OPTIMIZED LOCAL DISCRIMINANT BASIS ALGORITHM

Kamyar Hazaveh
Electrical & Computer Engineering Dept.
Ryerson University, Toronto, Canada
kamyar_hazaveh@ieee.org

Kaamran Raahemifar
Electrical & Computer Engineering Dept.
Ryerson University, Toronto, Canada
kraahemi@ee.ryerson.ca

Abstract- Local discriminant bases method is a powerful algorithmic framework for feature extraction and classification applications that is based on supervised training. It is considerably faster compared to more theoretically ideal feature extraction methods such as principal component analysis or projection pursuit. In this paper an optimization block is added to original local discriminant bases algorithm to promote the difference between disjoint signal classes. This is done by optimally weighting the local discriminant basis using steepest decent algorithm. The proposed method is particularly useful when background features in the signal space show strong correlation with regions of interest in the signal as in mammograms for instance.
Keywords: Best basis, Local discriminant basis, Local feature extraction, The steepest decent method, Pattern recognition, Time-frequency analysis, Wavelet packet and Mammography.

1. INTRODUCTION

In 1994, Coifman and Saito [8], [9], developed the local discriminant bases (LDB) method as Wickerhauser's best basis algorithm [1], [2] counterpart in feature extraction and signal classification applications. Both best basis and LDB methods use tree-structured collections of basis functions (atoms) called dictionaries. These are redundant sets of basis functions that are localized in time and frequency. Examples of time-frequency dictionaries include wavelet packets and local trigonometric bases. For a complete discussion of these basis function collections and their properties the interested reader is referred to [2], [20].

The best basis method can be tuned to adaptively choose the best set of basis functions so that the entropy of the signal coordinates is minimized. The same basis selection task is done by Karhunen-Loeve transform (KLT) [4] developed in 1965 that is better known as Principal Component Analysis (PCA). The advantage of best basis methods over KLT is twofold. By exploiting the structured nature of time-frequency dictionaries, Coifman and Wickerhauser developed a divide-and-conquer algorithm for best basis search with a computational cost of $O(n \log n)$ that is superior to KLT computational cost *i.e.* $O(n^3)$ [4]. Besides, in contrast to KLT, the best basis method is localized in time and frequency making it suitable for totally non-stationary process analysis. A local KLT has been developed by Coifman and Saito [5] but it suffers from the same high computational cost $O(n^3)$.

Later on, best basis method was used for de-noising [6], [7]. In 1994, Saito [8, Chapter 3] developed an algorithm for simultaneous de-noising and compression of signals. De-noising was the start of a new generation of best basis algorithms for signal classification problems. Compared to other supervised feature extraction solutions such as projection pursuit [23], [24], LDB is modest in the sense that it picks a set of good coordinates from a finite collection rather than a sequence of the absolutely best projections without constraints. The LDB concept has increasingly gained popularity and has been applied to a variety of classification problems including biomedical [7], geophysical [11], sonar [12], radar [10], [22] and military [13], [14], [22] application areas. In 1996, Coifman and Saito discovered a counterexample in which, LDB was unable to distinguish between two classes of synthetic signals. Thereafter several improved versions of the original LDB were developed [21].

The authors have noticed that an additional stage in LDB algorithm for assigning weights to selected basis functions helps to improve the accuracy of LDB method. This feature boosting is especially useful when background data has strong correlation with regions of interest in signal space under study [25]. The method of gradient decent is employed to find the optimal weight for the selected basis functions. The effectiveness of the proposed optimization block is proved by experiment, *i.e.* near 40% decrease in misclassification rate. Yoshida [25] has employed the same technique for improving the performance of matching pursuit method [26] for extraction of microcalcifications from mammograms. The usefulness of our proposed scheme is currently being tested on a mammography database.

This is how this paper is organized. In section 2 the LDB algorithm and its improved version are reviewed. Section 3 is devoted to the development of the new idea of boosting features by using optimally weighted basis functions. Simulation results are given in section 4. Section 5 discusses conclusions and future work.

