DIMENSIONALITY REDUCTION OF MULTI-SCALE FEATURE SPACES USING A SEPARABILITY CRITERION

Kamran Etemad and Rama Chellappa

Center for Automation Research
Department of Electrical Engineering
University of Maryland
College Park, Maryland 20742

ABSTRACT

An algorithm for classification task dependent multiscale feature extraction is suggested. The algorithm focuses on dimensionality reduction of the feature space subject to maximum preservation of classification information. It has been shown that, for classification tasks, class separability based features are appropriate alternatives to features selected based on energy and entropy criteria. Application of this idea to feature extraction from multi-scale wavelet packets is presented. At each level of decomposition an optimal linear transform that preserves class separabilities and results in a reduced dimensional feature space is obtained. Classification and feature extraction is performed at each scale and resulting "soft decisions" are integrated across scales. The suggested scheme can also be applied to other orthogonal or non-orthogonal multiscale transforms e.g. local cosine transform or Gabor transform. The suggested algorithm has been tested on classification and segmentation of some radar target signatures as well as textured and document images.

1. INTRODUCTION

The observation about the economy of clues in human's recognition, and the fact that classification systems with small number of parameters have better generalization, are computationally more cost effective and can be trained and adapted at higher speeds, are motivations for dimensionality reduction techniques. Thus, it is usually advantageous to sacrifice some information in order to keep the number of system parameters to a minimum. With this observation and also motivated by the success of wavelet based classification [1][2][3] systems and their biological plausibility, this paper addresses the optimal extraction of small sized feature sets from a tree structured wavelet packet decomposition of signals. The multi-scale dimensionality reduction idea can be used for both orthogonal and non-orthogonal library of local basis functions e.g. local Sine/Cosine functions, Gabor functions and even composite and redundant libraries of basis. It can also applied to other tasks e.g. classification of acoustic transients and biomedical and satellite images. The approach focuses on the exploitation of class specific differences obtained through inspection of a pre-defined class-separation [4] attainable from the wavelet packet tree and to find a linear map that provides the smallest set of features relative to which the given collection of signals shows the largest class separability. This in turn results in simple and efficient classification.

After multi-scale (e.g. wavelet or Gabor based) feature vectors are computed at each scale, the algorithm for dimensionality reduction can be applied either to each scale separately or to all of them together. The resulting feature vectors are used in a multi-resolution classification scheme where soft decisions made at different scales are combined to provide more confident results. The general non-linear mapping capability of multi-layer neural networks can be utilized to approximate membership functions by adjusting the network parameters to form the desired soft decision boundaries between clusters. Resulting soft classifiers are used in the multi-scale context dependent classification/segmentation system. In segmentation experiments soft decisions on a context area around each block are also incorporated in the voting process. The final decision is based on the majority of accumulated votes. Detailed description of this method is given in [6].

2. WAVELET PACKET BASED SIGNAL REPRESENTATIONS

The optimal representation of signals in time-frequency plane (or the so called *Phase Plane*[7]) is an active area of research, where the optimality is a task dependent issue. In most time-frequency decompositions, signals are projected onto a set of waveforms or time-frequency atoms [7]. A general family of time-frequency atoms

The support of this research by the Advanced Research Projects Agency (ARPA Order No. A550), under contract MDA 9049-3C-7217, is greatfully acknowledged.

0-7803-2431-5/95 $4.00 © 1995 IEEE

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

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

Therefore any element of the dictionary is of the form:

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

and can be identified by the triple $\gamma = (s, \xi, u) \in \mathbf{\Gamma} = (\mathbf{R}^+ \times \mathbf{R^2})$; where s, ξ and u represent scaling, modulation and translation parameters respectively [7]. These waveforms form a dictionary $\mathbf{D} = \{\mathbf{g_\gamma(t)} : \gamma \in \mathbf{\Gamma}\}$ of basis which may or may not be orthogonal or even complete and may or may not have a tree structure. The waveforms $\{g_{\gamma_n}\}$ must be selected adaptively based on the local properties of desired signals, so that the expansion coefficients provide the desired information most "efficiently".

