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 Institute, 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@lids.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 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,

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

where the set $\{g_\gamma | \gamma \in \Gamma\}$ 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), including method of frames [4], best orthogonal basis [3], matching pursuit [11], and basis pursuit [2]. The definition of "optimal" is application dependent.

For this work, the application of interest is feature extraction for object recognition [1,7,10,12]. 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. Thus, for our work, the "optimal" representation would be one which is hierarchical, stable, quickly computable, sparse, and physically interpretable. While the first three properties are self-explanatory, the last two need some explanation. A sparse representation is one in which a minimum number of dictionary elements are used to represent any function. If a function is synthesized as the sum of dictionary elements, then a sparsity preserving representation would be precisely those elements used to construct 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 or the underlying object. For example, Figure 1 shows a high resolution radar return from a Cessna 310

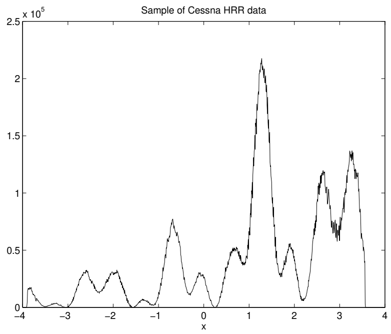


Figure 1: Sample of Cessna high resolution radar profile.

aircraft. Each of the peaks in the signal are related to physical features of the plane [13]. One physically interpretable representation is where each term of (1) corresponds to one of the peaks of the signal.

Existing adaptive approximation techniques do not always yield representations with the desired characteristics as summarized in [2]. The method of frames and best orthogonal basis tend towards solutions which are not sparsity preserving and are unable to resolve closely spaced features. Matching pursuit (MP) is also unable to resolve closely spaced features. Finally, basis pursuit (BP) produces representations which preserve sparsity and resolve closely spaced features, but is computationally complex. 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 MP algorithm. HRP was developed to overcome the shortcomings of the traditional MP algorithm by emphasizing local fit over global fit without significantly increasing the computational complexity of MP. This paper concentrates on the development of HRP in one dimension.

This paper is organized as follows. Section 2 summarizes two adaptive approximation schemes : MP and BP. Section 3 describes the HRP algorithm and discusses convergence issues. Sections 4 and 5 present numerical examples of the performance of the HRP algorithm.

2 Adaptive Approximation of Signals

In this section, a brief description of relevant adaptive schemes for signal approximation is presented. In particular, we describe the MP [11] and BP [2] algorithms. 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 $g_\gamma \in \mathcal{D}$. 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

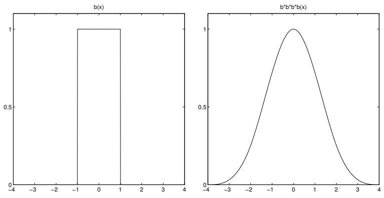
Matching pursuit (MP) is a recursive, adaptive algorithm for signal decomposition [11]. The matching pursuit algorithm builds up the signal representation one element at a time, picking the most contributive element at each step. The element chosen at the n-th step is the one which minimizes $||R^n f||$ 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

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

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



(a) (b)

Figure 2: Box Splines.(a) A box spline, $b(x)$ (b) A cubic box spline, $b * b * b * b(x)$.

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 $\|R^{n-1}f - R^n f\|$. The MP algorithm yields a cumulative decomposition of

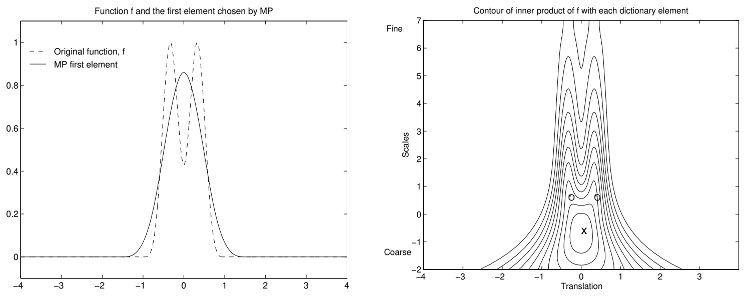
$$f = \sum_{i=0}^{n-1} < R^i f, g_{\gamma_i} > g_{\gamma_i} + R^n f \tag{6}$$

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 time-frequency localizations [11]. 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 (i.e. more localized in time). 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. 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 3a which shows the original function and the first element chosen by MP, g_{γ_0}. The projection graph is the inner product of the function with each dictionary element which is indexed by scale and translation. The contour of the projection graph shown in Figure 3b gives us more insight into the behavior of MP for this case. In fact, 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, shown in Figure 4a, 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 4b shows the first ten MP elements. Note that none of these elements correspond to the physical features of the function.

