SUBSET SELECTION WITH STRUCTURED DICTIONARIES IN CLASSIFICATION

Nuri F. Ince[1], Fikri Goksu[1], Ahmed H. Tewfik[1], Ibrahim Onaran[2], A. Enis Cetin[2]

[1]Department of Electrical and Computer Engineering, University of Minnesota
200 Union St. SE, 55455, Minneapolis, U.S.A.
phone: + (1) 612-625-5006, fax: + (1) 612-625-4583, email: firat@umn.edu
[2]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey

ABSTRACT

This paper describes a new approach for the selection of discriminant time-frequency features for classification. Unlike previous approaches that use the individual discrimination power of expansion coefficients, the proposed approach selects a subset of features by implementing a classifier directed pruning of an initial redundant set of candidate features. The candidate features are calculated from a structured redundant time-frequency analysis of the signal, such as an undecimated wavelet transform. We show that the proposed approach has a performance that is as good as or better than traditional classification approaches while using a much smaller number of features. In particular, we provide experimental results to demonstrate the superior performance of the algorithm in the area of impact acoustic classification for food kernel inspection. The proposed algorithm achieved 91.8% and 98.5% classification accuracies in separating open shell from closed shell pistachio nuts and discriminating between empty and full hazelnuts respectively. Traditional methods used in this area resulted in 82% and 97% classification accuracies respectively.

1. INTRODUCTION

The time-frequency (t-f) content of the signal provides important information for classification. Since the t-f plane is a high dimensional space, using all features from this plane is not a good strategy. In the last several years there has been a growing interest to explore the time-frequency plane for classification by using adaptive strategies. The local discriminant basis an algorithm of [1] has been proposed to achieve this task. The LDB algorithm first represents a given signal for each class in a redundant manner over a single pyramidal tree structure by using wavelet packets or cosine packets. In order to find a complete representation of the signals in this redundant dictionary, a bottom to top pruning algorithm is implemented. Then the energies of expansion coefficients between classes in each node of the tree structure are evaluated by a cost function. Based on the differences between the classes the pyramidal tree structure is pruned from bottom to top such that the discrimination power between expansion coefficients in the nodes of the tree is maximized.

$$\text{PrunedTree} = \arg\max D(\psi_{j,k})$$

where the $\psi_{j,k}$ is the dual tree structure with frequency, j, and time, k, indices and D is a cost function. Once the tree is pruned a complete representation of the signal is obtained which is tuned for discrimination. This is followed by sorting the expansion coefficients according to their discrimination power and inputting a top subset to a classifier for final decision. This powerful method has been used in several applications such as EEG and EMG classification and successful results have been obtained [2, 3]. However, the LDB algorithm has many drawbacks. First, the wavelet packets (WP)/cosine packets (CP) which are used to represent the signal in the nodes of the tree structure do not satisfy the shift invariance property. Briefly, a shift in the signal generates unpredictable changes in the expansion coefficients. This behavior is not appropriate for pattern recognition applications. Furthermore, since a single tree is used, this algorithm can only adapt along a single axis such as either time or frequency. Several studies have shown that the adaptation along both axes is crucial [2]. In addition to this, post processing methods such as PCA has been applied on the sorted expansion coefficients to improve the classification accuracy [3]. Finally, the pruning and feature sorting stages do not account for the interactions/relations in between different t-f cells. The obtained complete representation may not be the best subset for the classifier.

In this paper we tackle these problems by introducing a structured subset selection approach which is based on the evaluation of the real classification performance of the selected subsets of features. Unlike prior methods, our approach is a wrapper method that implements a *classifier directed pruning* of a highly redundant set of features as shown in Figure 1. The features are obtained by computing a redundant t-f analysis of signals, such as an undecimated wavelet transform (UDWT), over a dual tree structure. As compared to the decimated wavelet transform (WT), the UDWT yields a shift invariant signal representation. In [4] the author has shown that the classification results obtained with UDWT is superior to WT. A block diagram summarizing our proposed approach is given in Figure 1. We show that our proposed approach achieves similar or better performance than traditional approaches while using a much smaller number of features. Here, we illustrate this by testing our proposed approach in two impact acoustic classification problems for food safety.

