November 1996

LIDS-P-2371

Research Supported By:

ARO grant DAAL03-92-G-0115
ARPA grant F49620-93-1-0604
AFOSR grant F49620-95-1-0083
AFOSR grant F49620-96-1-0455
MURI grant GC123913NGD
French Consulate in Boston

HIGH RESOLUTION PURSUIT FOR FEATURE EXTRACTION

Seema Jaggi, William C. Karl, Stephane Mallat, Alan S. Willsky

High Resolution Pursuit for Feature Extraction[1]

Seema Jaggi[2], William C. Karl[3]

Stéphane Mallat[4], Alan S. Willsky

[1]This research was conducted with support provided in part by ARO under grant DAAL03-92-G-0115, ARPA under grant F49620-93-1-0604, AFOSR under grants F49620-95-1-0083 and F49620-96-1-0455, MURI under grant GC123913NGD, and the French Consulate in Boston.

[2]S. Jaggi and A. S. Willsky are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139.

[3]W. C. Karl is with the Department of Electrical and Computer Engineering and the Department of Biomedical Engineering, Boston University, Boston, MA.

[4]S. Mallat is with Ecole Polytechnique, Paris, France and Courant Institue, New York, New York.

High Resolution Pursuit for Feature Extraction

Contact Author : Seema Jaggi
77 Massachusetts Avenue, Room 35-427
Cambridge, MA 02139
Phone : 617-253-3816
Fax : 617-258-8553
Email: jaggi@mit.edu

Abstract

Recently, adaptive approximation techniques have become popular for obtaining parsimonious representations of large classes of signals. These methods include method of frames, matching pursuit, and, most recently, basis pursuit. In this work, high resolution pursuit (HRP) is developed as an alternative to existing function approximation techniques. Existing techniques do not always efficiently yield representations which are sparse and physically interpretable. HRP is an enhanced version of the matching pursuit algorithm and overcomes the shortcomings of the traditional matching pursuit algorithm by emphasizing local fit over global fit at each stage. Further, the HRP algorithm has the same order of complexity as matching pursuit. In this paper, the HRP algorithm is developed and demonstrated on 1D functions. Convergence properties of HRP are also examined. HRP is also suitable for extracting features which may then be used in recognition.

1 Introduction

Recently, adaptive approximation techniques have become popular for obtaining parsimonious representations of large classes of signals. In these adaptive approximation techniques, the goal is to find the representation of a function f as a weighted sum of elements of from an overcomplete dictionary. That is, f is represented as

$$f = \sum_{\gamma \in \Gamma} \lambda_\gamma g_\gamma \tag{1}$$

where the set $\{g_\gamma | \gamma \in \Gamma\}$ spans the space of possible functions but is redundant. Many possible representations of f exist in this redundant dictionary. Several methods have been suggested to find the "optimal" representation of the form of (1). These methods include method of frames [5], best orthogonal basis [4], matching pursuit [13], and, most recently, basis pursuit [3]. The definition of "optimal" is application dependent.

For this work, the application of interest is feature extraction. Feature extraction from one and two-dimensional signals is an important step in object recognition. Object recognition has applications in many varied fields including military, medical, and industrial. Object recognition based on template-matching is performed by comparing a given data signal to a set of model signals and determining which model signal the data signal most closely resembles. To do this, significant *features* are extracted from both the object and the templates, and recognition is performed by comparing these object and template features. In two dimensions, some of the imaging modalities that have been considered include

visual images [12], range images [9], MRI scans, and line drawings [2]. Template-matching object recognition in two dimensions has employed many different types of features such as edges [8], moments [1, 16], and curvature extrema [15]. In one dimension, the modalities under consideration include inverse synthetic aperture radar [14].

Thus, for our work, the "optimal" representation would be one which is sparse, hierarchical, stable, quickly computable, and physically interpretable. A sparse representation is one in which a minimum number of dictionary elements are used to represent any function. In particular, if a function is synthesized as the sum of dictionary elements, the "optimal" adaptive approximation representation would be precisely those elements used to construct the signal. In other words, the representation should be sparsity preserving. There should also exist a hierarchy in the representation so that a coarse approximation may be obtained by using only the most important elements in the sum (1). The representation should be stable so that small perturbations in the underlying signal do not drastically change the representation in (1). The signal representation should be obtained as efficiently as possible. Finally, we would like the terms in the sum (1) to be physically interpretable because of our underlying feature extraction motivation. To illustrate this point, consider the sample signal shown in Figure 1. This signal is a high resolution radar return from a Cessna 310 aircraft. Each of the peaks in the signal are related to physical features of the plane such as the joint between the wing and the fuselage or the tip of the nose [17]. The location and scale (width) of these peaks are directly related to physical attributes of subparts of the aircraft. Thus, for a signal such as the one shown in Figure 1, one example of a physically interpretable

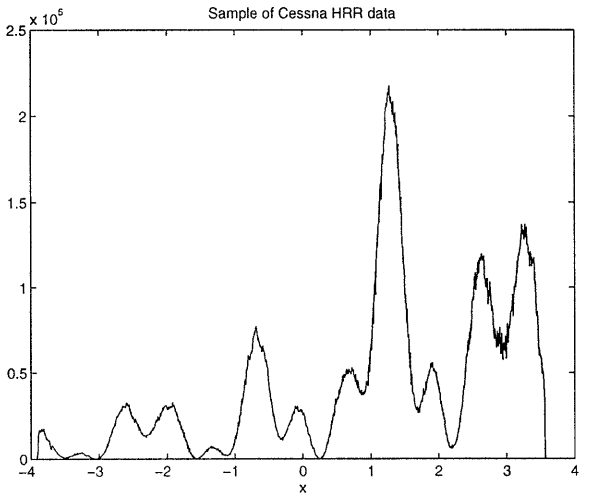


Figure 1: Sample of Cessna high resolution radar profile.

representation is one where each term of (1) corresponds to one of the peaks of the signal. In general, a physically interpretable representation is one in which each term of (1) relates directly to the geometric (e.g. size and location of subparts) characteristics of the function.

Existing adaptive approximation techniques do not always yield representations with the desired characteristics. An in-depth comparison of existing adaptive approximation techniques is given in [3]. To summarize the results of this comparison, the method of frames tends towards solutions which are not sparsity preserving and is unable to resolve closely spaced features. Best orthogonal basis also has problems preserving sparsity. Matching pursuit is also unable to resolve closely spaced features. That is, matching pursuit is unable to super-resolve features. This results from the fact that matching pursuit is a greedy algorithm which favors global over local fit. Finally, basis pursuit produces representations which preserve sparsity and resolve closely spaced features, but is computationally complex. In basis pursuit, the optimal solution is defined to be the one which minimizes the ℓ^1 norm

of the coefficients, λ_γ, in (1). To find this optimal solution, the minimization problem is translated to an equivalent large scale linear program, which is known to be computationally complex. Both matching pursuit and basis pursuit will be further explored in Section 2.

In light of the desired representation characteristics outlined above, an alternative to existing function approximation techniques is developed in this paper. This new technique, high resolution pursuit (HRP), is an enhanced version of the matching pursuit algorithm. HRP was developed to overcome the shortcomings of the traditional matching pursuit algorithm by emphasizing local fit over global fit without significantly increasing the computational complexity of matching pursuit. This paper concentrates on the development of HRP in one dimension.

This paper is organized as follows. Section 2 summarizes two adaptive approximation schemes : matching pursuit and basis pursuit. Section 3 describes the HRP algorithm and discusses convergence issues. Section 4 presents numerical examples on simulated and real data. Section 5 develops and demonstrates the HRP algorithm using a wavelet packet dictionary and compares HRP performance with that of basis pursuit.

2 Adaptive Approximation of Signals

In this section, a brief description of relevant adaptive schemes for signal approximation is presented. In particular, the two schemes that will serve as relevant background for this work are matching pursuit [13] and basis pursuit [3].

The following definitions will be used throughout the paper. Let f be a signal in a Hilbert

space $\mathcal{H}$. Let $\{g_\gamma | \gamma \in \Gamma\} = \mathcal{D}$ be a set of dictionary vectors with $\|g_\gamma\| = 1$ for all $g_\gamma \in \mathcal{D}$. Note, that such dictionaries generally include functions with a wide range of time-frequency characteristics. Thus, prior knowledge may be incorporated in the construction of the dictionary to yield the best signal decomposition. Further, this dictionary will be redundant (e.g. a dictionary that contains a wavelet frame). The function f will be decomposed as the weighted sum of dictionary elements as in (1). The signal representation is then given by

$$f = \sum_{i=0}^{n-1} \lambda_i g_{\gamma_i} + R^n f \tag{2}$$

where $R^n f$ is the residual in an n-term sum. Often, we choose to approximate f by the n-term sum in (2).

2.1 Matching Pursuit

Matching pursuit (MP) is a recursive, adaptive algorithm for signal decomposition [13]. The matching pursuit algorithm builds up the signal representation one element at a time, picking the most contributive element at each step. The element chosen at the n-th step is the one which minimizes $\|R^n f\|$ as defined in (2). In particular, the residual at stage n is given by

$$R^n f = R^{n-1} f - \lambda_n g_{\gamma_n} \tag{3}$$

where

$$\lambda_n = < R^{n-1}f, g_{\gamma_n} > \tag{4}$$

$$g_{\gamma_n} = \arg\max_{g_\gamma \in \mathcal{D}} | < R^{n-1}f, g_\gamma > |. \tag{5}$$

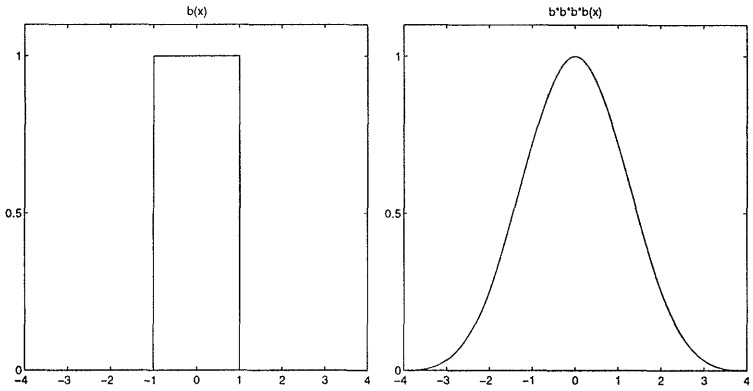
Thus, the element which minimizes $\|R^n f\|$ is the one which maximizes $| < R^{n-1}f, g_\gamma > |$. In other words, the standard inner product is used as the measure of similarity between the function and the dictionary elements and the "most similar" element is chosen at each stage. Note that the element which maximizes the similarity measure, $| < R^{n-1}f, g_\gamma > |$, is the same one which maximizes $\|R^{n-1}f - R^n f\|$. In Section 3, we describe an analogous interpretation of the HRP similarity measure. Specifically, the element which maximizes the HRP similarity measure is shown to be the one which maximizes $\|R^{n-1}f - R^n f\|$ subject to a set of constraints. The MP algorithm yields a cumulative decomposition of

$$f = \sum_{i=0}^{n-1} < R^i f, g_{\gamma_i} > g_{\gamma_i} + R^n f \tag{6}$$

and a cumulative energy equation of

$$\|f\|^2 = \sum_{i=0}^{n-1} | < R^i f, g_{\gamma_i} > |^2 + \|R^n f\|^2 \tag{7}$$

The MP approach works well for many types of signals. It has been shown to be especially useful for extracting structure from signals which consist of components with widely varying



(a) A box spline, $b(x)$.