Wavelet transforms and their generalization, called wavelet packets [1], are examples of tree structured local basis which provide signal analysis through smooth partitioning of the frequency axis. Wavelet packet analysis corresponds algorithmically to adaptive multi-rate filtering schemes and are numerically as fast as the FFT algorithms [1]. In designing wavelet packet trees, one either starts from the most refined sub-space decomposition and moves upward in the tree by merging "adjacent" nodes "appropriately", or starts from the root and performs iterative decomposition at each node to its subspaces if it is "appropriate". Depending on the task, the "appropriate" decision about further decomposition of subbands/nodes can be made based on different criteria. Decomposition toward maximizing "entropy" [1] or energy compaction, minimizing rate-distortion function or maximizing class separation [5], and also decomposition of subbands with dominant energies are some of suggested criteria [3,4]. In classification tasks, one may observe relatively high energy subbands on which the desired signals are quite similar and subbands of relatively low average energy that contain significant information about the differences between the signals, so class separability criteria seems to be a sensible choice. The tree structure obtained based on class separability may not be optimal or even sub-optimal for representing or approximating individual signals and it does not even need to provide a "complete basis".

3. DIMENSIONALITY REDUCTION

In order to design a simple and efficient classification and segmentation scheme one has to select features that are most effective in showing the salient differences between the signals. This selection may or may not be appropriate for other tasks such as approximation or compression. In other words the selection must give the best minimal set of features in terms of the separability of signal clusters in the feature space. Examples of quantitative measures of class separability are Bayes error, Bhattacharya distance, divergence based or variational distribution distances and scatter matrix based measures [4].

Unlike Mean Square Error(MSE) which is the most widely used criterion for signal representation, class separability measures are typically invariant under any non-singular, linear or non-linear, transformation. However any non-singular mapping used for dimensionality reduction results in losing some of classification information. Our objective is to find the mapping that for a given reduction in space dimension provides the maximum class-separability. In other words we are searching among all possible singular transformations for the best subspace which preserves class-separations as much as possible in the lowest possible dimensional space. A simplified and yet elegant way of formulating criteria of class separability is based on within and between class scatter matrices which are used widely in discriminant analysis of statistics. The *Within-class Scatter Matrix* shows the scatter of sample vectors (V) of different class around their respective mean/expected vectors M,

$$\begin{aligned} S_w &= \sum_{i=1}^{L} Pr\{C = C_i\}\Sigma_i \quad (3) \\ \text{where } \Sigma_i &= E[(V - M_i)(V - M_i)^T | C_i] \end{aligned}$$

The *Between-class Scatter Matrix* shows the scatter of the conditional mean vectors M_i's around the overall mean vector M.

$$S_b = \sum_{i=1}^{L} Pr\{C = C_i\}(M - M_i)(M - M_i)^T \quad (4)$$

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

$$\begin{aligned} J_1 &= tr(S_w{}^{-1} S_b) \quad (5) \\ J_2 &= ln|S_w{}^{-1} S_b| \quad (6) \end{aligned}$$

In our experiments J_1 is used but the same results hold for J_2. We denote the objective function computed over subspace V as J_V.

So we are seeking a linear transformation A from $\mathbf{R^n}$ to $\mathbf{R^m}$ with $m < n$ such that.

$$\begin{array}{rcl} A : X \subset \mathbf{R^n} & \rightarrow & Y \subset \mathbf{R^m} \\ min_A\{J_X & - & J_{A^T X}\} \end{array} \quad (7)$$

so A optimizes J_Y, i.e. minimizes the drop in cost $J_X - J_{A^T X}$ incurred by the reduction in the feature space dimensionality. It can be shown that for such an optimum A;

$$\{\lambda^Y{}_i\} \subset \{\lambda^X{}_j\} \; i = 1, ..., m \; , \; j = 1, ..., n \quad (8)$$

This observation and the fact that

$$J_Y = tr(S^Y) = \sum_{i=1}^{m} \lambda^Y{}_i \quad (9)$$

suggest that one can maximize (or minimize) J_Y by taking the largest(or smallest) m eigenvalues of S^X. Thus the corresponding eigenvectors form the transformation matrix A. In other words the optimal linear transformation from $\mathbf{R^n}$ to $\mathbf{R^m}$ based on our selected separation measure results from projecting the feature vectors X onto m eigenvectors corresponding to the m-largest eigenvalues of the separation matrix S^X. These optimal vectors/direction can be obtained from a rich enough training set and can be updated in time if needed. Note that the dimensionality m of resulting feature vectors is

$$rank(S) = min(n, L - 1) \quad (10)$$

or less, where L is the number of classes in our training set.