The paper is organized as follows. Details of our approach are available in the following sections. Section 2 describes

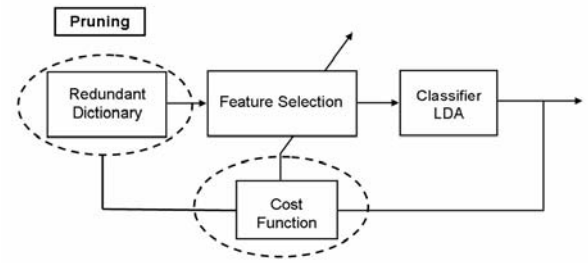


Figure 1-The block diagram of the proposed classification method

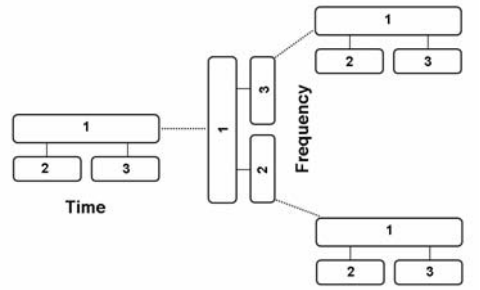


Figure 2- Dual tree structure.

the extraction of time frequency features using double tree undecimated wavelet transform. The search for the best combinations of features comprises Section 3. Both greedy search and pruning using the double tree structure with classifier feedback and another alternative without feedback are considered here for feature selection. Section 4 includes the materials used to record nuts sound signals from impact acoustics. In Section 5 we report the experimental results and compare them with previously available ones.

2. GENERATION OF STRUCTURED DICTIONARIES

Let us here describe the redundant dual tree structure where a schematic diagram is also given in Fig.2. The dual tree feature dictionary is constructed by using an undecimated wavelet transform which provides a pyramidal subband tree with different frequency resolutions. In each subband anthor pyramidal tree is utilized to obtain features with different time resolution. The details of obtaining the time-frequency features with distinct resolutions are described in the following section.

2.1 Undecimated Wavelet Transform

Discrete Wavelet Transform (DWT) and its variants have been extensively used in 1D and 2D signal analysis [5]. Here we consider the 2-band DWT. In the 2-band system, each output of the filters (subband) is downsampled by 2 for critical sampling. In our implementation of the DWT, down sampling is removed to make the UDWT shift invariant. This means that the number of samples in a subband at any level is same as that of in the original signal. The output at any level can be computed by using an appropriate filter derived from the wavelet tree structure. We find the equivalent filter for each subband using one of the Noble identities which involves downsampling.

$$\rightarrow \boxed{\downarrow K} \rightarrow \boxed{L(z) or H(z)} \rightarrow \equiv \rightarrow \boxed{L(z^K) or H(z^K)} \rightarrow \boxed{\downarrow \mathrm{K}} \quad (1)$$

In mathematical form what we are looking for is as in the following example; given the *Z*-transfrom of low pass filter *L(z)* and high pass filter *H(z),* what will be the equivalent filter corresponding to one of the subbands at the 3rd level? This means that this subband is computed using three filters each of which is either *L(z)* or *H(z)*. Assume they are *L(z), H(z)* and *L(z)* respectively. We have;

$$X(z) \rightarrow \boxed{L(z)} \rightarrow \boxed{\downarrow 2} \rightarrow \boxed{H(z)} \rightarrow \boxed{\downarrow 2} \rightarrow \boxed{L(z)} \rightarrow \boxed{\downarrow 2} \rightarrow Y(z) \quad (2)$$

And we want to find an *M(z)* such that the following holds;

$$X(z) \rightarrow \boxed{M(z)} \rightarrow \boxed{\downarrow 8} \rightarrow Y(z) \quad (3)$$

Using the Noble identity twice appropriately we reach the following equivalent input/output relationship;

$$X(z) \rightarrow \boxed{L(z)H(z^2)L(z^4)} \rightarrow \boxed{\downarrow 8} \rightarrow Y(z)\,. \quad (4)$$

This means $M(z)= L(z)H(z^2)L(z^4)$, and we can find corresponding filter *m[n]* in the time domain easily. In the way explained above we keep the output right after the equivalent filter *M(z)* before the down sampling. This constitutes the frequency branch of the dual tree UDWT.