2. LOCAL DISCRIMINANT BASIS ALGORITHM

In this section we review the general problem of feature extraction. Suppose that we have a space $X \subseteq R^n$ of input signals and a space Y of class labels. The goal is to construct a classifier d: $X \rightarrow Y$ that assigns the correct class label to each input signal. The optimal classifier is known to be the so-called Bayes classifier. However, Bayes classifier is impossible to construct due to high dimensionality of the real signals [21]. Examples of high dimensional signals are medical X-ray tomography ($n=512^2$), seismic signals ($n=4000$) and a speech segment ($n=1024$). Faced with the dimensionality and having such difficulty in constructing the Bayes classifier, the extraction of important features becomes essential. As Scott mentions in [3, Chapter 7] multivariate data in R^n are almost never n-dimensional and there often exists lower dimension structures of data. So based on the application whether compression or classification, the problem always has a lower intrinsic

dimension. It is important to note that intrinsic dimension is an application-oriented quantity. Coifman and Wickerhauser basis selection scheme followed by a simple thresholding on the amplitude of the coefficients can result in significant signal dimension reduction. Saito's LDB algorithm helps to reduce the dimensionality of the problem for the feature extraction and classification tasks. The feature extractor in LDB is formulated as $d = g o \Theta_m o \Psi$, where Ψ is an orthogonal transformation, Θ_m is a projection operator into m most important coordinates and g is a standard classifier. By a dimension reducer engine such as LDB, *i.e.* $\Theta_m o \Psi$, we select and keep the most important basis vectors according to the classification task and discard the nonessential coordinates. The classifier, g, can be *Linear Discriminant Analysis* (LDA) [15], *Classification and Regression Trees* (CART) [16], *k-nearest neighbour* (k-NN) [17] or *artificial neural networks* (ANN) [18].

LDB method first decomposes available training signals from different classes in a time-frequency dictionary, which is a large collection of bases functions, *i.e.* wavelet packets or local trigonometric basis. Then signal energies are accumulated for each class separately to form a time-frequency energy distribution per class. We assume that there are only two classes of signals. The generalization of the method to more signal classes is straightforward.

Let us assume that ω is a typical basis function in a time-frequency dictionary. The time-frequency distribution energies of class *1* and class *2* along ω are designated by Γ_ω^1 and Γ_ω^2 respectively.

In the original LDB algorithm, the tree-structured time-frequency dictionary is pruned by using a discrimination measure such as

- *l^2-distance:*

$W(\Gamma_\omega^1, \Gamma_\omega^2) = \| \Gamma_\omega^1 - \Gamma_\omega^2 \|^2$,

- *Relative entropy (or Kullback-Leibler divergence):*

$D(\Gamma_\omega^1, \Gamma_\omega^2) = \Gamma_\omega^1 \log(\Gamma_\omega^1 / \Gamma_\omega^2)$,

- *Symmetric relative entropy (or J-divergence):*

$J(\Gamma_\omega^1, \Gamma_\omega^2) = D(\Gamma_\omega^1, \Gamma_\omega^2) + D(\Gamma_\omega^2, \Gamma_\omega^1)$.

In the first step LDB is the children nodes at the lowest part of the tree in Fig.1. Then the discrimination measures of each two children nodes are compared to their parent's. If the sum of the discrimination measures of the children nodes is higher than their parent's, we keep the children nodes. Otherwise, the parent node is chosen as the LDB. Once a complete basis (LDB) is selected, we further choose $m(< n)$ atoms from the LDB. A typical m would be $n/10$. The simplest way of choosing m atoms from a collection of n atoms is to sort them in the order of decreasing discrimination power and to retain the first m atoms. It's important to note that the functionality of LDB lies in the over-complete nature of binary tree dictionaries [2]. This redundancy enables the introduction of different search algorithms with different discriminant measures to prune the binary tree in an optimal manner that is tailored to the particular application. It is possible to construct a simple classification problem that is intractable by original LDB algorithm [21]. Therefore it is sometimes necessary to consider the distribution of expansion coefficients for individual coordinates. The original LDB algorithm measure is based on the differences of mean class energy of projections. It is possible to use a measure based on the differences between probability distribution functions.

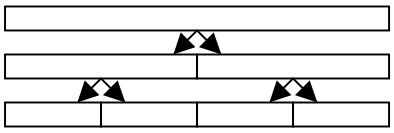

Fig.1 A binary tree-structured dictionary

For a complete treatment of the problems that lead to the introduction of Type II, Type III and Type IV LDB algorithms the interested reader is referred to [21]. Type II LDB has been applied to real world applications [10], [12]. The initial problem of Type II would be the estimation of pdf's from the available database. Suppose that p_ω^1 and p_ω^2 are the pdf's of *class 1* and *class 2* signals in ω direction respectively. Here is a discrimination measures proposed by Saito [21].