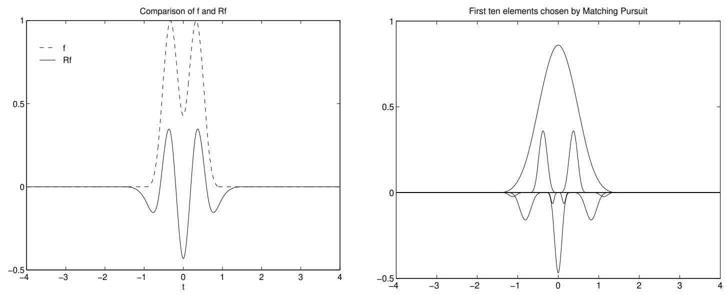
2.2 Basis Pursuit

The basis pursuit (BP) principle [2] is to find the decomposition given in (1) which minimizes the ℓ^1-norm of the coefficients λ_n. The examples presented in [2] indicate that basis pursuit yields decompositions which are sparse and physically meaningful. Thus, they do not exhibit 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



(a) (b)

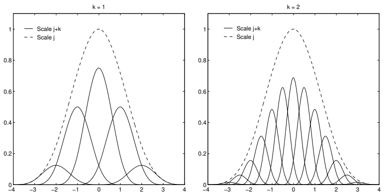
Figure 3: (a) The twin peaks function and first element chosen by MP. (b) The projection graph is the inner product of the function with each dictionary element which is indexed by scale and translation. This figure shows the contour of the projection graph. X marks maximum inner product. O marks location of true elements of function.



(a) (b)

Figure 4: (a) First residual generated by MP. (b) The first ten elements picked by MP.

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 [2]. Two algorithms are proposed in [2] 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 [5, 6]. Thus, the implementation of basis pursuit is computationally intensive.



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

Figure 5: Weighted sum of cubic b-splines at scale $j + k$ yields a cubic b-spline at scale j.

3 High Resolution Pursuit

The objective of high resolution pursuit is to obtain the computational speed of MP as well as the physically meaningful representations of BP. In this section, the HRP algorithm, which parallels the MP algorithm, is developed. First, a new, more locally-sensitive similarity measure is proposed. Second, the HRP algorithm is outlined. Third, to gain additional insight into the HRP algorithm, we discuss an alternative interpretation of HRP.

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. Note that any cubic b-spline may be written as the sum of finer scale cubic b-splines which are also dictionary elements. For example, using the notation introduced in Section 2.1, $g_{j,t}$ may be written as the weighted sum of finer scale cubic b-splines which are all at the same scale, $j + k$; that is,

$$g_{j,t} = \sum_{i=1}^{L} c_i g_{j+k,t_i} \tag{7}$$

This is illustrated in Figure 5 for $k = 1$ and $k = 2$. Following this idea, and for later convenience, 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,

$$I_\gamma(k) = \left\{ \gamma, (j+k, t_i) | g_\gamma = \sum_{i=1}^{L} c_i g_{j+k,t_i} \right\} \tag{8}$$

Thus, (7) can be written equivalently as

$$g_\gamma = \sum_{i \in I_\gamma(k)} c_i g_i \tag{9}$$

Since $g_{j,t}$ may be represented as the weighted sum of finer scale cubic b-splines, the inner product $< f, g_{j,t} >$ may also be expressed in terms of finer scale inner products,

$$< f, g_{j,t} >= \sum_{i=1}^{L} c_i < f, g_{j+k,t_i} >= \sum_{i \in I_\gamma(k)} c_i < f, g_i > \tag{10}$$

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 (10), $\{< f, g_i >\}_{i \in I_\gamma(k)}$, is sensitive to local structure, the (weighted) averaging process of (10) 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.

[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 $g_{j,t}$. For this work, we will use the definition given in (8).

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

$$
\begin{aligned}
S(f, g_\gamma) &= m(f, g_\gamma) s(f, g_\gamma) &(11)\\
s(f, g_\gamma) &= \min_{i \in I_\gamma(k)} \frac{| < f, g_i > |}{| < g_i, g_\gamma > |} &(12)\\
m(f, g_\gamma) &= \begin{cases} +1 & \text{if } \frac{<f, g_i>}{<g_i, g_\gamma>} > 0 \text{ for all } i \in I_\gamma(k) \\ -1 & \text{if } \frac{<f, g_i>}{<g_i, g_\gamma>} < 0 \text{ for all } i \in I_\gamma(k) \\ 0 & \text{otherwise} \end{cases} &(13)
\end{aligned}
$$

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.

