Discriminant Feature Selection for Texture Classification

Abhir H. Bhalerao and Nasir M. Rajpoot
Department of Computer Science, University of Warwick, UK
{abhir|nasir}@dcs.warwick.ac.uk

Abstract

The computational complexity of a texture classification algorithm is limited by the dimensionality of the feature space. Although finding the optimal feature subset is a NP-hard problem [1], a feature selection algorithm that can reduce the dimensionality of problem is often desirable. In this paper, we report work on a feature selection algorithm for texture classification using two subband filtering methods: a full wavelet packet decomposition and a Gabor type decomposition. The value of a cost function associated with a subband (feature) is used as a measure of relevance of that subband for classification purposes. This leads to a fast feature selection algorithm which ranks the features according to their measure of relevance. Experiments on a range of test images and both filtering methods provide results that are promising.

1 Introduction

In many real-world classification applications, a number of features are often computed to aid the supervised learning. Some of the candidate features, however, may be *irrelevant*, in that they may be adding noise to useful information, or *redundant* in the presence of other relevant features. A feature selection algorithm addresses this problem by selecting a subset of features which are directly relevant to the target concept, which is texture classification here. Eigenspace methods like the principal component analysis (PCA) can be used to transform the feature space to a set of independent and orthogonal axes and rank those axes by the extent of variation. This is one way to reduce the set of features with regard to the global data covariance. Fisher's linear discriminant analysis (LDA), on the other hand, finds the feature space mapping which maximises the ratio of between-class to within-class variation jointly for each feature (dimension). PCA can be further applied to find a compact subspace to reduce the feature dimensionality. The central theme of almost all feature subset selection methods is to reduce the computational complexity and improve the classification performance by discarding those features which may be irrelevant or redundant (see, for instance, [2] for a survey of such algorithms).

Despite the advances in feature selection, texture classification algorithms have often used as many features as possible [3] without consideration of whether the features are independent and discriminating, giving rise to unnecessary computation and sometimes hindering the subsequent classification process. In this paper, we present a fast feature selection scheme that uses a cost function associated with each feature as a *measure of relevance* of that feature to texture classification. Features are ranked so that the problem

BMVC 2003 doi:10.5244/C.17.80

of selecting a subset of features reduces to picking the first m features only. Such features can also be termed as *discriminant* features, offering the benefit that only the discriminant feature subset needs to be computed on the observation data, unlike PCA or LDA methods. This has significant advantage when dealing with texture classification problem in 3D [4] because of reduced complexity. We note that Boz [1] has proposed a similar algorithm which uses Quinlan's gain ratio [5] to sort the features according to their relevance. It is also worth noting that our approach to the problem separates feature selection from classification stage, therefore putting it in the *filter* category of Kohavi et al. [2].

The paper is organised as follows. In the next section, a statement of the problem is provided with a description of the cost functions that can be used to measure the feature relevance. In Section 3, two multiresolution feature space mappings investigated in this work and based on subband decomposition are described. Experimental results are presented and discussed in Section 4, and the paper concludes with a summary and future directions.

2 Discriminant Feature Selection

Texture classification can be cast into the general signal classification problem of finding the best mapping, $d : \mathscr{X} \rightarrow \mathscr{Y}$, of a set of input signals, $\mathscr{X} \subseteq \mathbb{R}^n$, on to the class labels, $\mathscr{Y} = \{1,...,K\}$. Even given a set of training samples, in most interesting problems, the dimensionality of the input space, n, is too great to estimate the optimal classifier, such as the Bayes classifier. Feature extraction allows us to discard much of the redundant information. Multivariate data often contain important structures (features) at lower dimensions and that the problem has some *intrinsic dimension* $m < n$ is of course goal dependent [6]. It is therefore prudent to employ a classifier, $k : \mathscr{F} \rightarrow \mathscr{Y}$, on a subspace, $\mathscr{F}$ of $\mathscr{X}$. We can write the feature space mapping $s : \mathscr{X} \rightarrow \mathscr{F}$. Furthermore, we can try to extract a *discriminant* subspace of $\mathscr{F}$ using a feature selector Θ_m that picks a set of dimensions m which we would hope is the intrinsic dimension of the problem:

$$d = k \circ \Theta_m \circ s \tag{1}$$

The focus of our study here is to investigate two popular multiresolution feature space mappings for textural data: a special kind of wavelet packet transform and a Fourier transform, subband filtering akin to a Gabor decomposition, and to employ a "greedy" feature selector which operates sequentially on the feature dimensions (subbands).