(b) A cubic box spline, $b * b * b * b(x)$.

Figure 2: Box Splines.

time-frequency localizations [13]. MP is a greedy algorithm in the sense that the element chosen at each step is the one which absorbs the most remaining energy in the signal. In practice, this results in an algorithm that sacrifices local fit for global fit and thus, is unable to meet our feature extraction goals.

To illustrate this drawback in MP, consider the following example constructed using cubic b-splines. Note that a cubic b-spline $g(x)$ (Figure 2b) can be obtained by convolving a box spline $b(x)$ (Figure 2a) with itself three times. Scaled versions of this cubic b-spline are of the form $g(2^j x)$. As $j \to +\infty$, the cubic b-splines become finer in scale and approach Diracs. A cubic b-spline function at scale j and translation t will be denoted $g_{j,t}$, or, equivalently, g_γ where γ is a joint index over scale and translation, $\gamma = (j, t)$.

The twin peaks function, f, illustrated in Figure 3, is the sum of two cubic b-splines at the same scale but different, nearby translates. Let the dictionary $\mathcal{D}$ consist of cubic

b-splines at a wide range of translates and scales, including those used to construct f. This dictionary is well suited for the signal under consideration. For the twin peaks example, the first element chosen by MP is one which does not match either of the two functions which are the true components of f. This is illustrated in Figure 3 which shows the original function and the first element chosen by MP, g_{γ_0}. The projection graph in Figure 4 gives us more insight into the behavior of MP for this case. The proximity of the two components of f leads to a maximum of the similarity function (the inner product) which is not at the correct translation and scale of either element. The first MP residual is shown in Figure 5. The residual has a large negative component at $t = 0$ where the original function was positive. Thus, instead of finding significant features of the signal, MP has effectively introduced new "non-features" which the algorithm will have to account for by fitting additional elements. This problem is further compounded as subsequent elements are chosen by MP in an effort to correct the initial mistake. Figure 6 shows the first ten elements chosen by MP to represent f. Here, note that the elements chosen by MP do not correspond to the physical features of the function. In fact, many of these are "non-features" which only serve to correct mistakes from previous stages.

2.2 Basis Pursuit

The basis pursuit (BP) principle [3] is to find the decomposition given in (1) which minimizes the ℓ^1-norm of the coefficients λ_n. The examples presented in [3] indicate that basis pursuit yields decompositions which are sparse and show super-resolution. Thus, they do not exhibit

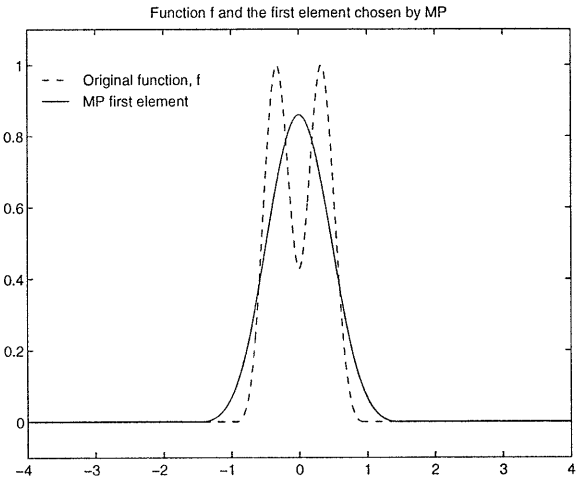


Figure 3: The twin peaks function and first element chosen by MP.

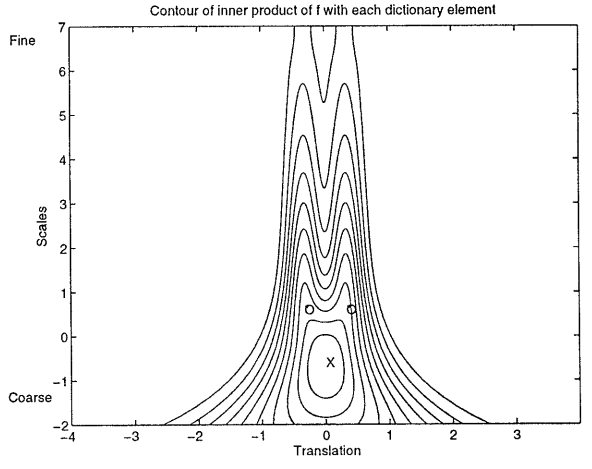


Figure 4: The projection graph is the inner product of the function with each dictionary element which is indexed by scale and translation. This figure shows the contour of the projection graph. X marks maximum inner product. O marks location of true elements of function.

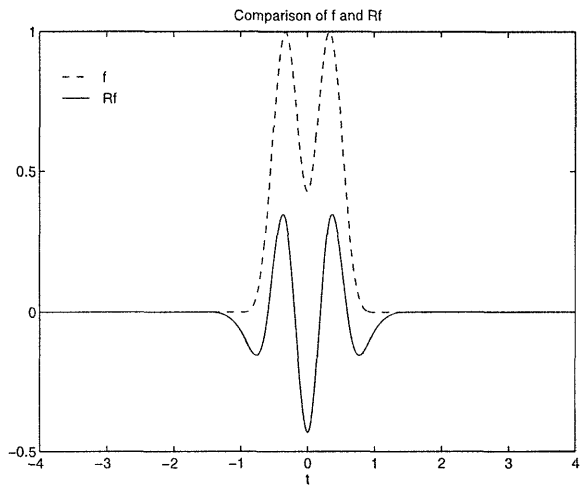


Figure 5: First residual generated by MP.

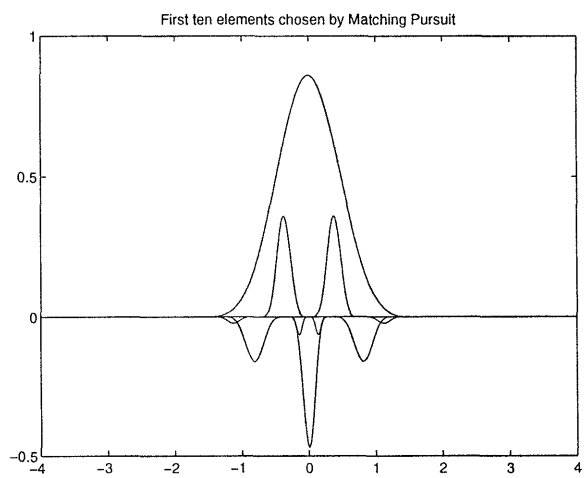


Figure 6: The first ten elements picked by MP.

problems in picking out the two adjacent cubic b-splines in the twin peaks example. An important drawback in the implementation of BP is that of computational complexity. Since basis pursuit decompositions are based on solving a large-scale optimization problem, there exist examples where the decomposition may not be completed in a reasonable amount of time, as stressed in [3]. Two algorithms are proposed in [3] to implement the basis pursuit principle : the simplex method and interior point methods. For a signal of length P and a dictionary of Q elements, the BP principle implemented using the simplex method requires an average of $\mathcal{O}(Q^2P)$ calculations, though it could require as many as $\mathcal{O}(2^P - 1)\mathcal{O}(QP)$ calculations. The complexity of interior point methods depends on the implementation. Interior point methods are typically polynomial in Q and P [6,7]. Thus, the implementation of basis pursuit is computationally intensive.

3 High Resolution Pursuit

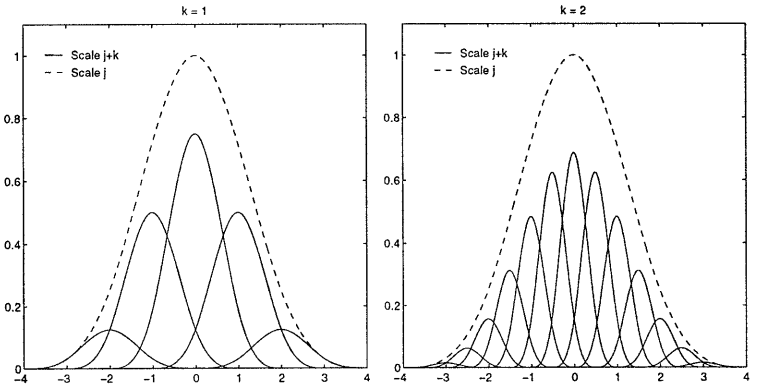
The objective of high resolution pursuit is to combine the computational speed of MP and the super-resolution of BP. The HRP algorithm, developed in this section, consists of the same procedure as MP and in fact has the same computational complexity as MP (see Section 6). In contrast to MP, HRP employs a similarity measure which emphasizes local fit over global fit, and is thus able to achieve super-resolution similar to that exhibited by BP. In this section, the HRP algorithm is developed and convergence issues are addressed.

3.1 The HRP Algorithm

In this section, the HRP algorithm, which parallels the MP algorithm, is developed. First, a new, more locally-sensitive similarity measure is proposed. This new similarity measure is proposed based on intuition derived from cubic b-spline dictionaries, but is easily extended to other dictionaries. Second, the HRP algorithm is outlined. The basic HRP procedure is to choose the element which maximizes the new similarity measure at each step. Third, to gain additional insight into the HRP algorithm, we discuss an alternative interpretation of HRP as a constrained maximization of $\|R^{n-1}f - R^n f\|$. This development is analogous to the MP algorithm development where the element which maximized the inner product similarity measure was shown to be the one which maximized $\|R^{n-1}f - R^n f\|$ without constraints.