In the HRP 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 (12) and (13). In the HRP algorithm, the element chosen at the n-th step, g_{γ_n} is given by

$$g_{\gamma_n} = \arg\max_{\gamma \in \Gamma} |S(R^n f, g_\gamma)|. \tag{14}$$

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

$$R^{n+1} f = R^n f - S(R^n f, g_{\gamma_n}) g_{\gamma_n}. \tag{15}$$

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 (15) under the following constraints :

$$|< R^n f, g_i >| \leq |< R^{n-1}f, g_i >| \qquad \text{for all } i \in I_\gamma(k) \tag{16}$$

$$\text{sign}(< R^n f, g_i >) = \text{sign}(< R^{n-1}f, g_i >) \qquad \text{for all } i \in I_\gamma(k). \tag{17}$$

These constraints are intuitively pleasing. The constraint in (16) 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 (17) 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 (16) and (17) 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 and insensitive to the local structures of the signal. At the other extreme when $k \to \infty$, the fine scale elements of $I_\gamma(k)$ approach Diracs and the HRP decomposition will be highly sensitive to noise in the signal. Thus, k can be used to build robustness into the HRP algorithm. The appropriate value of k is one which provides physically meaningful features but is still robust to noise. For our work k has been chosen empirically.

Finally, note that the HRP algorithm may be used with many dictionaries (e.g. wavelets, wavelet packet, and local cosine dictionaries) and 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 (8). In Section 5, the HRP algorithm is extended to wavelet packet dictionaries which also allow $I_\gamma(k)$ to be defined as in (8). 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.

3.1 Exponential Convergence

In this section, 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.

The following 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 :

$$g_\gamma[t] = \delta[t-r] \text{ for } 0 < r \leq P \tag{18}$$

where

$$\delta[t] = \begin{cases} 1 & \text{for } t = 0 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

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 (18),*

$$|S(R^n f, g_{\gamma_n})| \geq \frac{1}{\sqrt{P}} \|R^n f\| \tag{20}$$

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

$$|S(R^n f, g_{\gamma_n})| \geq |R^n f[r]| \quad \text{for any } r \tag{21}$$

This follows because, by definition,

$$g_{\gamma_n} = \arg\sup_{\gamma \in \Gamma} |S(R^n f, g_\gamma)|, \tag{22}$$

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

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

Further,

$$\|R^n f\|^2 = \sum_{r=1}^{P} |R^n f[r]|^2 \tag{24}$$

$$\|R^n f\|^2 \leq P(\sup_r |R^n f[r]|)^2 \tag{25}$$

which implies

$$\sup_r |R^n f[r]| \geq \frac{1}{\sqrt{P}} \|R^n f\|. \tag{26}$$

It follows that,

$$|S(R^n f, g_{\gamma_n})| \geq \frac{1}{\sqrt{P}} \|R^n f\| \tag{27}$$

□

The following theorem shows that for a complete dictionary which contains elements of the form given in (18), the HRP algorithm yields residuals whose energies decay exponentially.

Theorem 1 *For a dictionary* Γ *which contains elements of the form given in (18),*

$$\|R^{n+1}f\| \leq (1 - \frac{1}{P})^{1/2}\|R^n f\| \tag{28}$$

Proof : Note that

$$\|R^{n+1}f\|^2 = \|R^n f\|^2 - 2S(R^n f, g_{\gamma_n}) < R^n f, g_{\gamma_n} > + S^2(R^n f, g_{\gamma_n}) \tag{29}$$

From the definition of the similarity function, we know

$$|< R^n f, g_{\gamma_n} >| \geq S(R^n f, g_{\gamma_n}) \tag{30}$$

$$\mathrm{sign}(< R^n f, g_{\gamma_n} >) = \mathrm{sign}(S(R^n f, g_{\gamma_n})) \tag{31}$$