Let us define the feature selection problem in the context of subband decomposition as follows. Given a test image $\mathbf{x}$ that has been decomposed into n subbands, each of which can be regarded as a feature, the goal is to select m subbands such that the resulting misclassification error is minimal. Here we show that it is still possible to gauge the discrimination power of a subband independent of the classifier conditioned on K representative texture samples, $\mathscr{T} = \{(\mathbf{x}_1,y_1),...,(\mathbf{x}_K,y_K)\}$. Given the nature of subband decompositions under consideration, a subband can be regarded as being highly discriminant if it highlights the frequency characteristics of one class but not the other. In other words, if the coefficients of a particular subband light up (ie, are higher in magnitude) for one class but are relatively insignificant for the other class, the subband can prove to be helpful in terms of classification performance.

Rajpoot [7] has previously proposed to use a symmetric version of Kullback-Leibler distance, between the normalised energies of a subband for training images, as a measure

of relevance or an estimate of the discrimination power of a subband. In this paper, we investigate the effectiveness of a list of cost functions described below for our feature selection algorithm using two feature space mappings and a range of test images. Let F and G denote the transform coefficients of a particular subband for training images $\mathbf{x}_1$ and $\mathbf{x}_2$, belonging to two different texture classes, respectively. We considered three ways of forming pseudo-distributions f from the subband coefficients F:

$$f_1(x,y) = F(x,y), \qquad f_2(x,y) = |F(x,y)|^2, \qquad f_3(x,y) = |F(x,y)|^2/\|\mathbf{x}_1\|^2 \tag{2}$$

Similarly, the pseudo-distributions g_i ($i = 1,2,3$) can be obtained by using the subband coefficients $G(x,y)$ and texture image $\mathbf{x}_2$. Four discriminant measures were used in our experiments: symmetric Kullback-Leibler (KL) divergence, Fisher distance (FD), Bhattacharyya distance (BD), and Euclidean distance (ED) – denoted respectively by KL_i, FD_i, BD_i, and ED_i – defined as follows.

$$KL_i(F,G) = I(f_i,g_i) + I(g_i,f_i), \tag{3}$$

$$FD_i(F,G) = \frac{|\mu(f_i) - \mu(g_i)|^2}{\sigma^2(f_i) + \sigma^2(g_i)}, \tag{4}$$

$$BD_i(F,G) = \frac{1}{4}\left[\log\left(\frac{(\sigma^2(f_i) + \sigma^2(g_i))^2}{4\sigma^2(f_i)\sigma^2(g_i)}\right) + FD_i(F,G)\right], \tag{5}$$

and

$$ED_i(F,G) = \|f_i - g_i\|_2 \tag{6}$$

where

$$I(f,g) = \sum_x \sum_y f(x,y) \log \frac{f(x,y)}{g(x,y)}$$

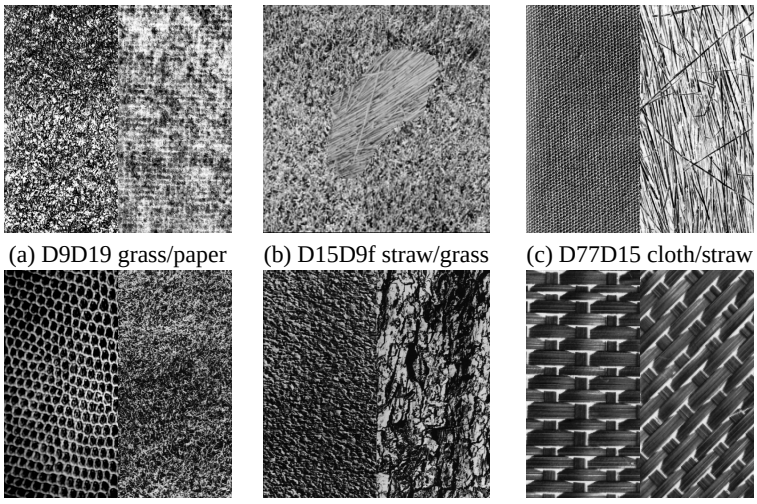
is the relative entropy between f and g, $\|\cdot\|_2$ denotes the l_2-norm, and $\mu(f)$ and $\sigma^2(f)$ respectively are the mean and variance of f. The feature selector, Θ_m, operates by selecting the m subbands with the largest discrimination power using one of the above four measures.

