a dually enrolled graduate student at Florida State University. He completed the requirements for a Master's degree in statistics and a Ph.D. in mathematics in April 1998. During his graduate studies at FSU Ivo has been a teaching and research assistant in mathematics and a predoctoral fellow in industrial engineering. He worked, under the supervision of Prof. De Witt Sumners (mathematics), Prof. Fred Huffer (statistics) and Prof. Samuel Awoniyi (industrial engineering), on various projects including modeling human brain anatomical and functional data, and developing algorithms for function optimization and solving non-linear inequalities.

As an FSU graduate student Ivo has visited and worked with researchers from several other research institutions. Among them are the Institute for Mathematics and its Applications (IMA) at the University of Minnesota,

the Veteran's Affairs Medical Center (VAMC) in Minneapolis, the Medical School at the University of Chicago, and the Laboratory of Neuroimaging (LONI) and the Brain Research Institute (BRI) at UCLA.

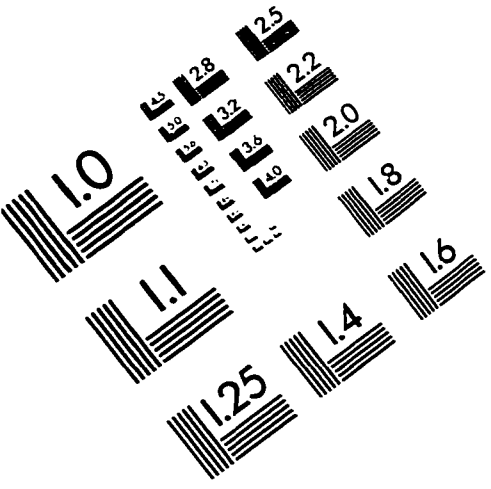
Ivo's interests include engineering, implementing and testing mathematical and statistical methods for modeling natural phenomena and problems arising in science, business and economy.

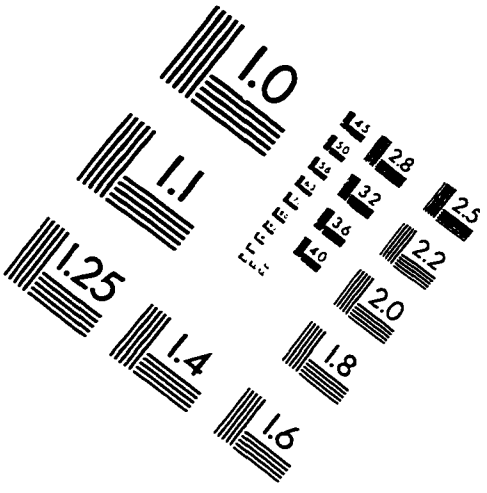
For the past five years (1993-1998) Ivo has been a member of the American Mathematical Society, Pi Mu Epsilon National Mathematics Honorary Society, FSU Congress of Graduate Students and FSU Stochastic Inference Group.

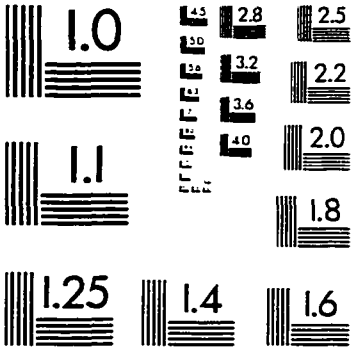
He received a 1996-97 Dissertation Research Grant, from the Office of Research and COGS at FSU.

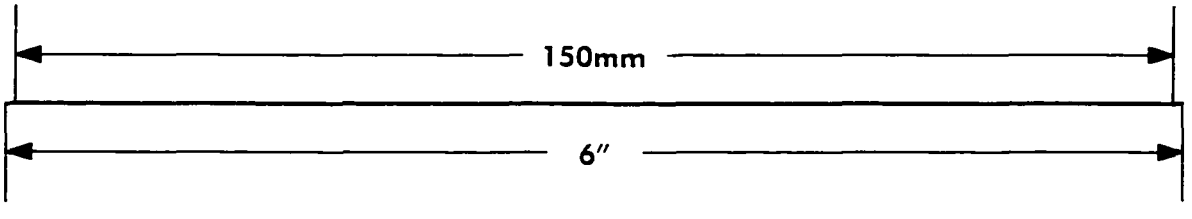
Ivo's extracurricular activities include organizing the FSU Mathematics Graduate Student Seminar (1996-97), judging in the science category at the Capital Regional Science Fair for high school students (1994-97) and officiating and coaching the FSU Water Polo Club (1993-98).

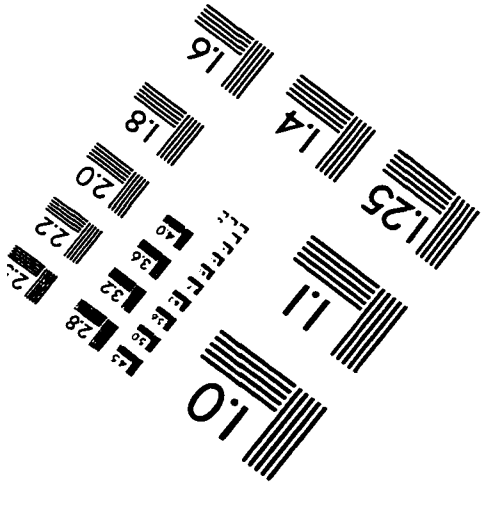
IMAGE EVALUATION TEST TARGET (QA-3)

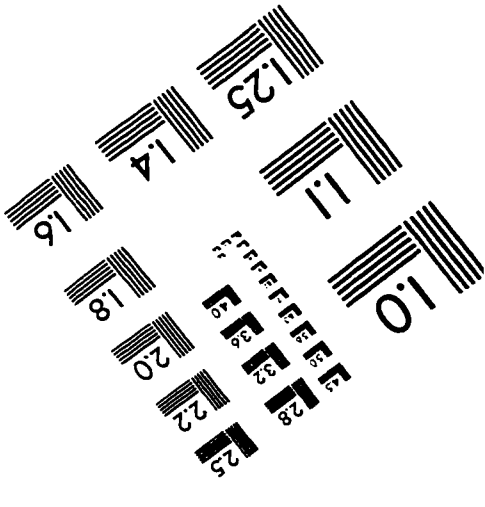












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989