2.2 Dyadic Time Segmentation

One can use the UDWT coefficients in each subband. However since this causes a high dimensional space, this is not a practical approach. In order to decrease the dimensionality and preserve the energy-based information in each subband, like in the frequency decomposition tree, every subband is segmented into time segments at each level with a pyramidal tree structure successively. In each time segment the sum of the squares of the samples, energy, is computed as one feature to be used in off-line training. The time segmentation explained above forms the second branch of the double tree. From now on we keep the index information of the dual tree structure to be used in the later stage for dimension reduction via pruning.

To summarize this section the reader is referred to the double tree structure in Figure 2. This double tree uses 1-level in both planes. The vertical middle boxes are the frequency subbands. Box 1 represents unfiltered original signal, box 2 represents low pass filtered signal and box 3 high pass filtered signal. Each of these subbands are segmented in time into 3 segments, as shown. Segment 1 covers the whole subband, segment 2 covers the first half of it and segment 3 the second part of it. Please note that The dual tree structure satisfies these two conditions:

- For a selected node in the frequency tree mother band covers the same frequency band width (BW) of its children

$$BW_{Mother} \supset (BW_{Child1} \cup BW_{Child2}) \quad (5)$$

- This same condition is also satisfied in time axis. For a selected node, the number of time samples (TS) of mother is equal to the number of its children.

$$TS_{Mother} \supset (TS_{Child1} \cup TS_{Child2}) \quad (6)$$

These two properties allow us to prune the tree structure. The parameters for deciding the number of features are the number of levels for frequency and time segmentation. Let **T** be the number of levels in time and **F** the number levels in frequency, there will be $\mathbf{2^{(F+1)}-1}$ subbands (including the original signal) and $\mathbf{2^{(T+1)}-1}$ time segments for each subband. This will make the total number of features $\mathbf{(2^{(F+1)}-1)x(2^{(T+1)}-1).}$

3. SUBSET SELECTION

As emphasized before the selection of features from a redundant dictionary is critical. The redundancy comes from both the UDWT coefficients and dual tree structure. As explained in the previous section the dual tree has total $\boldsymbol{NF}=\mathbf{(2^{(F+1)}-1)x(2^{(T+1)}-1)}$ number of features for each sound signal where **F** is the frequency level and **T** is the time level. The LDB approach uses feature sorting after pruning, which does not account for the interrelations between features. Furthermore, the subset of a given feature order, which is obtained with a cost function, is not necessarily the best subset for the classifier. Here we reshape the strategy for feature selection by evaluating the actual classification performance. We quantify the efficiency of each feature set by evaluating its classification accuracy by a cost measure and we use this cost to reformulate our dictionary.

Three different types of methods are considered for feature selection. The structure in Figure 1 is general for all three methods with slight variations. The left most box is the dictionary of feature vectors. (LDA) on the right is used both for classification and extracting the relationship among combinations of features. This output is fed to a cost function to measure the discrimination power for that combination of features between classes. This measure will be used to select the best feature combination among other feature combinations. In this particular study, Fisher Discrimination (FD) function is used as a cost function. The three different strategies as explained above are:

- *Type-I:* Sequentially select features from the dictionary by evaluating their classification performance with a cost function.
- *Type-II*: Sequentially select features from the dictionary by evaluating their classification performance with a cost function. In each step the dictionary is pruned for the selected index.
- *Type-III*: Prune the dual tree using the old LDB approach-based sequential feature selection.

3.1 Type-I

Type I is a Sequential Forward Selection (SFS) method. All of the feature vectors from the dictionary (from each class) enter to the LDA one by one and a corresponding cost is measured for each using the cost function. After this search is done over all ***NF*** feature vectors, the best feature is selected by comparing cost values of each feature vector. In the next step the second best feature vector which will do the best in combination with the first selected one is searched over the remaining feature vector set one by one. This procedure is run until a desired number of features are reached. Type-I uses all the boxes and connections in Figure 1 except the feedback from the cost function to the dictionary. Since no dimension reduction is implemented in the entire feature space, this approach has high computational complexity.