Saito et al. [6] warn that the approach of sequentially measuring the efficacy of each dimension of the feature space independently may be "too greedy" as 2D and higher dimensional structures in the feature space may be missed. The principal advantage, however, is that we can make a feature selection solely on training data which reduces the complexity of the final classifier on test samples. When compared with traditional multivariate feature projection methods like PCA or LDA, this advantage is significant.

3 Subband Decompositions

3.1 Full Wavelet Packet Decomposition

The wavelet packet (WP) decomposition [8] extends the standard dyadic discrete wavelet transform such that the high frequency wavelet subbands are decomposed as well as the low frequency subbands, to an arbitrary number of levels. The frequency intervals of varying bandwidths can be adaptively selected, e.g. by picking the most energetic subbands, to extract specific frequencies present in the given signal from this overcomplete



(a) D9D19 grass/paper (b) D15D9f straw/grass (c) D77D15 cloth/straw

(d) D9D3 grass/reptile (e) D4D12 cork/bark (f) D65D65r fence/fence

Figure 1: Test images created from Brodatz album examples show combinations of granular, oriented and periodic textures (natural and synthetic).

representation. The full wavelet packet decomposition (FWP), sometimes also known as the uniform wavelet subband decomposition, is a special type of wavelet packet transform wherein all subbands are decomposed up to a given number of levels.

The discrete wavelet packet transform of a 1D discrete signal $f(x)$, $x = 0,1,\ldots,N-1$ can be computed as follows. The wavelet packet coefficients are defined as

$$\begin{aligned} w_0^0(x) &= f(x) \qquad (7) \\ w_{2j}^l(x) &= \sum_k g(k-2x)\, w_j^{l-1}(k) \\ w_{2j+1}^l(x) &= \sum_k h(k-2x)\, w_j^{l-1}(k) \\ & x = 0,\ldots,N2^{-l}-1 \qquad (8) \end{aligned}$$

where $l = 1,2,\ldots,L$; $L = \log_2 N$, $w_j^l(x)$ is the transform coefficient corresponding to the wavelet packet function which has relative support size 2^l, frequency $j2^l$ and is located at $j2^l$. In other words, l, j and x can be regarded as the scale, frequency and position indices of the corresponding wavelet packet function respectively. The coefficients $\{h_x\}$ and $\{g_x\}$ correspond to the lowpass and highpass filters respectively for a two-channel filter bank and the transform is invertible if appropriate dual filters $\{\tilde{h}_x\}$, $\{\tilde{g}_x\}$ are used on the synthesis side.

The FWP can be used as a feature space mapping by generating a feature image from each of its subbands at a given level. A feature image corresponding to a subband can be generated by computing the magnitude of contribution of the coefficients of that subband to the input image, and smoothing it with a Guassian filter. As can be seen from Figure 2(b), which shows sixteen features for the grass/straw image computed from the subbands of a two-level FWP, some of the features may be adding merely noise to the performance of a texture classification algorithm.

3.2 Wilson-Spann Decomposition

The Wilson-Spann decomposition [9] uses a set of real, band-limited basis functions which cover the Fourier half-plane. The decomposition is multiresolution and resembles Gabor filtering and has been shown to perform well in representing image features exhibiting granularity and local orientation. In [9], the authors proposed using finite prolate spheroidal sequences (FPSS) because they are optimal in maximising energy in both spatial and spatial-frequency domains. For ease of implementation, we use instead truncated Gaussian functions of similar shape, centred in the frequency subband. The subband coefficients $s_j^l(\mathbf{u})$ for subband j at level (or scale) l of the decomposition of a 2D signal $f(\mathbf{x}), \mathbf{x} = (x,y)^T$, $x,y = 0,1,\ldots,N-1$, can be computed by applying a set of band-limiting operators Λ_j^l to the filtered discrete Fourier transform coefficients:

$$s_j^l(\mathbf{u}) = \Lambda_j^l \left[g(\mathbf{u}; \mathbf{m}_j^l, \Sigma_j^l) F(\mathbf{u}) \right] \tag{9}$$