This implies

$$\|R^{n+1}f\|^2 \leq \|R^n f\|^2 - S^2(R^n f, g_{\gamma_n}) \tag{32}$$

Lemma 1 then implies

$$\|R^{n+1}f\|^2 \leq \|R^n f\|^2 - \frac{1}{P}\|R^n f\|^2 \tag{33}$$

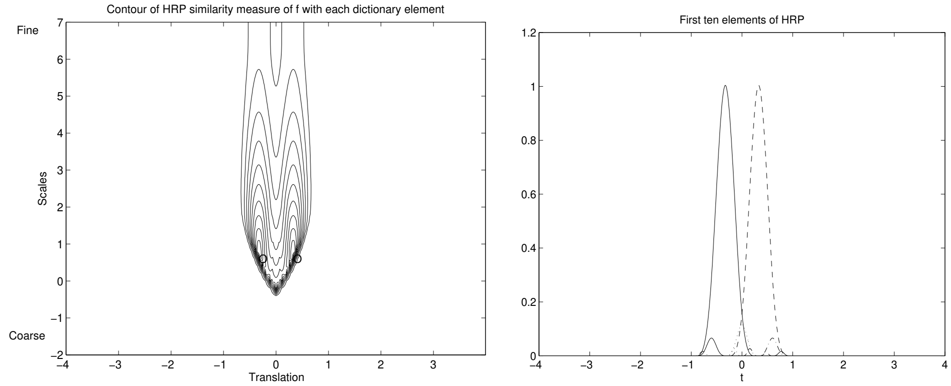
$$= \|R^n f\|^2(1 - \frac{1}{P}) \tag{34}$$

□

4 HRP with B-Spline Dictionaries

4.1 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 6



(a) (b)

Figure 6: (a) 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. (b) First ten elements for the twin peaks example using HRP.

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 3)b 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 6. 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[6].

As discussed earlier, the HRP decomposition will be affected by the depth at which the family $I_\gamma(k)$ is constructed. Figure 7a-c show the coherent features of the HRP decomposition with depths zero, one, and two, respectively. At a depth of zero, HRP reduces to MP. By choosing the parameter k, we can choose among the three decompositions shown in Figure 7. Figure 8 compares the residual norms for MP and HRP for the twin peaks example up to 1024 elements. We can identify three

[6]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.

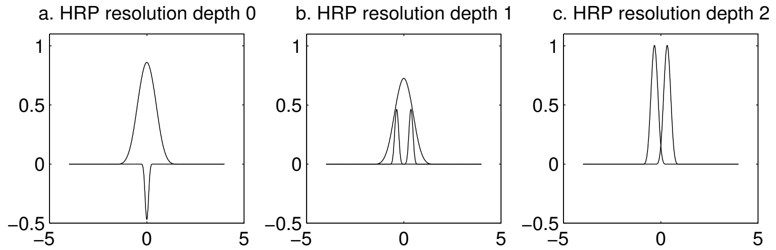


Figure 7: Changes in the HRP decomposition of the twin peaks signal as the resolution depth (i.e. the value of k) is changed. Each subfigure shows the first few elements of the HRP decomposition for a different value of k. (a) $k = 0$. (b) $k = 1$. (c) $k = 2$.

distinct regions of convergence. In the first region, from approximately element 1 through 10, both algorithms generate residuals whose norms decay at very similar rates. In the next region, from approximately element 10 to 200, both algorithms produce residuals whose norms decay at an exponential rate, but the MP residual norms are lower than HRP residual norms. In these first two regions, both algorithms behave as greedy procedures and favor coarse features over fine features. Around element 200, we see a third region where the MP residuals continue to decay at an exponential rate, but the HRP residuals decay at a rate much faster than exponential. Once the residuals only have structure at the finest scale (i.e. Diracs). HRP will only extract Diracs. On the other hand, MP will continue to extract coarser features. Thus, HRP more accurately reflects the true structure of the residual. In other words, MP continues to behave as a greedy procedure, but HRP ceases to behave in a greedy way[7].

4.2 The Gong Signal

The dashed function in Figure 9 is the envelope of a gong signal, a signal which has a sharp attack followed by a slow decay. The ideal decomposition would capture the attack with elements well

[7] In the feature extraction/object recognition context, it will be necessary to develop a stopping rule so that the algorithm does not overfit the noise. This type of stopping rule is application dependent, and therefore, beyond the scope of this paper.

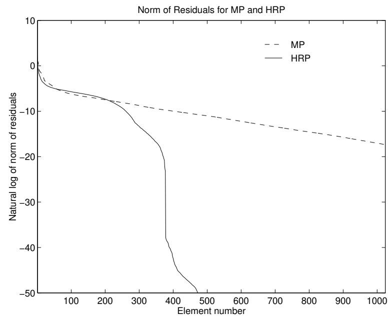


Figure 8: Comparison of MP and HRP residual norms for twin peaks example.