Let us begin by developing our intuition about the MP similarity measure using cubic b-spline dictionaries. For the case of cubic b-spline dictionaries, the inner product (the MP similarity measure) of f with dictionary element g_γ can be shown to be a weighted average of the inner products of f with finer scale dictionary elements. Recall the notation introduced in Section 2.1, where elements of the cubic b-spline dictionary at scale j and translation t are denoted $g_{j,t}$, or, equivalently, g_γ where γ is a joint index over scale and translation. Any cubic b-spline may be written as the sum of finer scale cubic b-splines composing g_γ which are also dictionary elements. For example, $g_{j,t}$ may be written as the weighted sum of finer scale cubic b-splines which are all at the same scale, $j + k$; that is,

$$g_{j,t} = \sum_{i=1}^{L} c_i g_{j+k,t_i} \tag{8}$$



(a) $k = 1$. (b) $k = 2$.

Figure 7: Weighted sum of cubic b-splines at scale $j + k$ yields a cubic b-spline at scale j.

This is illustrated in Figure 7 for $k = 1$ and $k = 2$. Following this idea, and for convenience later, let us define for each element in the cubic b-spline dictionary, g_γ, an associated set of indices, $I_\gamma(k)$. The functions which are indexed by $I_\gamma(k)$ are the function g_γ and the dictionary elements at the finer scale $j + k$ which when properly weighted and summed yield g_γ[5]. That is,

$$I_\gamma(k) = \left\{ \gamma, (j + k, t_i) | g_\gamma = \sum_{i=1}^{L} c_i g_{j+k,t_i} \right\} \tag{9}$$

Thus, (8) can be written equivalently as

$$g_\gamma = \sum_{i \in I_\gamma(k)} c_i g_i \tag{10}$$

Since $g_{j,t}$ may be represented as the weighted sum of finer scale cubic b-splines, the inner

[5]Of course, one could imagine combinations of finer scale cubic b-splines that are not all at the same scale which also sum to $g_{j,t}$. In some cases, this may be a way to incorporate prior knowledge. For this work, we will use the definition given in (9).

product $< f, g_{j,t} >$ may also be expressed in terms of finer scale inner products,

$$< f, g_{j,t} >= \sum_{i=1}^{L} c_i < f, g_{j+k,t_i} > . \tag{11}$$

or, equivalently,

$$< f, g_\gamma >= \sum_{i \in I_\gamma(k)} c_i < f, g_i > \tag{12}$$

In other words, the inner product of f and g_γ may be interpreted as the weighted average of the inner product of f with high resolution dictionary elements.

The above interpretation of the MP similarity measure yields intuition about what form a new, more locally-sensitive similarity measure might take. Even though each of the high resolution correlations in (11), $\{< f, g_i >\}_{i \in I_\gamma(k)}$, is sensitive to local structure, the (weighted) averaging process of (11) renders $< f, g_\gamma >$ relatively insensitive to local structure. One can imagine that some other combination of the high resolution correlations, $\{< f, g_i >\}_{i \in I_\gamma(k)}$, might yield a new measure of similarity between f and g_γ, which is more sensitive to local mismatch. Intuitively, this new similarity measure should be dominated by worst local fit. For example, the minimum of $\{< f, g_i >\}_{i \in I_\gamma(k)}$ is dominated by worst local fit.

The similarity measure we propose is essentially the minimum over $\{| < f, g_i > |\}_{i \in I_\gamma(k)}$. Our new similarity measure, $S(f, g_\gamma)$, is given by

$$S(f, g_\gamma) = m(f, g_\gamma) s(f, g_\gamma) \tag{13}$$

$$s(f, g_\gamma) = \min_{i \in I_\gamma(k)} \frac{| < f, g_i > |}{| < g_i, g_\gamma > |} \tag{14}$$

$$m(f, g_\gamma) = \begin{cases} +1 & \text{if } \frac{<f,g_i>}{<g_i,g_\gamma>} > 0 \text{ for all } i \in I_\gamma(k) \\ -1 & \text{if } \frac{<f,g_i>}{<g_i,g_\gamma>} < 0 \text{ for all } i \in I_\gamma(k) \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

The denominator of $s(f, g_\gamma)$ is a normalization factor which yields $S(g_\gamma, g_\gamma) = 1$. The term $m(f, g_\gamma)$ is included to assure that oscillatory functions yield a similarity measure of zero with coarse scale dictionary elements.

The HRP algorithm is analogous to the MP algorithm. At each step, the similarity function between $R^n f$ and each element g_γ for all $g_\gamma \in \mathcal{D}$ is calculated. For HRP, the similarity between the n-th residual, $R^n f$, and a dictionary element, g_γ, is given by $S(R^n f, g_\gamma) = m(R^n f, g_\gamma) s(R^n f, g_\gamma)$ as defined in (14) and (15). In the HRP algorithm, the element chosen at the n-th step, g_{γ_n} is given by

$$g_{\gamma_n} = \arg\max_{\gamma \in \Gamma} |S(R^n f, g_\gamma)|. \tag{16}$$

The $n+1$-st residual is then generated as

$$R^{n+1} f = R^n f - S(R^n f, g_{\gamma_n}) g_{\gamma_n}. \tag{17}$$

Additional insight may be gained through the following alternative interpretation of the HRP algorithm. As we now discuss the element which solves a *constrained* maximization of $\|R^{n-1} f - R^n f\|$ is the same one which maximizes the HRP similarity measure, $|S(R^{n-1} f, g_\gamma)|$.

This is analogous to the development of MP in Section 2.1 where we noted that the element which maximized $\|R^{n-1}f - R^n f\|$ was the same one which maximized the inner product similarity measure. Consider the maximization of $\|R^{n-1}f - R^n f\|$ where $R^n f$ is given in (17) under the following constraints :

$$|< R^n f, g_i >| \leq |< R^{n-1} f, g_i >| \qquad \text{for all } i \in I_\gamma(k) \tag{18}$$

$$\text{sign}(< R^n f, g_i >) = \text{sign}(< R^{n-1} f, g_i >) \qquad \text{for all } i \in I_\gamma(k). \tag{19}$$

These constraints are intuitively pleasing. The constraint in (18) captures the idea that the projection of the residual should decrease both globally and locally. In other words, if g_γ is well matched to f, then the projection of the residual onto g_γ should decrease, and the projection of the residual onto all the local structures which make up g_γ (i.e. g_i for $i \in I_\gamma(k)$) should decrease. The constraint in (19) captures the idea that the decomposition should not introduce "non-features" such as those introduced by MP in the twin peaks example. It is important to note that the two constraints effectively balance one another and together imply that the projection onto all local structures of g_γ must decrease, but not so much as to introduce a change in sign. The element which maximizes $\|R^{n-1}f - R^n f\|$ under constraints (18) and (19) is the same one which maximizes $|S(R^{n-1}f, g_\gamma)|$. This result is shown in Appendix A.

One further note about the parameter k which essentially controls the depth of the resolution of the HRP algorithm. The HRP decomposition will change as a function of k, as will be illustrated in Section 4.1. When k is set to zero, the HRP decomposition will

be identical to the MP decomposition. At the other extreme when $k = \infty$, the fine scale elements of $I_\gamma(k)$ will be Diracs and the HRP decomposition will be highly sensitive to noise in the signal. For our work k has been chosen empirically. In general, k should be regarded as a means to incorporate prior knowledge.

Finally, note that the HRP algorithm is not limited to dictionaries where coarse scale elements may be constructed as the weighted sum of finer scale elements. In the preceding discussion, we have concentrated on cubic b-spline dictionaries which have the property that coarse scale elements may be exactly constructed as the weighted sum of finer scale elements and, thus, we were able to define $I_\gamma(k)$ as given in (9). In Section 5, the HRP algorithm is extended to wavelet packet dictionaries which also allow $I_\gamma(k)$ to be defined as in (9). For general dictionaries, however, it may not be possible to represent coarse scale elements exactly as the sum of finer scale elements. In this case, it would be necessary to specify for each dictionary element g_γ a local family I_γ which consists of finer scale functions which somehow capture the local behavior of g_γ.

3.2 Exponential Convergence

In this subsection, the properties of the HRP algorithm for finite discrete functions $f[t]$ for $0 < t \leq P$ are studied. The main result of this subsection shows that if the dictionary Γ is complete then the HRP algorithm produces residuals whose norms decay exponentially.

To prove the exponential convergence of the norm of the residuals produced by HRP, the following lemma is needed. This lemma proves that at each step the similarity function must

be bounded below by a fraction of the energy of the current residual. A crucial element of this proof is the assumption that the dictionary contains all elements $g_\gamma[t]$ of the form :

$$g_\gamma[t] = \delta[t - r] \text{ for } 0 < r \leq P \tag{20}$$

where

$$\delta[t] = \begin{cases} 1 & \text{for } t = 0 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

Note that by definition $S(f, \delta[t - r]) = f[r]$.

Lemma 1 *For a dictionary* Γ *which contains elements of the form given in (20),*

$$|S(R^n f, g_{\gamma_n})| \geq \frac{1}{\sqrt{P}} \|R^n f\| \tag{22}$$

Proof : The similarity function will always be greater than the value of $R^n f$ at any particular point. That is,

$$|S(R^n f, g_{\gamma_n})| \geq |R^n f[r]| \quad \text{for any } r \tag{23}$$

This follows because, by definition,

$$g_{\gamma_n} = \arg \sup_{\gamma \in \Gamma} |S(R^n f, g_\gamma)|, \tag{24}$$

and $\delta[t-r] \in \Gamma$ and $S(R^n f, \delta[t-r]) = R^n f[r]$. This implies

$$|S(R^n f, g_{\gamma_n})| \geq \sup_r |R^n f[r]| \tag{25}$$

Further,

$$\begin{aligned} \|R^n f\|^2 &= \sum_{r=1}^{P} |R^n f[r]|^2 & (26) \\ \|R^n f\|^2 &\leq P(\sup_r |R^n f[r]|)^2 & (27) \end{aligned}$$

which implies

$$\sup_r |R^n f[r]| \geq \frac{1}{\sqrt{P}} \|R^n f\|. \tag{28}$$

It follows that,

$$|S(R^n f, g_{\gamma_n})| \geq \frac{1}{\sqrt{P}} \|R^n f\| \tag{29}$$

□

The following theorem shows that for a complete dictionary which contains elements of the form given in (20), the HRP algorithm yields residuals whose energies decay exponentially.