where $g(\mathbf{u})$ is an isotropic Gaussian function with centre frequency $\mathbf{m}_j^l$ and bandwidth determined by Σ_j^l, and $F(\mathbf{u})$ is the DFT of $f(\mathbf{x})$. Various arrangements of these filters are possible: a centre-surround tessellation with a low-pass region in the interval $[\pi/2, \pi/2]$ and 6 bands of size $[\pi/4, \pi/4]$ creates a set of subbands that have been demonstrated to capture well textural variation in frequency and orientation. At the next level of the decomposition, the central LP region is itself tessellated by a low-LP region size $[\pi/4, \pi/4]$ bandwidth surrounding oriented filters size $[\pi/8, \pi/8]$, and so on (Figure 2(a)).

4 Experiments and Discussion

The discriminant feature selection algorithm described in Section 2 was tested on texture examples taken from the Brodatz texture catalogue. We created two-class images containing pairs of textures with granularity (D9=grass, cork=D4, bark=D12), principal orientation (D15=straw) and periodicity (D3=reptile, D103=burlap, D65=fence), as shown in Figure 1.

Subband decomposition of these images using both the FWP (2 levels using Duabechies-8 filters) and the Wilson-Spann scheme (for 3 and 5 levels) were made. Given a pair of training samples, a set of discriminant values were calculated using ED_3, KL_2, FD_2 and BD_1 for corresponding subband decomposition. These values were then sorted into decreasing order. We then used a *k*-means classifier to classify the two-class test image taking increasing numbers of subband features in the order defined by the discriminant value of each feature. The percentage misclassification error was then calculated given a binary mask for the true result. Figure 3 show error plots for the test images D15D9f (grass/straw), D9D3 (reptile/grass), and D65D65r (fence/fence) for both subband decompositions. These plots are representative of the results overall and we can begin to draw some conclusions on the effectiveness of particular cost functions and subband decomposition given a pair of textures.

For textures containing a combination of granularity and orientation, such as straw/grass, the plots show FD and BD perform better as a feature relevance measure than KL or ED, which performs least well (Figures 3(a) and 3(b)). BD is able to bring the misclassification error to about 5% for 4 bands. The error then reduces marginally with the addition of more features to the classifier. Images containing a natural periodic texture, like reptile

skin against a granular/random texture like grass (Figure 1(d)), show the mean-variance cost measures, FD and BD, to significantly outperform the point-wise measures, KL and ED. But again, a relatively small subset of features is able to achieve a good classification. Note that in the case of the FWP, Figure 3(c), the plots for BD and FD show poorer results beyond 10 features indicating that bands 11-16 do not have useful discriminating power and are working against the classifier.

Figure 3(f) reveals that FD does less well when a periodic texture is paired with another periodic texture (Figure 1(f) which has fence against rotated fence). In this case, both the point-wise measures, KL and ED, can identify the subbands where the spectral distributions vary (likely to be a periodic pattern in the Fourier domain), whereas, the average energy difference, which is what FD estimates, cannot reveal weather a band is discriminating. The first term of BD measure takes into account the variability which explains its better performance for these images. For the FWP (Figure 3(e)), the KL and ED continue to perform worse which can be explained by the transform coefficients having spatial coordinates. A point-wise, pseudo distribution measure like KL, is fooled by the energy voids in a periodic texture aligning with energy peaks from the other texture at the same coordinate. KL, however, is widely used in WP basis selection [10] because it is additive and can be used to compare the relative energy when choosing child subbands over parent for an image region. Figure 4 shows four most discriminant feature images and corresponding classification results for the straw/grass image. It can be seen from this figure that the features selected by our algorithm prove to be relevant to the task at hand, ie texture classification.

As regards the optimality or otherwise of the sequential ranking, we have experimented with reversing the discriminant ranking order and comparing with random feature selection orders. Figure 5 shows a plot of the misclassification curve for an example image with different feature selections. Note that the 'best' performance is obtained by the decreasing discriminant ranking of our feature selection, where as the 'worst' is given by reversing the best ranking order. The random feature selections fall somewhere in between. This indicates that the misclassification curves are approximately bounded by the two decreasing and increasing discriminant rank orders. It is not surprising to see all orders converge rapidly beyond $n/2$ features as the number of possible ways to select the remaining features decreases - there is $n!$ possible rank orders at the start and $(n/2)!$ half way along and so on. If it is to be believed that the random orders fall along the centroid of this distribution, then it would appear that the likelihood that our 'best' rank ordering could have been obtained simply by chance is small. The decreasing rank order is probably close therefore to the optimal feature selection order.