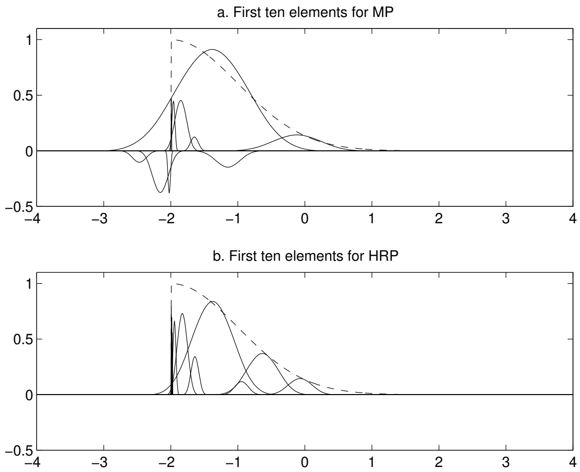


Figure 9: First ten elements for the gong example for MP and HRP.

localized in time and would not place elements prior to the attack of the signal. Figure 9 shows the first ten elements of the MP and HRP decompositions for the gong signal. HRP 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."

4.3 High Resolution Radar Examples

Recall the profile of a Cessna 310 airplane shown in Figure 1 where each of the peaks in this signal correspond to physical features of the airplane. Figure 10 shows the HRP decomposition with $k = 2$ of the signal shown in Figure 1. HRP extracts each of the significant features separately.

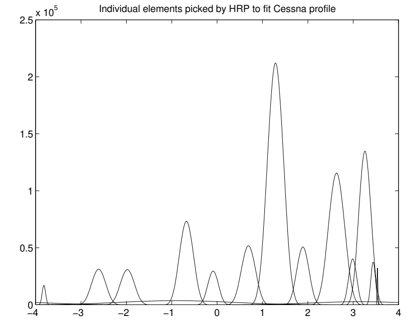


Figure 10: Elements extracted by HRP at depth 2.

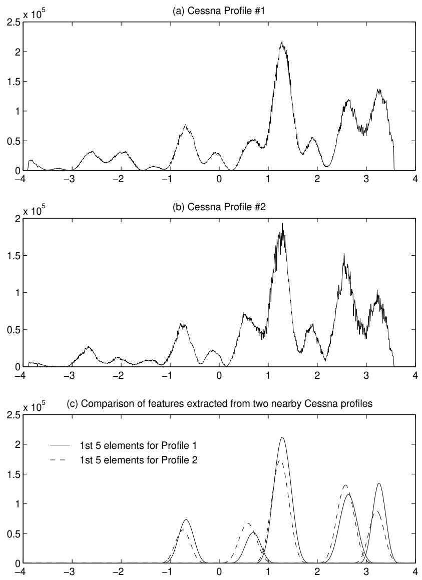


Figure 11: Comparison of two nearby Cessna profiles. (a) Cessna profile # 1. (b) Cessna profile #2. (c) Comparison of elements extracted from the two Cessna profiles.

	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.

The HRP algorithm produces features which are robust to noise. First, consider noise due to small differences in the imaging geometry. Figure 11a 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 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 11c shows a graphical comparison of the features extracted for the two Cessna profiles. Second, consider noise due to a simulated specular flash. Figure 12a 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 12b. Third, HRP is robust to additive Gaussian noise. Figure 13a shows a Cessna profile corrupted with additive Gaussian noise with $\sigma = 10^4$. Figure 13b shows the first five HRP features extracted from the noisy profile as well as the first five HRP features extracted from the noiseless profile. Again, a very similar set of features is extracted in the presence of Gaussian noise. To illustrate the stability of the HRP features in the presence of additive Gaussian noise, we created 100 noisy Cessna profiles in the same way and extracted HRP features from each of them. For each of these profiles, the scale of the first HRP feature was 18.4 and the translation of the first HRP features varied between 1.28 and 1.30. Figure 13c shows a histogram of the translation parameter of the first HRP feature for the 100 noisy realizations. This result indicates that the HRP features are robust in the presence of additive Gaussian noise.

5 HRP with Wavelet Packet Dictionaries

In this section, the structure of wavelet packet dictionaries will be described and the HRP algorithm using wavelet packet dictionaries will be demonstrated.