Theorem 1 *For a dictionary Γ which contains elements of the form given in (20),*

$$\|R^{n+1} f\| \leq (1 - \frac{1}{P})^{1/2} \|R^n f\| \tag{30}$$

Proof : Note that

$$\|R^{n+1}f\|^2 = \|R^n f\|^2 - 2S(R^n f, g_{\gamma_n}) < R^n f, g_{\gamma_n} > + S^2(R^n f, g_{\gamma_n}) \tag{31}$$

From the definition of the similarity function, we know

$$| < R^n f, g_{\gamma_n} > | \geq S(R^n f, g_{\gamma_n}) \tag{32}$$

$$\text{sign}(< R^n f, g_{\gamma_n} >) = \text{sign}(S(R^n f, g_{\gamma_n})) \tag{33}$$

This implies

$$\|R^{n+1}f\|^2 \leq \|R^n f\|^2 - S^2(R^n f, g_{\gamma_n}) \tag{34}$$

Lemma 1 then implies

$$\|R^{n+1}f\|^2 \leq \|R^n f\|^2 - \frac{1}{P}\|R^n f\|^2 \tag{35}$$

$$= \|R^n f\|^2(1 - \frac{1}{P}) \tag{36}$$

□

4 HRP with B-Spline Dictionaries

In the previous section, the formulation of HRP was developed. In this section, the HRP algorithm is applied to several simulated examples. Finally, HRP is used to examine high-resolution radar returns from a Cessna plane.

4.1 Simulated Examples

4.1.1 Twin Peaks Revisited

Recall the twin peaks example of Section 2 for which MP yielded unintuitive results. The twin peaks signal is constructed as the sum of two dictionary elements at scale 32 and translation $t = \pm 0.3281$. The contour plot of the HRP similarity function for fitting the first element is shown in Figure 8 and clearly shows two maxima at the scale and translations which correspond to the features of the original signal. This is in contrast to the analogous contour plot for MP (see Figure 4) which had a single maxima at scale 40 and translation $t = 0$.

The coherent structures of this signal are captured by the first two elements of the HRP approximation. The first ten elements of the HRP decomposition are shown in Figure 9. Since HRP chooses two reasonable elements in the first stages, subsequent elements serve to refine the fit rather than to correct mistakes from previous stages. One can imagine that, in a feature extraction setting, the first two elements would provide a good approximation to the signal and could be used as features of the signal.

As discussed earlier, the HRP decomposition will be affected by the depth at which the

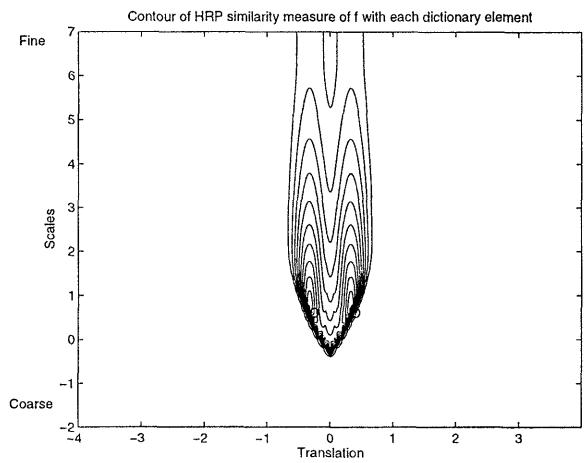


Figure 8: The HRP similarity graph is the HRP similarity measure between the function and each dictionary element which is indexed by scale and translation. This figure shows the contour of the HRP similarity graph. O marks location of true elements of the function which are the same as the maxima of the HRP similarity graph.

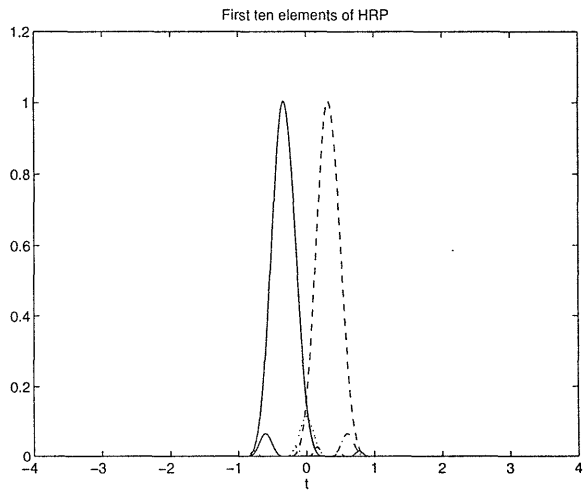


Figure 9: First ten elements for twin peaks example using HRP.

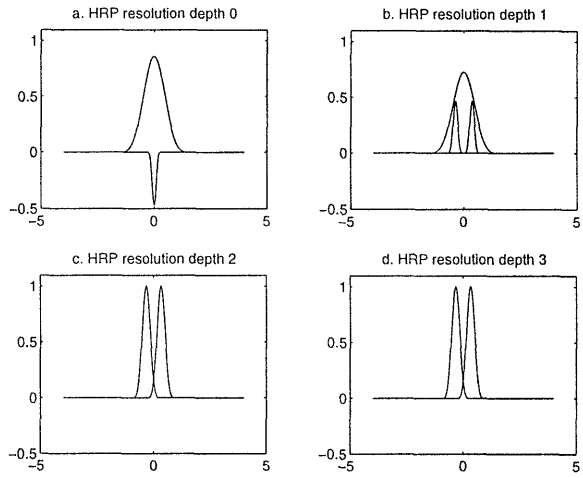


Figure 10: Changes in the HRP decomposition of the twin peaks signal as the resolution depth (i.e. the value of k) is changed. Each subfigure shows the first few elements of the HRP decomposition for a different value of k. (a) $k = 0$. (b) $k = 1$. (c) $k = 2$. (d) $k = 3$.

family $I_\gamma(k)$ is constructed. Figure 10a-d show the coherent features of the HRP decomposition with depths zero, one, two and three, respectively. At a depth of zero, HRP reduces to MP and the signal is decomposed as a coarse scale feature plus a negatively weighted fine scale feature near the center. At a depth of one, HRP gives the decomposition in Figure 10b which may be interpreted as a coarse scale feature plus fine scale details at $t \approx \pm 0.25$. Finally, at a depth of two or higher, HRP gives the decomposition shown in Figure 10c-d, which is interpreted as the sum of two positively weighted fine scale features. In real data applications, the depth of $I_\gamma(k)$ may be used to incorporate prior knowledge into the decomposition.

Figure 11 compares the residual norms for MP and HRP for the twin peaks example up to 1024 elements. We can identify three distinct regions of convergence. In the first region, from approximately element 1 through 10, both algorithms generate residuals whose norms decay at a very similar rate. In this region, both algorithms are extracting coarse scale structures

and the norm of the residuals decays quickly. Both algorithms are behaving in a greedy way by picking coarse features instead of fine features. In the next region, from approximately element 10 to 200, both algorithms produce residuals whose norms decay at an exponential rate. In this region, the MP residual norms are lower than HRP residual norms. This is to be expected since the MP criterion is to minimize the norm of the residual at each step. In this second region, both algorithms continue to behave as greedy procedures and continue to favor coarse features over fine features. The final region starts at approximately element 200. In this final region, the MP residuals continue to decay at an exponential rate, but the HRP residuals decay at a rate much faster than exponential. In this region, the residuals only have structure at the finest scale (i.e. Diracs). HRP will only extract Diracs at this stage; MP, on the other hand, will continue to extract coarser features. In other words, MP continues to behave as a greedy procedure, but HRP ceases to behave in a greedy way. This behavior is simply an extension of the behavior shown in Figure 12 which shows that MP often extracts coarse scale structures from signals which have only fine scale structure, but HRP extracts fine scale structure. The implication of this behavior is that once HRP attains the Dirac extraction mode, the residual will converge to zero in N iterations, where N is the number of samples of the signal.

4.1.2 The Gong Signal

The dashed function in Figure 13 is the envelope of a gong signal. This type of signal has a sharp attack followed by a slow decay. The ideal decomposition would capture the attack

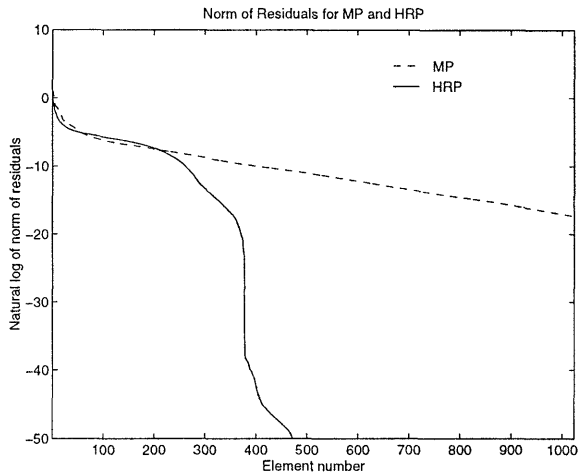


Figure 11: Comparison of MP and HRP residual norms for twin peaks example.

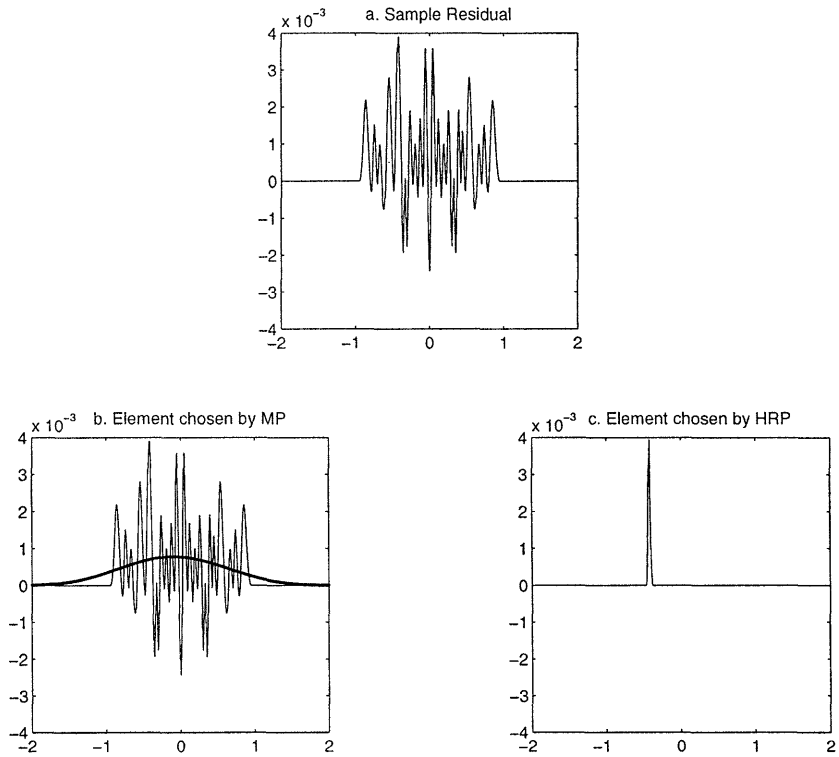


Figure 12: Comparison of MP and HRP on a residue with only fine scale structure. (a) Sample Residual. (b) MP chooses an element with coarse scale structure when the signal has only fine scale features. (c) HRP chooses an element with fine scale structure.