3.2 Type-II

Type II is a wrapper method with an additional pruning module for dimension reduction. As in type-I, feature vectors are fed to the LDA one by one and corresponding costs are measured. By comparing the cost values over all feature vectors the best feature is selected.

After the first feature is selected now we use the feedback path from the cost function to the dictionary as in Figure 1. The index of the selected feature corresponds to a node on the double tree which has a frequency tree index and a time tree index in that subband. In the frequency tree the nodes (subbands) which overlap the selected frequency index are removed. Similarly in the time tree the nodes which overlap the selected time index are removed. This way only "good" potential feature vectors are kept in the dictionary. Hence, the dictionary is pruned based on the last selected feature. Now the next feature which will do best in combination with the first selected one is searched on the pruned dictionary. This procedure is run until the desired number of features is reached. Therefore, the only difference between type II and type I is that pruning is done on the dictionary based on the selected features.

3.3 Type-III

This type is the simplest one. It does not use the LDA nor a feedback path when Figure 1 is considered. Instead, using cost function FD, a cost value is computed for each node on the double tree individually. Then a pruning algorithm is run on the double tree from bottom to top according to a rule to find the nodes with maximum discrimination power measured by the FD cost function.

As stated before, the selected features using one of the types described so far are used to classify a nut sound signal in the validation phase. Selected features are calculated on the sound signal and a binary decision is given.

4. CLASSIFICATION OF FOOD KERNEL IMPACT ACOUSTICS

In order to evaluate the performance of the adaptive subset selection approaches, we test them on two classification problems based on impact acoustic signals for nut kernel inspection and separation. The first one is open-closed pistachio separation. The second one is empty-full hazelnut classification. The quality of nut shells is a critical issue which may cause low consumer acceptance in the nut industry. For instance, closed shell pistachio nuts could be rejected by consumers because they are difficult to open and may contain immature kernels. Currently closed shell pistachio nuts are separated from open shell nuts by mechanical devices. These devices, called "pinpickers," can inadvertently damage the kernel of open shell nuts by inserting a needle into the kernel meat. In addition, according to [6], approximately 5 to 10%

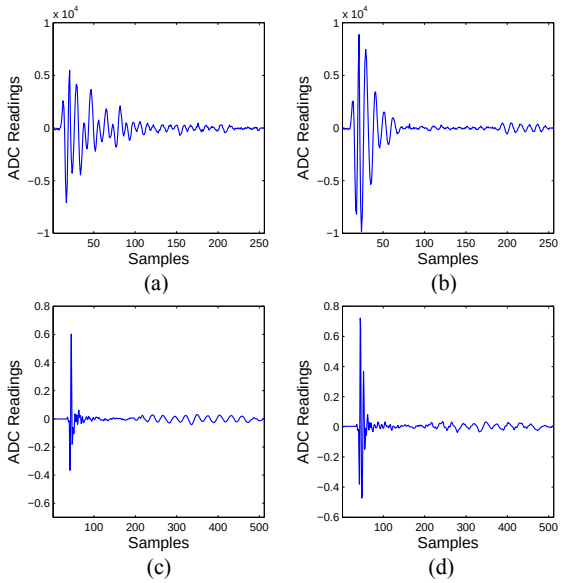


Figure 3- Sample impact acoustic signals. (a) Open and (b) Closed pistachio acoustic. (c) Empty and (d) Full hazelnuts acoustic.

of all U.S. open shell pistachio nuts are incorrectly classified by mechanical devices as having a closed shell, costing the industry $3.75 to $7.5 million per a year in lost revenue. Therefore high accuracy classification systems are needed in the industry.

Hazelnuts, another widely used nut in the food market, are one of the main ingredients used in chocolate and flavored coffee industries. One of the major quality attributes of raw bulk hazelnuts is the ratio of kernel weight to shell weight. Empty hazelnuts and hazelnuts containing undeveloped kernels negatively affect this ratio. If the ratio of kernel weight to gross weight is less than 0.5 then some buyers reject the product. Currently, raw hazelnuts are processed by an "airleg" which is a pneumatic device to separate empty hazelnuts from fully developed ones. However, these devices have high classification error rates.