If the input signal representation has been obtained through linear operations or "projections" [7] one can absorb the matrix A into those operations. For example if projections of signals onto a set of multiscale "templates" $\{\phi_i, i = 1, .., n\}$ are used, then application of A to those templates, $\{A^T \phi_i, i = 1, .., m < n\}$ provides few number of "composite waveforms" on which the projections of input signals show the largest differences i.e.

$$\begin{array}{rcl} V & = & \{v_i\} = \{< s, \phi_i >\} \quad (11) \\ U & = & \{u_i\} = AV = \{Av_i\} = \{< s, A\phi_i >\} \quad (12) \end{array}$$

The original library of multi-scale basis can be a redundant dictionary composed of several wavelet packet basis, Local Sine/Cosine functions or family of Gabor functions. Also "composite" waveforms generated using this method are task dependent and do not in general have any specific structure like wavelet tree structure. They can be stored as a set of multi-scale signal templates/vectors to be used in signal projection and feature computation processes.

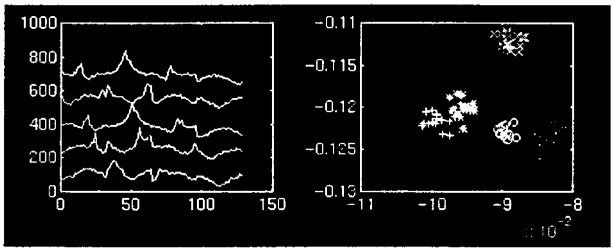

Figure 1: Example of radar target signatures for five different classes of objects (left), biases added for clarification, and separated clusters in the selected feature space (right).

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

$$\begin{array}{rcl} \phi_{(\sigma,f,d)} & = & exp(-\frac{(t-d)^2}{2\sigma}) \; cos(2\pi f(t-d)) \quad (13) \\ \Phi & = & \{\phi_\gamma\}_\Gamma \; , \; \Gamma = \{\gamma\} = \{(\sigma, f, d)\} \quad (14) \end{array}$$

and features are computed based on inner products or projections then a small set of multiscale templates for classification can be obtained based on linear combinations of Gabor wavelets according to rows of the matrix A. Resulting composite templates may not be symmetric and may not resemble any known local basis.

4. EXPERIMENTS

To show the effectiveness of the suggested feature extraction, it is applied to image texture classification and segmentation as well as classification of radar target signatures. Both balanced and pruned wavelet packet trees are used in these tests. On each subband/node second and third central moments μ_2 and μ_3 are considered as feature elements. For texture classification and segmentation tests these moments are computed over small windows on the decomposed image.

Figure 1, shows the separated clusters for five classes of radar targets where only the two most important features are used. Figure 2, compares the cluster separations in the feature space when the feature vectors are selected using the suggested class separability based linear map or energy based approaches. The distance between clusters allows us to have good classification results even in the presence of small noise. In both tests simple neural networks is used as "soft classifier". These networks have 2 input, 3 hidden and 3 or 5 output units for five classes of targets or three classes of textures respectively. Results shows only 0% to 2% misclassification.

The task of signal segmentation is more involved than classification. The window size W should be large

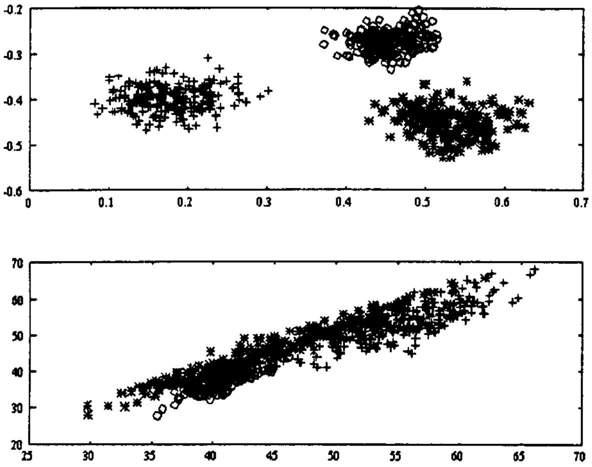

Figure 2: Clusters in the feature space for three textures. Features obtained from separability based method (top) and features obtained from highest energy subbands (bottom).