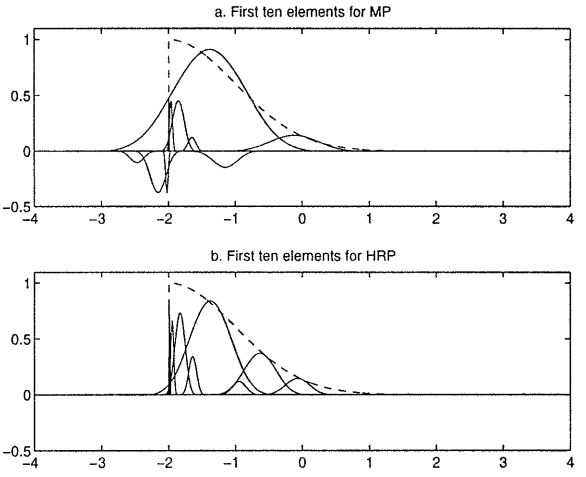


Figure 13: First ten elements for the gong example for MP and HRP.

with elements well localized in time and would not place elements prior to the attack of the signal. Figure 13 shows the first ten elements of the MP and HRP decompositions for the gong signal shown in the dashed line. HRP captures the attack of the signal and does not place elements before the attack. On the other hand, MP places elements prior to the attack which results in subsequent negatively weighted elements which are "non-features."

Figure 14 compares the norms of the MP and HRP residuals. Once again, three regions of convergence are evident. The first region, which extends from element 1 through 10, both algorithms extract the important signals structures and decay at similar rates. In the second region, from element 10 to 500, both algorithms show exponential convergence. In the final region, above element 500, HRP shows a convergence rate much faster than exponential. Again, this results from the fact that HRP enters a mode where it extracts only Diracs.

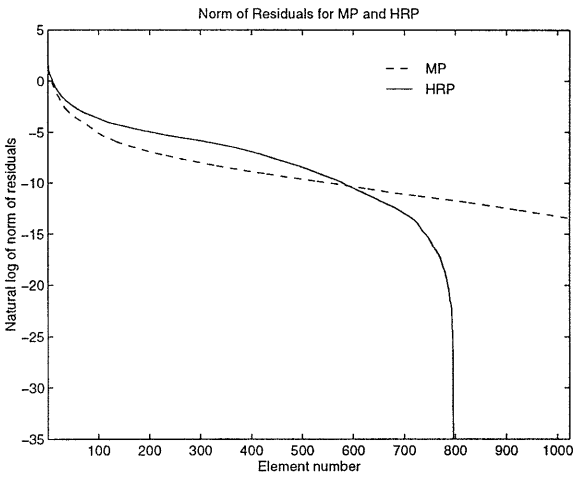


Figure 14: Comparison of MP and HRP residual norms for the gong example.

4.2 High Resolution Radar Examples

Recall the profile of a Cessna 310 airplane shown in Figure 1. Each of the peaks in this signal correspond to physical features of the airplane. In fact, the locations and widths of the peaks in the signal have a direct relation with the geometry of the subparts of the airplane.

Figure 15 shows the HRP decomposition with $k = 2$ of the signal shown in Figure 1. Each of the significant features of the signal is extracted separately by the HRP algorithm. For comparison, the HRP decompositions with resolution depths of zero and one are shown in Figures 16a and b, respectively. Prior knowledge may be used to determine which resolution depth is most appropriate.

The HRP algorithm produces features which are robust to noise. First, consider noise due to small differences in the imaging geometry. Figure 17a and b show two high range resolution signatures of the Cessna plane at slightly different viewing angles. The two signals are very similar in their coherent structures, but they are not identical. The HRP algorithm

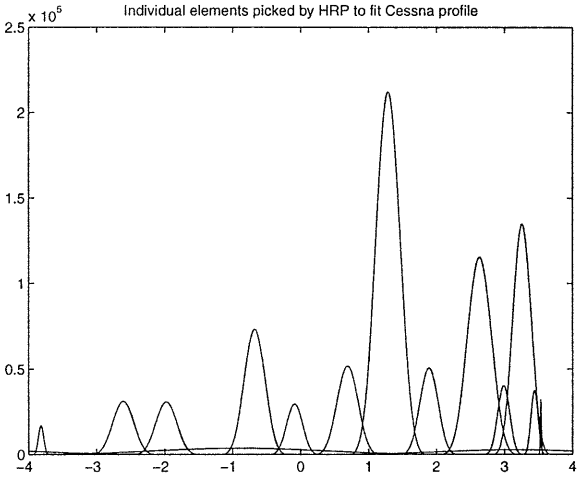


Figure 15: Elements extracted by HRP at depth 2.

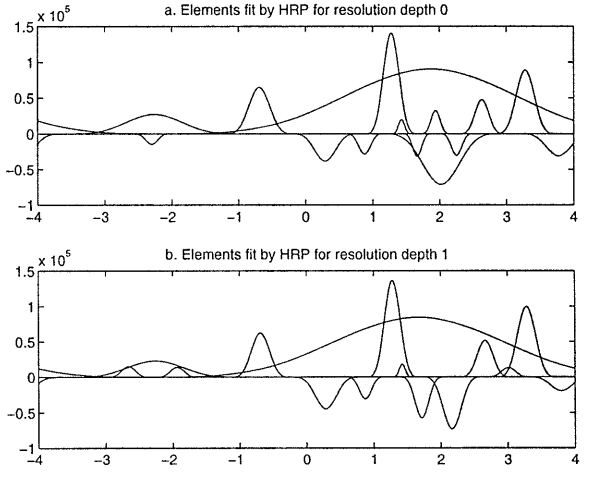


Figure 16: Elements extracted by HRP at depth 0 and 1

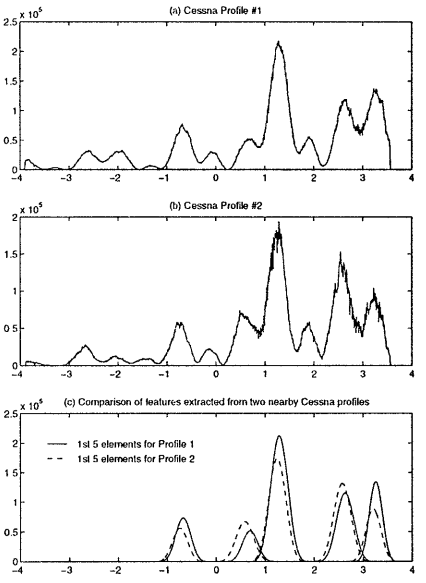


Figure 17: Comparison of two nearby Cessna profiles. (a) Cessna profile # 1. (b) Cessna profile #2. (c) Comparison of elements extracted from the two Cessna profiles.

with resolution depth of two extracts very similar set of features for the two signals. Table 1 lists the first five features (scales and translations) extracted from the two signals. Figure 17c shows a graphical comparison of the features extracted for the two Cessna profiles. Second, consider noise due to a simulated specular flash. Figure 18a shows a Cessna high range resolution signature plus a simulated specular flash. The HRP decomposition in the presence of this type of noise is identical except for an additional feature corresponding to the specular flash, as illustrated in Figure 18b. Third, HRP is robust to additive Gaussian noise. Consider the same Cessna profile corrupted by Gaussian noise as shown in Figure 19. Table 2 lists the first five features (scales and translations) extracted from the noisy signal. Again, a very similar set of features is extracted in the presence of Gaussian noise.

	j_1	t_1	j_2	t_2	j_3	t_3	j_4	t_4	j_5	t_5
Profile #1	18.4	1.29	13.9	3.25	18.4	2.63	16.0	-0.67	16.0	0.69
Profile #2	18.4	1.24	16.0	3.20	18.4	2.57	16.0	-0.74	18.4	0.58

Table 1: Comparison of first five elements extracted from two nearby Cessna profiles. Variables j_i are scales and t_i are translations.

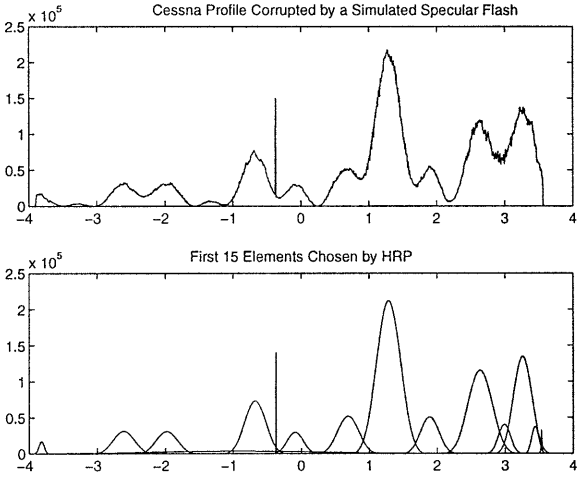


Figure 18: HRP decomposition of Cessna profile corrupted by a simulated specular flash.