Recently, a new non-contact system based on impact acoustic emission has been proposed for food kernel inspection which overcomes several limitations of the approaches summarized above. Initially this system was used to separate pistachio nuts with closed shells from those with open shells which are collected from California region of USA. In this system the pistachio nuts are impacted onto a steel plate and the resulting acoustic signals are recorded. During the off-line training (learning) phase a total of 359 features are extracted. They are based on the observations from open and closed shell pistachio nut sound signals and use the absolute value of the signal magnitude, the absolute value of the gradient, or both. Among those 359 features the best 3 of them are selected using both linear and non-linear discriminant analyses. The selection of only the 3 best features is due to the real-time processing constraint. This constraint requires that the nut under examination be decided on before the next nut comes into the system. Using with these 3 selected features on the validation set the classification accuracy was approximately 97% with a throughput of 40 nuts per second, proving better performance than traditional mechanical devices.

This same impact acoustics system was applied to empty-full hazelnut inspection and successful results were obtained in [7]. The authors used different signal features for better classification accuracies. The features are extracted from the time domain and frequency domain individually. Combinations of several subsets of these features are trained and tested using a support vector machine classifier. An improved classification accuracy of 97% was obtained. Both pistachio and hazelnuts studies using impact acoustics emphasized the importance of signal processing stages. In particular, it has been shown that having a priori information about relevant time segment and frequency indexes has important effects on classification accuracy. However the adjustment of these parameters is demanding. Furthermore the differences between nuts from region to region make it necessary to develop an adaptive classification system which can adjust its parameters for the given signal (nut type).
After this point we tackle these problems by using the adaptive time-frequency plane feature subset selection and classification algorithm that we described above. In particular, we test our proposed system to discriminate between open and closed shell pistachio nuts and full and empty hazelnuts collected from the Gaziantep and Blacksea region of Turkey, respectively. The recordings used in this study were provided by Bilkent University, Turkey.

The pistachio and hazelnut impact acoustics are recorded with a highly directional microphone. Output of the microphone is digitized with 44kHz sampling frequency by a sound card attached to a PC and stored for further analysis. For each type of pistachio, 200 recordings were obtained. Each recording was 256 samples long. For hazelnuts 230 recordings were obtained for each class and each signal was 512 samples long. Sample signals recorded with the described setup are given in Figure 3. The sound signals available this way are analyzed in an off-line manner for feature extraction. After features are selected and the decision rules are set the system is run for validation. In real time implementation a decision is given and either the nut is diverted by an air valve to one stream or no action is taken and the nut goes to another stream. Obviously, a timing constraint is present and this is also one of the reasons it is necessary to give a good decision as fast as possible. Although we do not validate the performance of our data in real-time these time constraints have to be taken into consideration for practice.

5. RESULTS

We tested the proposed approach on the pistachio and hazelnut acoustic signals. We used a 2 times 2 fold cross validation method to estimate the classification performance. Basically half of the data set is used for training and the rest for testing. Then the test and train sets are swapped. This experiment is repeated 2 times. We use a frequency level F=3 and time level T=4 for the dual tree. For UDWT, Daubechies wavelet filter of order 6 is used. After calculating the energy features in the nodes of the dual tree, they are converted to

TABLE-1: Open-Closed shell pistachio nut classification errors for the proposed types. The results from base approaches which are used in this area is given as well.

Type	Minimum Error	# of Features
Type I	8.25%	17
Type II	**8.25%**	**8**
Type III	11.5%	31
BA	18%	3

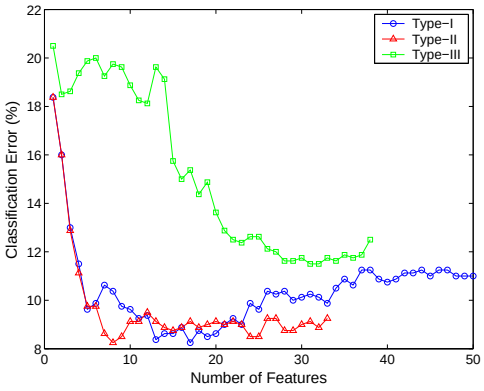
TABLE-2: Empty-Full hazelnut classification errors for the proposed types. The results from base approaches which are used in this area is given as well.