- *Relative entropy (or Kullback-Leibler divergence):*

$D(p_\omega^1, p_\omega^2) = \int p_\omega^1 \log(p_\omega^1 / p_\omega^2)\, dx$.

3. OPTIMALLY WEIGHTED LDB ALGORITHM

The basis functions selected by LDB capture some of the common features or background components as well. This becomes particularly important when background structures are highly correlated to the features of interest. In this section we introduce a method of boosting desired distinguishing features among different classes by optimally weighting the basis functions.

Let us assume that $\psi_\gamma, 1 \le \gamma \le N$ is the collection of basis functions that are obtained as the output of LDB algorithm. We study the effect of assigning a weight sequence such as $\omega_\gamma, 1 \le \gamma \le N$ to these basis functions. The goal is to promote the discrimination power of $\omega_\gamma \psi_\gamma, 1 \le \gamma \le N$ hence improving the classification accuracy. To this end, we minimize the difference between samples of each signal class with the so-called *teacher signals* [25]. Teacher signals for each class are the average of the signal class windowed at the regions of interest; therefore there are a limited number of them. The error function is given in (1). In this expression, α_γ^k's are the LDB coordinated of signal S^k. The corresponding teacher signal is designated by T^k,

$$E(\omega) = \frac{1}{K} \sum_k \sum_x (S^k(\omega, x) - T^k(x))^2, \quad (1)$$

$$S^k(\omega, x) = \sum_\gamma \alpha_\gamma^k \omega_\gamma \psi_\gamma(x). \quad (2)$$

In (1) and (2) k ranges over the samples used in the optimization phase whereas x ranges over time or pixels in an image processing application and K is the number of samples used for optimization. The error function $E(\omega)$ can be minimized by a gradient decent algorithm to yield an optimal set of weights that maximally separate different signal classes. The partial derivative of $E(\omega)$ in terms of a weight ω_γ is given by (3),

Therefore the weight update formula for gradient decent algorithm will be as in (4),

$$\frac{\partial E}{\partial \omega_\gamma} = \frac{2}{K}\sum_k \sum_x (S^k(\omega, x) - T^k(x))\alpha_\gamma^k \psi_\gamma(x), \quad (3)$$

$$\omega_\gamma \rightarrow \omega_\gamma - \eta \frac{\partial E}{\partial \omega_\gamma}. \quad (4)$$

Here η is a user-defined learning rate. For the purpose of illustration we consider example 5.2 form [9] known as "triangular waveform classification" problem [16]. The problem is to classify three classes of the signals generated by the statistical process given in (5),

$$\begin{aligned} c(i) &= (6+\varsigma).\chi_{[a,b]}(i) + \varepsilon(i) \\ b(i) &= (6+\varsigma).\chi_{[a,b]}(i).\frac{i-a}{b-a} + \varepsilon(i) \\ f(i) &= (6+\varsigma).\chi_{[a,b]}(i).\frac{b-i}{b-a} + \varepsilon(i) \end{aligned} \quad (5)$$

a is an integer-valued uniform random variable on the interval [16,32]. $b-a$ also obeys an integer-valued uniform distribution on [32,96]. ς and ε are the standard normal variates, and $\chi_{[a,b]}(i)$ is the characteristic function on $[a,b]$. $c(i)$ is called the 'cylinder' class whereas $b(i)$ and $f(i)$ are known as 'bell' and 'funnel' classes. The statistical averages of the each class in the above process over a set of 300 training signals are depicted in Fig.2. The statistical average of each class will be windowed around the regions of interest to be used as teacher signal in the proposed optimization block. The same procedure can be applied to obtain teacher images that locate microcalcifications in mammograms.

4. OPTIMIZED LDB SIMULATION RESULTS

In the beginning, local discriminant analysis was used to capture LDB for a set of 300 training signals described above. The wavelet packet was generated by a Coiflet 4 mother wavelet. Symmetric relative entropy (*J*-divergence) was used as discrepancy measure in our study. Fisher's Linear Discriminant Analysis [15] was fixed as the classifier in this study. Two sets of teacher signals were generated by windowing the statistical averages in the intervals [40,55] and [40,60] to study the effect of window size. These two sets of teacher signals are referred to as 15-point and 20-point windows in all the graphs. The number of samples used in optimization phase was fixed to 21 equally distributed between three classes (*K*=21). Another set of 300 signals was generated as a test bed for examining the effectiveness of optimization process. Fig.3 shows the results of using the first 12 most discriminative LDB vectors. The weight update process was performed on 15 and 20 first most discriminant LDB vectors as opposed to 128 to save processor's time. The experiment continued up to 100 iterations. Fig.4 and Fig.5 show similar results for classification based on 10 and 8 most discriminative LDB vectors respectively.