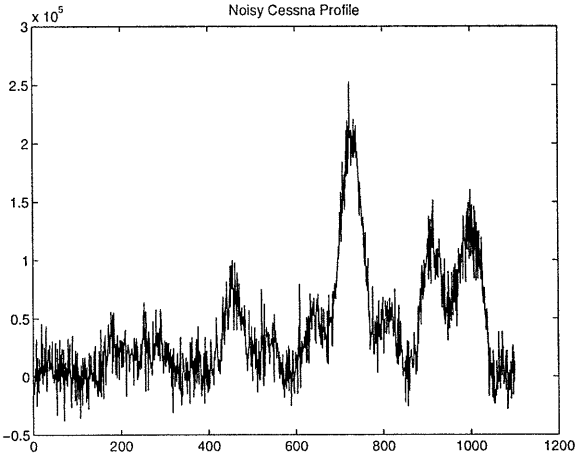


Figure 19: Cessna profile with simulated additive sensor noise.

	j_1	t_1	j_2	t_2	j_3	t_3	j_4	t_4	j_5	t_5
Noisy Profile	18.4	1.29	16.0	3.22	16.0	2.62	16.0	-0.67	16.0	0.68

Table 2: Features extracted from noisy Cessna profile.

5 HRP with Wavelet Packet Dictionaries

In this section, high resolution pursuit using wavelet packet dictionaries will be considered. In Section 4, HRP with cubic b-spline dictionaries was applied to several simulated signals and high range resolution radar returns. For the signals considered in Section 4, the cubic b-spline dictionary was appropriate since it was well matched to the signals being analyzed. However, the cubic b-spline dictionary is not critical to the HRP algorithm. In this section, we consider wavelet packet dictionaries. The wavelet packet dictionary consists of the basis elements used in the wavelet packet decomposition. In this section, the structure of wavelet packet dictionaries will be described and the HRP algorithm using wavelet packet dictionaries will be demonstrated.

5.1 The Wavelet Packet Dictionary Structure

The wavelet packet dictionary is a redundant dictionary consisting of the functions used to generate the wavelet packet decomposition. This section will highlight the structure of the wavelet packet dictionary as it relates to HRP. More complete reviews of the wavelet packet decomposition may be found in [11, 18].

The wavelet packet decomposition is an extension of the wavelet decomposition. Recall that the wavelet decomposition of a function is the projection of that function onto translated, scaled

versions of the mother wavelet, $\psi(x)$. Let

$$\psi_j(x) = \sqrt{2^j}\psi(2^j x). \tag{37}$$

The set of wavelet functions at scale $j \in \mathbf{Z}$ is given by $\{\psi_j(x - 2^{-j}k)\}_{k\in\mathbf{Z}}$, and is an orthogonal basis for the space $\mathbf{W}_j$. For the scaling function $\phi(x)$, let

$$\phi_j(x) = \sqrt{2^j}\phi(2^j x). \tag{38}$$

The set of scaling functions at scale $j \in \mathbf{Z}$ is given by $\{\phi_j(x - 2^{-j}k)\}_{k\in\mathbf{Z}}$, and is an orthogonal basis for the space $\mathbf{V}_j$. The spaces $\mathbf{W}_j$ and $\mathbf{V}_j$ are orthogonal to one another. Linear combinations of the scaling functions at scale j yield the wavelet and scaling functions at the next coarser scale, $j - 1$. These linear combinations are specified the conjugate mirror filters h_1 and h_2 [6]. That is,

$$\phi_{j-1}(x) = \sum_{n=-\infty}^{+\infty} h_1[n]\phi_j(x - 2^{-j}n) \tag{39}$$

$$\psi_{j-1}(x) = \sum_{n=-\infty}^{+\infty} h_2[n]\phi_j(x - 2^{-j}n) \tag{40}$$

These coarser scale functions are bases for a high frequency space, $\mathbf{W}_{j-1}$, and a low frequency space, $\mathbf{V}_{j-1}$, which are contained in $\mathbf{V}_j$ and thus orthogonal to $\mathbf{W}_j$. The wavelet transform is constructed by repeatedly dividing the spaces $\mathbf{V}_j$. As a result, the wavelet transform yields poor resolution for high frequencies. In contrast, the wavelet packet transform is constructed by dividing $\mathbf{W}_j$ as well as $\mathbf{V}_j$. Generalizing the wavelet notation, the wavelet packet decomposition of a function is the

[6] We have used h_1 and h_2 to refer to the conjugate mirror filters which are usually referred to as h and g. This notation was used to avoid confusion with our dictionary elements g.

projection of the function on to a set of spaces $\mathbf{W}_{j,f}$ where j is scale and f is a frequency index. Each space $\mathbf{W}_{j,f}$ has a corresponding orthogonal basis $\{\psi_{j,f}(x - 2^{-j}k)\}_{k\in\mathbf{Z}}$. Linear combinations (specified by h_1 and h_2) of the basis functions of the space $\mathbf{W}_{j,f}$ yield the basis functions of the spaces $\mathbf{W}_{j-1,2f}$ and $\mathbf{W}_{j-1,2f+1}$. That is,

$$\psi_{j-1,2f}(x) = \sum_{n=-\infty}^{+\infty} h_1[n]\psi_{j,f}(x - 2^{-j}n) \tag{41}$$

$$\psi_{j-1,2f+1}(x) = \sum_{n=-\infty}^{+\infty} h_2[n]\psi_{j,f}(x - 2^{-j}n) \tag{42}$$

Where the wavelet transform divided the frequency axis into large intervals at high frequencies and small intervals at low frequencies, the wavelet packet transform divides the frequency axis into intervals of different sizes in a way that does not depend on the frequency. Figure 20 shows sample elements from the Haar wavelet packet dictionary. Note the following important properties of the wavelet packet dictionary. First, elements of the wavelet packet dictionary will still be labeled g_γ, where γ is now a joint index over scale, translation, and frequency. This is in contrast to the cubic b-spline dictionary which was indexed only by scale and translation. One convenient representation of members of a wavelet packet dictionary is on a time-frequency plane. The time-frequency plane representation of the Haar wavelet packet dictionary elements is shown in Figure 21. The scale determines the dimensions of the rectangle and the frequency and translation determine the location. Second, dictionary elements at a given scale are the weighted sum of elements at a finer scale. Recall that the HRP algorithm developed in Section 3 required only that each dictionary element, g_γ, have an associated set $I_\gamma(k)$ which contains γ plus the indices of the finer scale elements which when properly weighted and summed yield g_γ. Thus, the wavelet packet dictionary is appropriate for use with HRP. Third, the collection of functions of the same scale

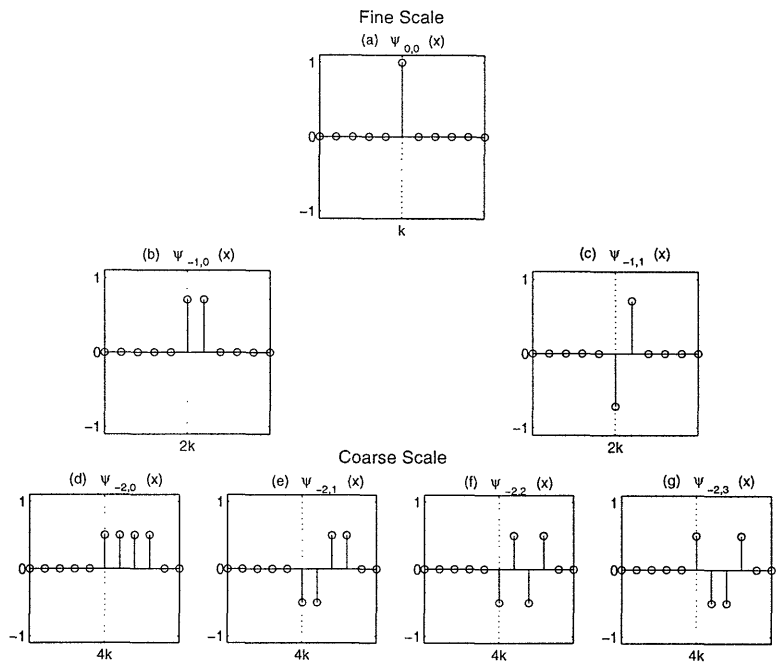


Figure 20: Representative elements of Haar wavelet packet dictionary. Note that the finest resolution has been designated $j = 0$.

(size) is a basis for $\Re^P$, where P is the length of f. Note that the entire dictionary is a collection of bases and is therefore redundant.

5.2 Simulated Examples

In Section 4, the twin peaks and gong examples were used to show that HRP with a cubic b-spline dictionary is able to extract signal structure. In this section, we show that HRP with wavelet packet dictionaries is also able extract signal structure. However, the HRP algorithm will *not* be able to resolve two elements which have the same scale and translation characteristics but differ in frequency. In this section, we highlight the strengths and weaknesses of HRP with wavelet packet dictionaries.

The HRP algorithm with wavelet packet dictionaries proceeds exactly as before. Again, for each

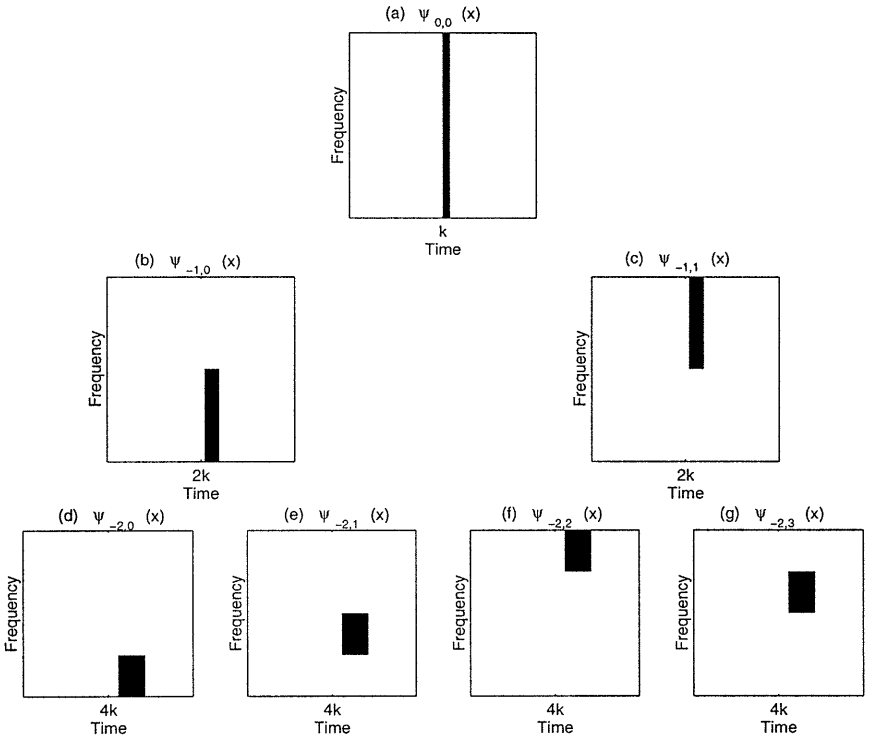


Figure 21: Time Frequency representations of the Haar Wavelet Packet Dictionary

element of the dictionary g_γ, the associated set $I_\gamma(k)$ is chosen to be the indices of the elements at scale $j+k$ which when properly weighted and summed yield g_γ, plus γ, itself. At each step, the similarity function between the current residual and each element in the dictionary, $S(R^n f, g_\gamma)$, is calculated. The inner product $< g_i, g_\gamma >$ which appears in the denominator of $S(R^n f, g_\gamma)$ is determined directly from the quadrature mirror filters h_1 and h_2. The element chosen at the n-th step, g_{γ_n}, is the one which maximizes $|S(R^n f, g_\gamma)|$.