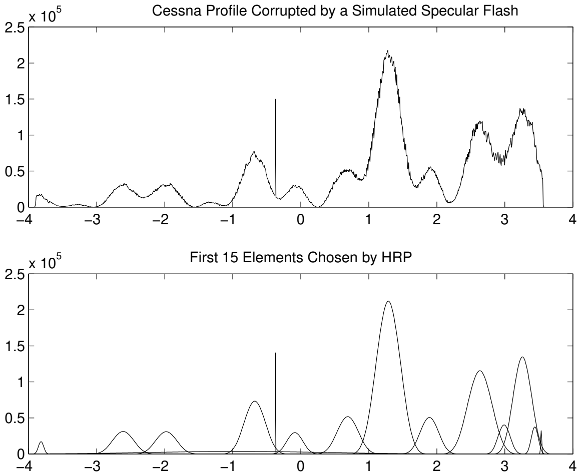
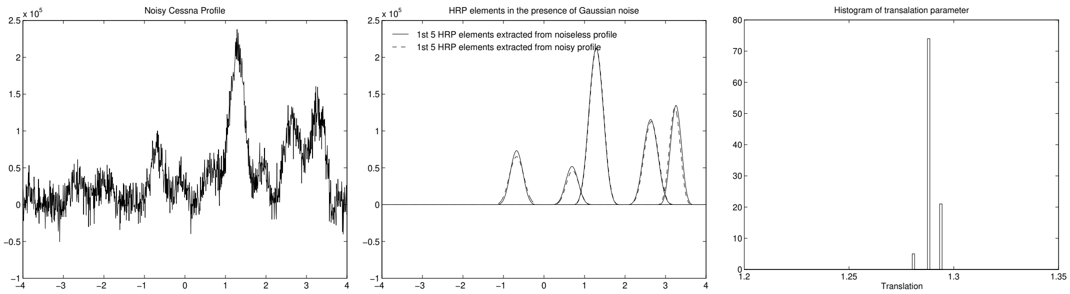


Figure 12: HRP decomposition of Cessna profile corrupted by a simulated specular flash.



(a) (b) (c)

Figure 13: (a) Cessna profile corrupted by additive Gaussian noise. (b) Comparison of HRP features extracted from a Cessna profile corrupted by Gaussian noise with HRP features extracted from the noiseless profile. (c) Histogram of the translation parameter of the first HRP element extracted from 100 noisy Cessna profiles. Noise in each profile is additive Gaussian noise with $\sigma = 10^5$.

5.1 The Wavelet Packet Dictionary Structure

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 [9, 14]. The wavelet packet dictionary is a collection of the wavelet functions, $\psi_j(x)$, and scaling functions, $\phi_j(x)$, used in the wavelet packet decomposition. 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 by the conjugate mirror filters h_1 and h_2 [8]. That is,

$$\phi_{j-1}(x) = \sum_{n=-\infty}^{+\infty} h_1[n]\phi_j(x - 2^{-j}n) \tag{35}$$

$$\psi_{j-1}(x) = \sum_{n=-\infty}^{+\infty} h_2[n]\phi_j(x - 2^{-j}n) \tag{36}$$

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. 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.

5.2 Simulated Examples

In this section, we show that HRP with wavelet packet dictionaries is also able extract signal structure and highlight the strengths and weaknesses of HRP with wavelet packet dictionaries. Note that the HRP algorithm with wavelet packet dictionaries exactly as before and the intuition developed for cubic b-spline

[8] 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.

dictionaries translates in a straightforward way to wavelet packet dictionaries.

5.2.1 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 14a. This example is similar to an example considered in [2]. This signal is the sum of four elements : a Dirac, a sinusoid and two wavelet packet atoms which are closely spaced in time. The dictionary used is a Symmlet wavelet packet dictionary. We will use a time-frequency plane to display the decompositions chosen by each technique. In the time-frequency plane, each element is represented by a rectangle where the weight of the element in the decomposition determines the darkness of the rectangle, the scale of the element determines the dimensions of the rectangle, the frequency and translation determine the location. Figure 14b shows the time-frequency plane representation of the elements chosen by HRP. The four elements of the signal are clearly visible : the horizontal line is the sinusoid, the vertical line is the Dirac, and the two rectangles are the two wavelet packet atoms which are closely spaced in time. Thus, the HRP decomposition consists of precisely those elements used to synthesize the signal and is a sparsity preserving decomposition. Similarly, the BP decomposition shown in Figure 14c resolves all four elements. The HRP and BP decompositions are identical, but HRP improves on the BP computation time by a factor of four. In contrast, the MP decomposition shown in Figure 14d is not sparsity preserving. The MP algorithm is able to extract the sinusoid and the Dirac, but is unable to resolve the two elements which are closely spaced in time. Instead of choosing the two wavelet packet elements, MP chooses five elements that are clustered around the correct location but do not match the physical features of the signal.