enough to cover spatial variations that characterize the difference between signals but small enough so that it can have a good temporal/spatial localization. Because of the down sampling involved in the transform the corresponding window sizes on sub-bands at the k^{th} level of the tree are $W/(2^k)$. Therefore the depth of the tree is limited by the size of input window and the nature of the signals to be classified. Also the order of the filters in filter bank implementations should be smaller than window size to avoid the dominance of the window boundary effects on the resulting feature computation. Therefore for segmentation tasks filters with smaller number of taps are preferred. Choosing small size windows for good localization results in higher classification errors. So in order to improve the final performance additional techniques are needed. Figure 3. shows the segmentation results for the same three textures; where window size of 16×16 pixels, with 8 pixel overlaps, is chosen and decision integration is used [6]. Also Figure 4. shows an application of suggested scheme to document page segmentation where text and pictures are treated as two different textures.

5. RESULTS AND DISCUSSION

A method for efficient classification of a variety of 1D and 2D signals such as radar signatures and textured and document images has been suggested. The idea of applying dimensionality reduction from linear discriminant analysis to multi-scale feature spaces is suggested and studied. The method is tested on wavelet packet based features. Also very good segmentation results with small number of features are reported. The results obtained in several signal and image classification tests show that multi-scale separation based feature extraction is an appropriate general framework which provides good results using simple feature sets and classifiers. Also suggestions on extensions of these ideas to libraries of redundant and non-orthogonal local basis are given.

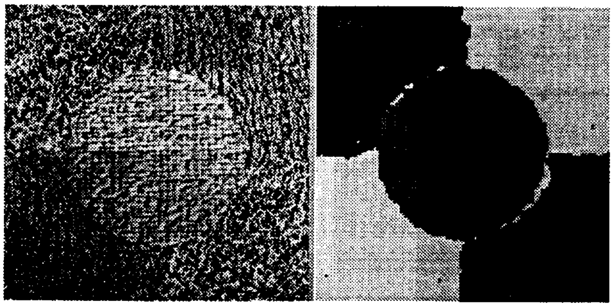

Figure 3: Example of a texture segmentation using the reduced two dimensional feature space.

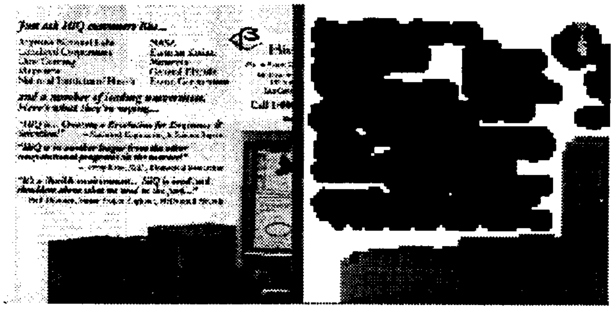

Figure 4: Example of a document page segmentation using our texture based approach.

6. REFERENCES

[1] R.R. Coifman and M.V. Wickerhauser,"Entropy based algorithms foe best basis selection", IEEE Trans. Information Theory, Vol. 38, March 92, pp. 713-718.

[2] R.E. Learned, W.C. Karl and A.S. Willsky," Wavelet packet based transient signal classification", Proc. IEEE Conf. on time scale and time frequency analysis, pp. 109-112, 1992.

[3] T. Chang and C.C.J. Kuo," Texture analysis and classification with tree structured wavelet transform" , IEEE Trans. Image Processing, Vol. 2, October 93, pp. 429-440.

[4] K. Fukunaga,"Introduction to Statistical Pattern Recognition" 2nd ed. Academic Press, 1990.

[5] K.Etemad, R. Chellappa,"Separability based tree structured basis selection for textures classification", Proc. IEEE-ICIP94, Austin, Texas, pp.442-445,1994.

[6] K.Etemad, R. Chellappa and D.Doermann,"Document page decomposition by integration of distributed soft decisions", Proc. IEEE-ICNN94, Orlando, FL, pp. 4022-4027, 1994.

[7] S.Mallat and Z. Zhang, "Matching Pursuit with Time Frequency Dictionaries", IEEE Trans. Signal Processing, the Special Issue on Wavelets and Signal Processing, Vol. 41, pp. 3397-3415, December 1993.