5. CONCLUSIONS AND DISCUSSION

The original and improved versions of local discriminant bases algorithm are reviewed. An additional feature boosting stage is added to the algorithm. The authors have applied the new technique on classification of synthetic data. Optimal weighting of local discriminant basis improved the misclassification error with reasonable number of iterations. The optimization block can further improve the efficiency and accuracy of the local discriminant basis algorithm specially when the number of selected LDB vectors is small and background data has strong correlation with signal space. The explicit improvement in accuracy proves that the process is worthwhile as a one-time optimization performed right after training phase. Authors are studying the effect and importance of varying different parameters in the experiment such as the learning rate η, number of sample signals used in the optimization phase (*K*) as well as other optimization algorithms, different wavelet packet dictionaries and discrepancy measures. The usefulness of the feature-boosting module is being studied in classification of mammograms where background structure shows considerable correlation with microcalcification patterns.

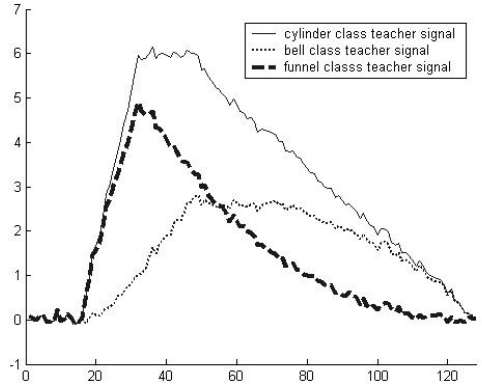

Fig.2 Statistical averages of different signal classes.

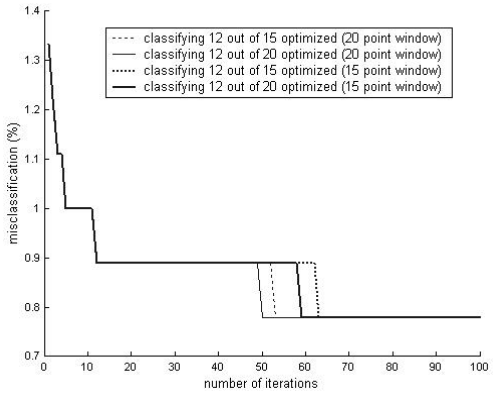

Fig.3 The narrow curves show the misclassification rate when teacher signals were windowed at [40,60]. Others are obtained when windowing is done at [40,55].

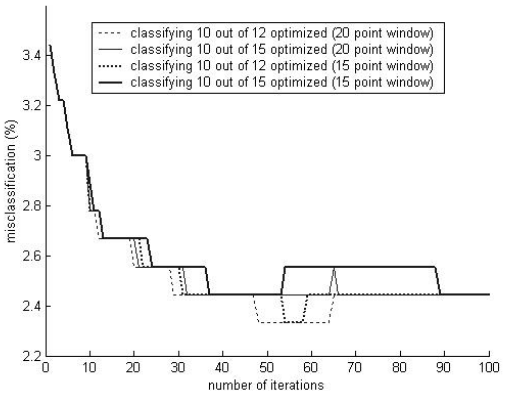

Fig.4 Repeating the experiment with 10 LDB vectors for classification.

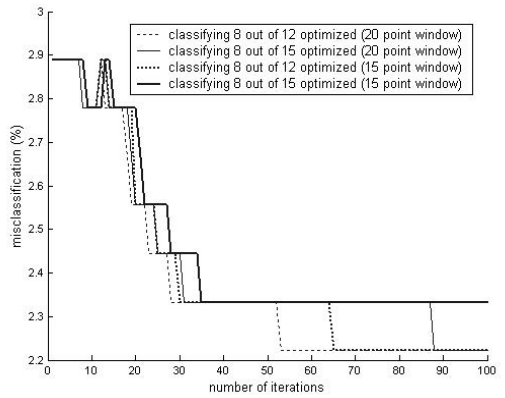

Fig.5 Repeating the experiment with 8 LDB vectors for classification.