Further, the intuition developed for cubic b-spline dictionaries translates in a straightforward way to wavelet packet dictionaries. For wavelet packet dictionaries, any element may expressed as the weighted sum of finer scale elements. This is the same as for the cubic b-spline dictionary. It follows that the inner product $< f, g_\gamma >$ is just a weighted sum of finer scale inner products. The HRP similarity measure developed in Section 3 is still interpreted as the combination of finer scale

inner products which is dominated by worst local fit.

5.3 The Carbon Signal

Just as was the case for cubic b-spline dictionaries, HRP is able to resolve two elements from a wavelet packet dictionary which are closely spaced in time. Consider the signal carbon shown in Figure 22a. This example is similar to an example considered in [3]. This signal is the sum of four elements : a Dirac, a sinusoid and 2 wavelet packet atoms which are closely spaced in time. The dictionary used is a Symmlet wavelet packet dictionary. Figure 22b shows the time-frequency plane representation of the elements chosen by MP. MP is able to extract the sinusoid and the Dirac, but is unable to resolve the two elements which are closely spaced in time. In contrast, HRP is able to resolve all four elements as shown in Figure 22d. Finally, for comparison, the BP decomposition of this signal is given in Figure 22c. The HRP and BP decompositions are identical, but HRP improves on the BP computation time by a factor of 4.

In the wavelet packet dictionary, it is also possible to construct a signal which is the sum of dictionary elements which share scale and translation characteristics but differ in frequency characteristics. HRP is unable to resolve elements which are closely spaced in frequency. The HRP similarity measure is defined in terms of finer scale elements which cover a wider frequency range. The finer scale elements yield even less frequency resolution than the original coarse scale element. It follows that HRP as we have developed it will be unable to resolve elements which are closely spaced in frequency. One can imagine, however, developing an algorithm analogous to HRP to resolve elements close in frequency.

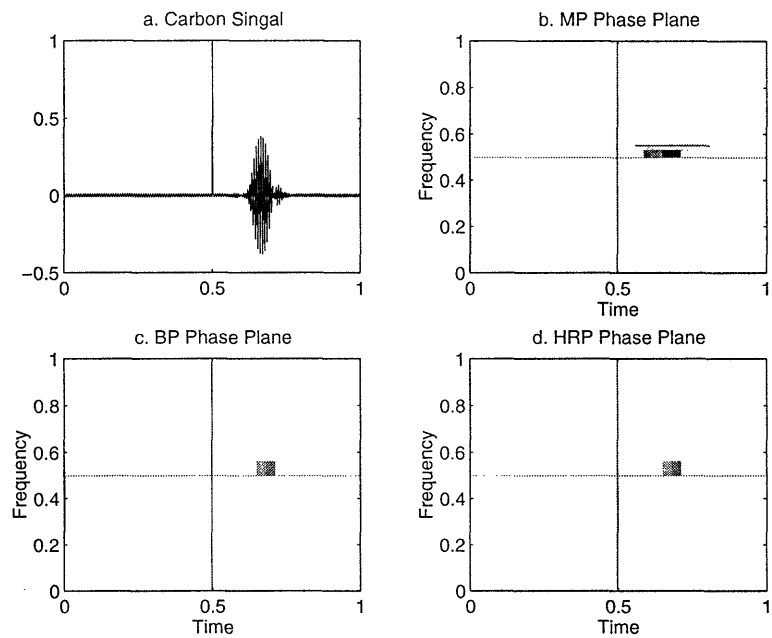


Figure 22: Results for the carbon signal. (a) The carbon signal which consists of the sum of four dictionary elements. (b) The MP decomposition. Note that nearby elements are blurred. (c) The BP decomposition. Note that all four elements are resolved. (d) The HRP decomposition. Again, all four elements are resolved.

5.4 The Gong Signal

Figure 23a shows a gong signal. As was mentioned in Section 4.1.2, this type of signal with a sharp attack followed by a slow decay is important in several signal processing applications. Again, the ideal decomposition would capture the attack with elements well localized in time and would capture the correct frequency of the modulation. Further, the ideal decomposition would not introduce elements prior to the attack of the signal. That is, it would not introduce a pre-echo effect which is particularly disturbing for audio signals.

Figures 23b-d show the time-frequency plane results for MP, BP, and HRP, respectively. The partial reconstructions for three, five and ten elements each of the three methods are shown in Figure 24. The signal was analyzed using a wavelet packet dictionary constructed from the Daubechies six tap wavelet. MP captures the point of the attack and identifies the correct frequency, but introduces several elements prior to the attack of the signal which results in the addition of subsequent "non-features" in the reconstruction. Although the elements before the attack have a small weight, they significantly impact the reconstruction. Thus, the MP reconstruction exhibits this pre-echo effect. BP performs very well since it captures the attack, does not place elements prior to the attack of the signal, and captures the correct frequency of the modulation. HRP captures the point of the attack and does not introduce elements prior to the attack of the signal. However, HRP does not do as well as BP in capturing the correct frequency of the modulation. Comparing the rates of decay of the three methods (see Figure 25), we see that BP decays at a rate faster than HRP. In conclusion, HRP does not surpass BP in the quality of the decompositions. However, HRP provides reasonable decompositions without the intensive computation that may be required by BP.

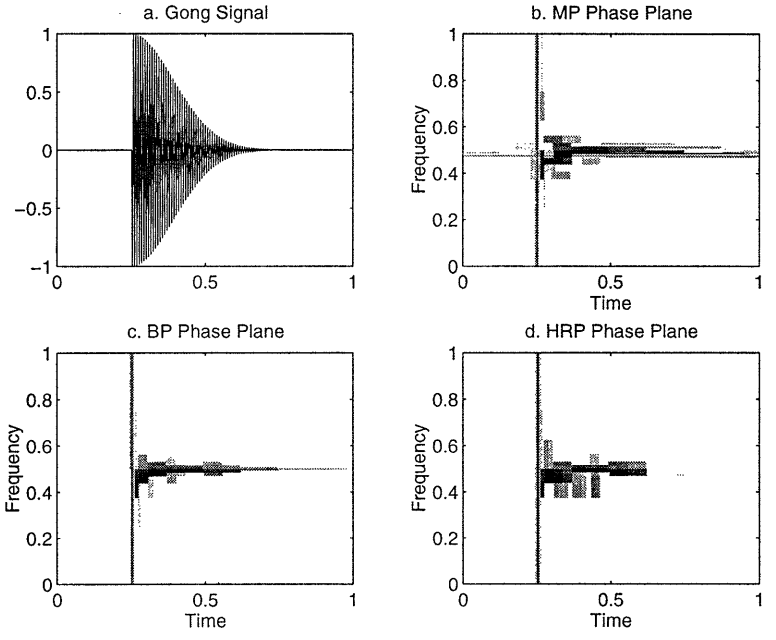


Figure 23: (a) The gong signal. (b) Time-Frequency plane for MP. (c) Time-frequency plane for BP. (d) Time-Frequency plane for HRP.

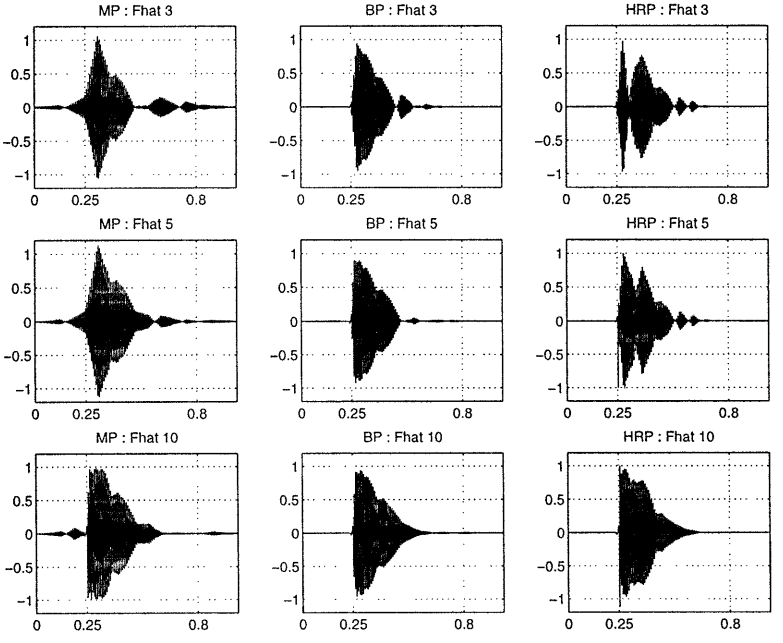


Figure 24: Partial reconstructions for MP, BP, and HRP with 3, 5 and 10 elements. In the MP reconstruction, we see the elements prior to the attack of the signal have a significant impact on the reconstruction.

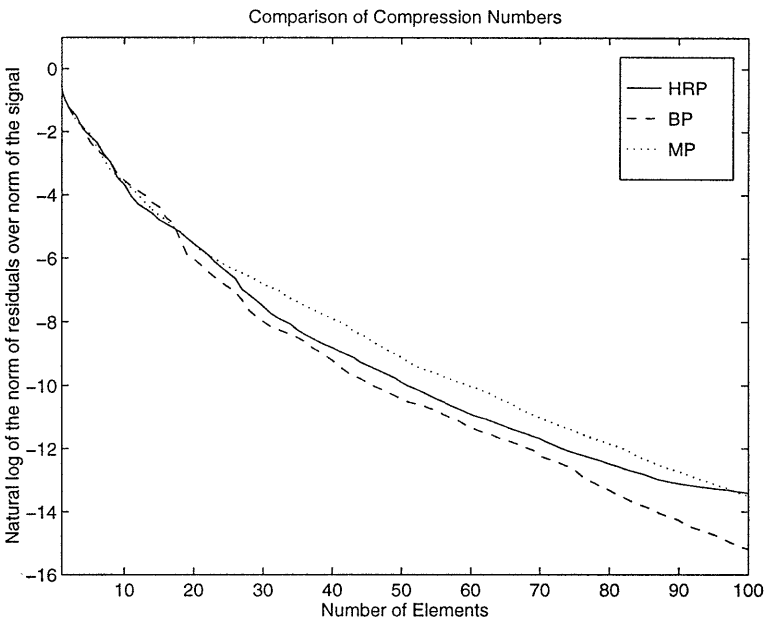


Figure 25: Rates of decay of the three methods.