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

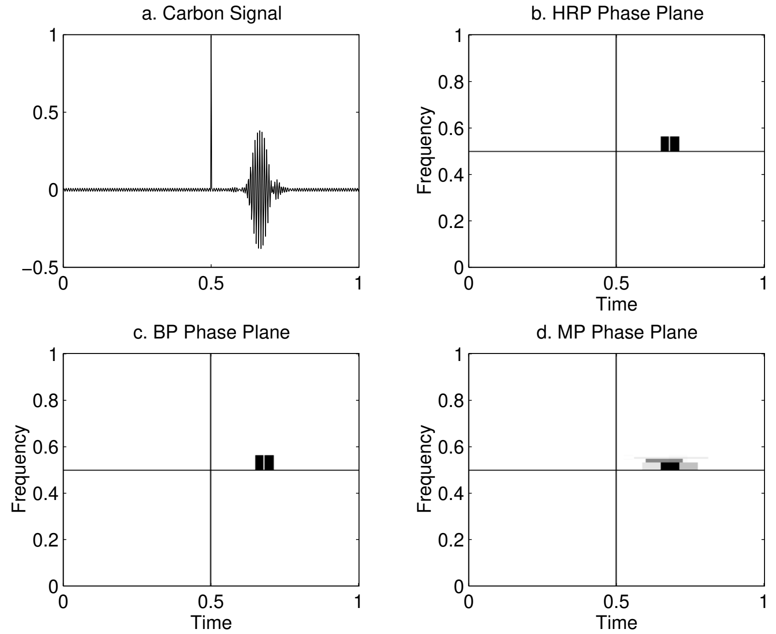


Figure 14: Results for the carbon signal. (a) The carbon signal which consists of the sum of four dictionary elements. (b) The HRP decomposition which resolves all four elements. (c) The BP decomposition which also resolves all four elements. (d) The MP decomposition which blurs the elements of the signal.

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. To summarize, the BP algorithm provides better decompositions but requires more computations in general. In a number of cases where either resolution in space or resolution in frequency is desired, HRP does as well as BP but with a much smaller computational burden.

5.2.2 The Gong Signal

Figure 15a shows a gong signal. As was mentioned in Section 4.2, 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

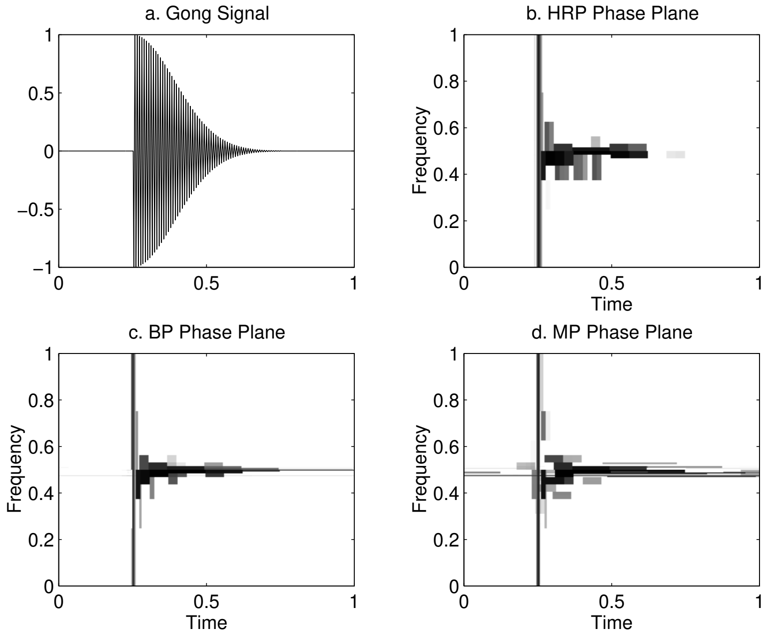


Figure 15: (a) The gong signal. (b) Time-Frequency plane for HRP. (c) Time-Frequency plane for BP. (d) Time-frequency plane for MP.