Type	Minimum Error	# of Features
Type I	1.52%	47
Type II	**1.74%**	**9**
Type II	1.96%	49
BA	3%	21

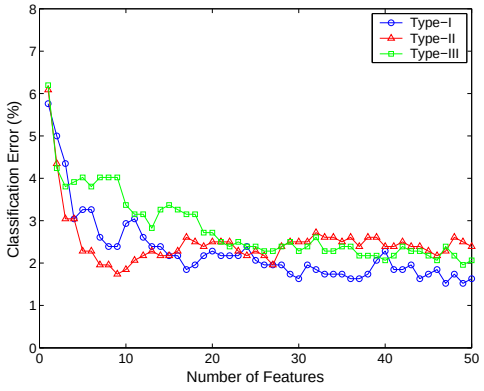
log scale. Table 1 and Table 2 show the classification accuracies obtained with the proposed approaches for pistachio and hazelnuts respectively. The classification errors obtained with those base line algorithms proposed in the same area are given as BA. The Type-I and Type-II approaches which use the feedback from the classifier has outperformed the type-III and BA approaches. As indicated before, the type-III only uses the same tree pruning method of original LDB algorithm and does not account for the interactions between features. The obtained classification accuracies strongly support that the evaluation of the classification performance of the combined features in the training stage is important and should be preferred to evaluating the individual discrimination power of the features. Furthermore we note that Type-II approach used always small number of features then the other approaches to achieve the minimal error. The classification error curves versus the number of features are given in Figure 4. This can be explained by the effect of pruning. Since each selected feature index is used to prune the dual tree structure the orthogonalization of the representation space helps the classifier to use small number of features to achieve minimal error rate. Although the obtained classification accuracies of type-I and type-II are comparable, the number of used features carries significant importance for real-time applications. It reduces computational complexity. Many times, PCA is used to decorrelate the time-frequency features which induce an additional complexity [2, 3]. Here the type-II could perform a similar dimension reduction without need any post processing stage. In addition it is a critical property to achieve higher generalization capability. In practice type-II may generalize more than type-I and type-III.

6. CONCLUSION

In this paper we described a new adaptive time-frequency plane feature extraction and classification algorithm. We tested the new approach on impact acoustic signals for food inspection. The results we obtained show that the algorithm is superior to previous algorithms applied in this area. It uses a small number of features to achieve minimal classification error by orthogonalization of redundant dictionary with *classifier directed pruning*. The use of small number of features for classification makes the algorithm computationally efficient. Based on the observations from previously mentioned algorithms above, our proposed algorithm does not depend on a priori knowledge of time-frequency content of the signals under examination. Its adaptation capability to different type of signals makes the algorithm a universal method.



(a)



(b)

Figure 4- Classification curves for (a) Pistachio (b) Hazelnut classification.

7. REFERENCES

[1] N. Saito, R. R. Coifman, F. B. Geshwind, F. Warner, "Discriminant feature extraction using empirical probability density and a local basis library," *Pattern Recognition*, pp. 1842-52, Vol. 35, 2002.

[2] N. F. Ince, S. Arica, A. H. Tewfik, "Classification of single trial motor imagery EEG recordings with subject adapted non-dyadic arbitrary time-frequency tilings," *J. of Neural Engineering*, pp. 235–244, July 2006.

[3] K. Englehart, B. Hudgins, P. A. Parker, M. Stevenson, "Classification of the myoelectric signal using time-frequency based representations," *Med. Eng. Phys.*, pp. 431-438, Vol.21, 1999.

[4] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Trans. Image Proc.*, pp. 1549–60, Vol.4(11), Nov.1995.

[5] M. Vetterli, "Wavelets, approximation, and compression," *IEEE Signal Proc. Magazine*, pp. 59–73, Sept. 2001.

[6]T. C. Pearson, "Detection of pistachio nuts with closedshells using impact acoustics, " *Applied Eng. in Agric.* 17(2): 249-253. 2001.

[7] I.Onaran, T.C. Pearson, Y. Yardimci, A. E. Cetin, "Detection of underdeveloped hazelnuts from fully developed nuts by impact acoustics," *Trans. of ASABE*, Vol.49(6), 2006.