6. REFERENCES

[1] R. R. Coifman and M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory* 38 (1992), no.2, pp 713-719.

[2] M.V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A K Peters, Ltd. Wellesley, MA, 1994.

[3] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York, 1992.

[4] S. Watanabe, "Karhunen-Loeve Expansion and Factor Analysis: Theoretical Remarks and Applications," in *Trans. 4th Prague Conf. Inform. Theory*, Publishing House of Czechoslovak Academy of Sciences, pp 635-660, 1965.

[5] R. R. Coifman, N. Saito, "The Local Karhunen-Loeve Bases," *Time-Frequency and Time-Scale Analysis, Proceedings of the IEEE-SP International Symposium on*, pp 129 –132, 1996.

[6] P. Qua, Z. Lei, Z. Hongcai, D. Guanzhong, "Adaptive wavelet based spatially de-noising," Signal *Processing Proceedings, 1998 Fourth International Conference on*, pp 486-489,1998.

[7] R. R. Coifman and M.V. Wickerhauser, "Adapted waveform "de-noising" for medical signals and images," IEEE Engineering in Medicine and Biology Magazine, v.14, n.5, pp 578 –586, 1995.

[8] N. Saito, *Local Feature Extraction and Its Application Using a Library of Bases*, Ph.D. Thesis, Dept. of Math, Yale University, New Haven, CT 06520 USA, Dec. 1994.

[9] N. Saito and R. R. Coifman, "Local discriminant bases and their applications," J. Mathematical Imaging and Vision 5 (1995), no.4, pp 337-358, Invited paper.

[10] G. Kronquist and H. Storm, "Target Detection with Local Discriminant Bases and Wavelets," *Proc. of SPIE*, v.3710, pp 675-683, 1999.

[11] N. Saito, R. R. Coifman, "Extraction of Geological Information from Acoustic Well-Logging Waveforms Using Time-Frequency Wavelets," *Geophysics*, v.62, n.6, 1997.

[12] L. S. Rogers, C. Johnston, "Land Use Classification of SAR Images Using a Type II Local Discriminant Basis for Preprocessing," *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing*, v.5, pp 2729-2732, 1998.

[13] M. L. Cassabaum, H. A. Schmitt, H. W. Chen, J. G. Riddle, "Application of Local Discriminant Bases Discrimination Algorithm for Theater Missile Defense," *Proc. of SPIE*, v.4119, pp 886-893, 2000.

[14] M. L. Cassabaum, H. A. Schmitt, H. W. Chen, J. G. Riddle, "Fuzzy Classification Algorithm Applied to Signal Discrimination for Navy Theater Wide Missile Defense," *Proc. of SPIE*, v.4120, pp 134-145, 2000.

[15] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Ann. Eugenics*, Vol. 7, pp 179-188, 1936.

[16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Chapman and Hall, Inc., New York, 1993.

[17] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, San Diego, 1990.

[18] B. D. Ripley, "Statistical aspects of neural networks," *Networks and Chaos: Statistical and Probabilistic Aspects*, O. E. Barndorff-Nielson, J. L. Jensen, D. R. Cox, and W. S. Kendall, eds., pp 40-123, Chapman and Hall, Inc., New York, 1993.

[19] P. Frossard, P. Vandegheynst, "Redundancy in non-orthogonal transforms," *Proc. of IEEE International Symposium on Inform. Theory*, n.01CH3725, p 196, 2001.

[20] Mallat, S.G., *A Wavelet Tour of Signal Processing*, San Diego Toronto: Academic Press, 1998.

[21] N. Saito and R. R. Coifman, "Improved Local Discriminant Bases Using Empirical Probability Density Estimation," Proc. *on Statistical Computing*, American Statistic. Assoc., 1996.

[22] C. M. Spooner, "Application of local discriminant bases to HRR-based ATR," *Signals, Systems and Computers, Conference Record of the Thirty-Fifth Asilomar Conference on*, v.2, pp 1067 -1073, 2001.

[23] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Trans. Comput.* 23, pp 881-890, 1974.

[24] J. B. Kruskal, "Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'", *Statistical Computation* (R. C. Milton and J. A. Nelder, eds.), Academic Press, New York, pp 427-440, 1969.

[25] H. Yashida, "Matching pursuit with optimally weighted wavelet packets for extraction of microcalcifications in mammograms," *Applied Signal Processing*, no.5, pp 127-141, 1998.

[26] S. G. Mallat, Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397-3415, 1993.