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 15b-d show the time-frequency plane results for HRP, BP, and MP, respectively. The signal was analyzed using a wavelet packet dictionary constructed from the Daubechies six tap wavelet. We begin by discussing the BP decomposition. BP captures the point of the attack by placing several elements which are concentrated in time around $t = 0.25$ where the attack of the signal begins. In addition, BP does not place any elements prior to the $t = 0.25$ and therefore the decomposition does not exhibit a pre-echo effect. The elements in the BP decomposition with $t > 0.25$ capture the correct frequency of the modulation and are well concentrated around this frequency. Thus, BP gives a decomposition which qualitatively displays the structure of the signal. The HRP decomposition again captures the point of the attack by placing several elements which are concentrated in time around $t = 0.25$. HRP does not include any elements prior to the attack of the signal. However, qualitatively one might say that HRP does not do as well as BP in

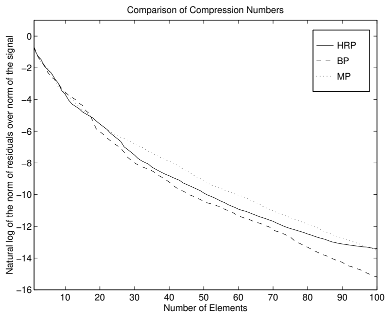


Figure 16: Rates of decay of the three methods.

capturing the correct frequency of the modulation since the HRP elements with $t > 0.25$ are not as well concentrated around the correct frequency of the modulation. MP introduces several elements prior to the attack of the signal. That is, the MP decomposition includes several elements with $t < 0.25$, prior to the attack of the signal. As a result, there will be 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. Further, the MP decomposition is not as well concentrated around the correct frequency of the modulation as BP. Comparing the rates of decay of the three methods (see Figure 16), 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.

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 (15), the inner products $< Rf, g_i >$ for all $g_i \in H$ can be computed as

$$< Rf, g_i >=< f, g_i > -S(f, g_{\gamma_0}) < g_{\gamma_0}, g_i > . \tag{37}$$

Since each of the terms on the right hand side of (37) 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

Existing approaches from function approximation did not meet our feature extraction goals. MP failed to 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. The authors would also like to thank David Donoho's group at Stanford University for providing us with the WaveLab and Atomizer software packages.

A The HRP Similarity Measure

The element which maximizes $\|R^n f - R^{n-1} f\|$ under constraints (16) and (17) also maximizes the new similarity measure $|S(f, g_\gamma)|$ as given in (12) and (13). Consider the first stage residual Rf and let $R_\gamma f$ be the residual produced by choosing some dictionary element g_γ. That is,

$$R_\gamma f = f - S(f, g_\gamma) g_\gamma. \tag{38}$$

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

$$\|R_\gamma f - f\| = |S(f, g_\gamma)|. \tag{39}$$

We begin by showing that for any dictionary element, $S(f, g_\gamma)$ as defined in (12) and (13) maximizes $\|R_\gamma f - f\|$ under constraints (16) and (17). Assume for now that

$$\frac{< f, g_i >}{< g_\gamma, g_i >} > 0 \qquad \text{for all } g_i \in I_\gamma(k) \tag{40}$$

For any dictionary element, constraint (16) may be simplified as follows

$$\begin{aligned} | < R_\gamma f, g_i > | &\leq | < f, g_i > | \qquad \text{for all } g_i \in I_\gamma(k) & (41) \\ | < f, g_i > - S(f, g_\gamma) < g_\gamma, g_i > | &\leq | < f, g_i > | & (42) \\ |1 - S(f, g_\gamma) \frac{< g_\gamma, g_i >}{< f, g_i >}| &\leq 1 & (43) \\ 0 \leq S(f, g_\gamma) &\leq \frac{2 < f, g_i >}{< g_\gamma, g_i >} & (44) \end{aligned}$$

where the last line follows because of (40). Further, for any dictionary element, constraint (17) may be simplified as

$$\begin{aligned} \text{sign}(< Rf, g_i >) &= \text{sign}(< f, g_i >) \qquad \text{for all } g_i \in I_\gamma(k) & (45) \\ < Rf, g_i >< f, g_i > &\geq 0 & (46) \\ (< f, g_i > - S(f, g_\gamma) < g_i, g_i >) < f, g_i > &\geq 0 & (47) \\ S(f, g_\gamma) &\leq \frac{< f, g_i >}{< g_\gamma, g_i >} & (48) \end{aligned}$$

where the last line follows because of (40). 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 (12) and (13) maximizes $\|R_\gamma f - f\|$ under constraints (16) and (17).

Further, the single dictionary element which maximizes $\|R_\gamma f - f\|$ under constraints (16) and (17) is the same one which maximizes $|S(f, g_\gamma)|$.

References

[1] R. Bergevin and M. 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.

[2] S. Chen and D. Donoho. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, To appear. Also available via ftp at playfair.stanford.edu.

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

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

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

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

[7] A. Gupta, G. Funka-Lea, and K. 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.

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

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

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

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

[12] M. Menon, E. Boudreau, and P. Kolodzy. An automatic ship classification system for isar imagery. *Linoln Laboratory Journal*, 6(2), 1993.

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

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