6 HRP Computational Complexity

The HRP algorithm may be efficiently implemented by sampling the scale/shift space. Recall the notation for the dictionary is $\{g_\gamma | \gamma \in \Gamma\}$. Suppose we construct a reduced dictionary $\{g_\gamma | \gamma \in \Gamma_R\}$. For the cubic b-spline dictionary, the reduced dictionary has scales j which are integers in the range $0 \leq j \leq \log_2(P)$, where P is the length of the signal, and 2^j evenly spaced translations. This reduced dictionary has a total of $C = 2P - 1$ elements. Let H be the set of functions which form the subfamilies for all elements of the reduced dictionary, $H = \{g_i\}$ for $i \in I_\gamma$ and $\gamma \in \Gamma_R$. The HRP algorithm is *initialized* by computing $< f, g_i >$ for all $g_i \in H$ and $< g_\gamma, g_i >$ for all $\gamma \in \Gamma$ and all $g_i \in H$. This initialization requires a one-time computation of $\mathcal{O}(P^2(\log_2(P))^2)$ operations using the FFT. The HRP similarity measure $S(f, g_\gamma)$ for $\gamma \in \Gamma_R$ may then be computed in $\mathcal{O}(KC)$ operations where K is the cardinality of the set $I_\gamma(k)$. The element which maximizes $|S(f, g_\gamma)|$ over the reduced dictionary is an approximation to the element which maximizes $|S(f, g_\gamma)|$ over

the unreduced dictionary. The element which maximizes $|S(f, g_\gamma)|$ unreduced dictionary, g_{γ_0}, could then be found using a Newton search strategy. Using (17), the inner products $< Rf, g_i >$ for all $g_i \in H$ can be computed as

$$< Rf, g_i > = < f, g_i > - S(f, g_{\gamma_0}) < g_{\gamma_0}, g_i > . \tag{43}$$

Since each of the terms on the right hand side of (43) has been previously stored, the calculation of $< Rf, g_i >$ for all $g_i \in H$ takes $\mathcal{O}(KC)$ operations. Extending this argument, we see that each iteration takes $\mathcal{O}(KC) = \mathcal{O}(2PK)$ operations. The number of iterations will typically be much smaller than P.

For the wavelet packet dictionary, the size of the reduced dictionary is $C = P \log_2(P)$. This reduced dictionary has scales j which are integers in the range $0 \leq j \leq \log_2(P)$, $2^{-j}P$ frequency bins for scale j, and 2^j evenly spaced translations for every scale and frequency bin. HRP using the wavelet packet dictionary can be initialized in $\mathcal{O}(P^2 \log_2(P))$ operations by computing $< f, g_i >$. Each iteration for HRP with the wavelet packet dictionary requires the computation of $S(R^n f, g_\gamma)$, the computation of $< g_{\gamma_n}, g_i >$, and the computation of $< Rf, g_i >$. This is a total of $\mathcal{O}(KC) = \mathcal{O}(KP \log_2(P))$ operations per iteration where K is the cardinality of the set $I_\gamma(k)$. Again, the number of iterations will be much smaller than P.

7 Conclusion

To summarize, our initial goal was a novel feature extraction routine. Existing approaches from function approximation did not meet our feature extraction goals. MP failed to super-resolve closely

spaced features and BP was computationally intensive. An alternative function approximation approach, HRP, was developed and demonstrated in this paper. In the same flavor as MP, HRP picks the most contributive element at each step. However, in HRP, the similarity function is modified to guide the decomposition away from blurring adjacent features. The HRP similarity measure developed in this work is one which is dominated by the worst local fit. We have demonstrated the HRP algorithm on simulated and real 1D functions. Further, the exponential convergence of HRP for finite discrete functions was proven. Future research directions include a demonstration of object recognition using HRP features and the extension of the HRP algorithm to 2D functions.

Acknowledgments

The authors would like to thank Rome Laboratory for collecting the Cessna range profiles used in this paper. We would also like to thank Jody O'Sullivan and Steve Jacobs for providing us with the preprocessed range profiles.

High resolution pursuit using wavelet packets was developed and demonstrated using the WaveLab and Atomizer software packages developed by David Donoho's group at Stanford University.

A The HRP Similarity Measure

The element which maximizes $\|R^n f - R^{n-1} f\|$ under constraints (18) and (19) also maximizes the new similarity measure $|S(f, g_\gamma)|$ as given in (14) and (15). Consider the first stage residual Rf and let $R_\gamma f$ be the residual produced by choosing some dictionary element g_γ. That is,

$$R_\gamma f = f - S(f, g_\gamma) g_\gamma. \tag{44}$$

where $S(f, g_\gamma)$ is a scalar. It follows that

$$\|R_\gamma f - f\| = |S(f, g_\gamma)|. \tag{45}$$

We begin by showing that for any dictionary element, $S(f, g_\gamma)$ as defined in (14) and (15) maximizes $\|R_\gamma f - f\|$ under constraints (18) and (19). Assume for now that

$$\frac{< f, g_i >}{< g_\gamma, g_i >} > 0 \qquad \text{for all } g_i \in I_\gamma(k) \tag{46}$$

For any dictionary element, constraint (18) may be simplified as follows

$$| < R_\gamma f, g_i > | \leq | < f, g_i > | \qquad \text{for all } g_i \in I_\gamma(k) \tag{47}$$

$$| < f, g_i > - S(f, g_\gamma) < g_\gamma, g_i > | \leq | < f, g_i > | \tag{48}$$

$$|1 - S(f, g_\gamma) \frac{< g_\gamma, g_i >}{< f, g_i >}| \leq 1 \tag{49}$$

$$0 \leq S(f, g_\gamma) \leq \frac{2 < f, g_i >}{< g_\gamma, g_i >} \tag{50}$$

where the last line follows because of (46). Further, for any dictionary element, constraint (19) may be simplified as

$$\text{sign}(< Rf, g_i >) = \text{sign}(< f, g_i >) \qquad \text{for all } g_i \in I_\gamma(k) \tag{51}$$

$$< Rf, g_i >< f, g_i > \geq 0 \tag{52}$$

$$(< f, g_i > - S(f, g_\gamma) < g_i, g_i >) < f, g_i > \geq 0 \tag{53}$$

$$S(f, g_\gamma) \leq \frac{< f, g_i >}{< g_\gamma, g_i >} \tag{54}$$

where the last line follows because of (46). The same derivation can be followed through for the case where $\frac{<f,g_i>}{<g_\gamma,g_i>} < 0$ for all $g_i \in I_\gamma(k)$. For the case where the ratio $\frac{<f,g_i>}{<g_\gamma,g_i>}$ does not have the same sign for all $g_i \in I_\gamma(k)$, the only value of $S(f,g_\gamma)$ which meets both constraints is zero. Thus, for any dictionary element, $S(f,g_\gamma)$ as defined in (14) and (15) maximizes $\|R_\gamma f - f\|$ under constraints (18) and (19).

Further, the single dictionary element which maximizes $\|R_\gamma f - f\|$ under constraints (18) and (19) is the same one which maximizes $|S(f,g_\gamma)|$.

References

[1] Yaser Abu-Mostafa and Demetri Psaltis. Recognitive aspects of moment invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), November 1984.

[2] Robert Bergevin and Martin Levine. Part-based description and recognition of objects in line drawings. In *Intelligent Robots and Computer Vision III : Algorithms and Techniques*, pages 63–74. SPIE, 1989.

[3] Shaobing Chen and David Donoho. Atomic decomposition by basis pursuit. Technical report, Statistics Dept., Stanford University, May, 1995. Available via ftp at playfair.stanford.edu.

[4] R. R. Coifman and M. V. Wickerhauser. Entopy-based algorithms for best-basis selection. *IEEE Trans. Info. Theory*, 38:713–718, 1992.

[5] I. Daubechies. Time-frequency localization operators: a geometric phase space approach. *IEEE Trans. Info. Theory*, 34(4):605–612, 1988.

[6] D. den Hertog. *Interior Point Approach to Linear, Quadratic and Convex Programming.* Kluwer Academic Publishers, 1994.

[7] S-C. Fang and S. Puthenpura. *Linear Optimization and Extensions : Theory and Alogrithms.* Prentice Hall, 1993.

[8] W. Eric L. Grimson. On the recognition of parameterized 2D objects. *International Journal of Computer Vision*, 3:353–372, 1989.

[9] Alok Gupta, Gareth Funka-Lea, and Kwanguyeon Wohn. Segmentation, modeling and classification of the compact objects in a pile. In *Intelligent Robots and Computer Vision III : Algorithms and Techniques*, pages 98–108. SPIE, 1989.

[10] P. J. Huber. Projection pursuit. *The Annals of Statistics*, 1985(2):435–475, 1985.

[11] B. Jawerth and W. Sweldens. An overview of wavelet based multiresolution analysis. *SIAM Review*, 36(3):377–412, September 1994.

[12] Z-Q. Liu and Terry Caelli. Multiobject pattern recognition and detection in noisy background using a hierarchical approach. *Computer Vision, Graphics and Image Processing*, 44:296–306, 1988.

[13] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, Dec 1993.

[14] Murali Menon, Eric Boudreau, and Paul Kolodzy. An automatic ship classification system for isar imagery. *Linoln Laboratory Journal*, 6(2), 1993.

[15] Whitman Richards, Benjamin Dawson, and Douglas Whittington. Encoding contour shape by curvature extrema. In Whitman Richards, editor, *Natural Computation.* MIT Press, 1988.

[16] Cho-Huak Teh and Roland Chin. On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4), July 1988.

[17] D. R. Wehner. *High Resolution Radar.* Artech House, 1987.

[18] M. V. Wickerhauser. Lectures on wavelet packet algorithms. Technical report, Washington University, 1991.