5 Conclusions

The primary goal of a feature selection algorithm in supervised classification applications is to reduce the dimensionality of problem by selecting only relevant features and discarding features that are irrelevant or redundant. In this paper, we described a simple yet effective feature selection algorithm for supervised texture classification problems by associating a measure of relevance with each of the features. Although the algorithm can be used with any type of features, its effectiveness was demonstrated by applying it to features extracted from two subband decompositions: a full wavelet packet decom-

position and a Gabor type representation. From our experiments, we can conclude that measures based on BD (or FD) seem to be the best for feature selection from FWP subbands. For Fourier domain subbands, like Wilson-Spann, BD is best but, KL can work marginally better when energy distribution is periodic. Experimental results also show that the ranking of features according to a sorted measure of relevance produces plausible results even using a simple *k*-means classifier. Investigation into a more sophisticated classifier and the extension of measures to a multi-class problems are matters of further interest. Combining a discriminant feature selection method with a basis selection and compact representation also remains to be investigated.

References

[1] O. Boz. Feature Subset Selection by Using Sorted Feature Relevance. In *Proc. Intl. Conf. on Machine Learning and Applications*, June 2002.

[2] R. Kohavi and G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[3] T. Randen and J.H. Husoy. Filtering for Texture Classification: A Comparative Study. *IEEE Trans. on PAMI*, 21(4), April 1999.

[4] C. C. Reyes-Aldasoro and A. H. Bhalerao. Volumetric Texture Classification and Discriminant Feature Selection for MRI. In *Proc. Information Processing In Medical Imaging IPMI'03*, June 2003.

[5] J.R. Quinlan. Learning Decision Tree Classifiers. *ACM Computing Surveys*, 28:71–2, 1996. (invited contribution to 50th anniversary issue).

[6] N. Saito, R. Coifman, F. B. Geshwind, and F. Warner. Discriminant Feature Extraction Using Empirical Probability Density Estimation and a Local Basis Library. *Pattern Recognition*, 35:2841–2852, 2002.

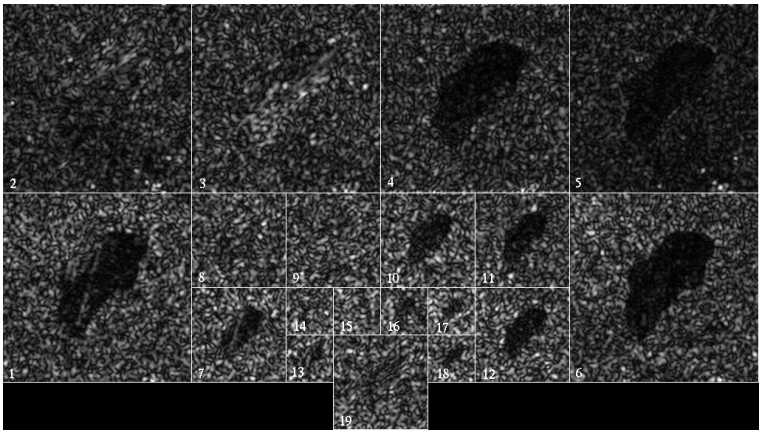
[7] N.M. Rajpoot. Texture Classification Using Discriminant Wavelet Packet Subbands. In *Proc. IEEE Midwest Syposium on Circuits and Systems*, Aug. 2002.

[8] R.R. Coifman and Y. Meyer. Othornormal Wave Packet Bases. *preprint, Deptt. of Maths., Yale Uni., USA*, 1990.

[9] R. Wilson and M. Spann. Finite Prolate Spheroidal Sequences and Their Applications: Image Feature Description and Segmentation. *IEEE Trans. PAMI*, 10(2):193–203, 1988.

[10] N. Saito and R.R. Coifman. Local discriminant bases. In A.F. Laine and M.A. Unser, editors, *Mathematical Imaging: Wavelet Applications in Signal and Image Processing II*, volume 2303, 1994.

(a)



(b)

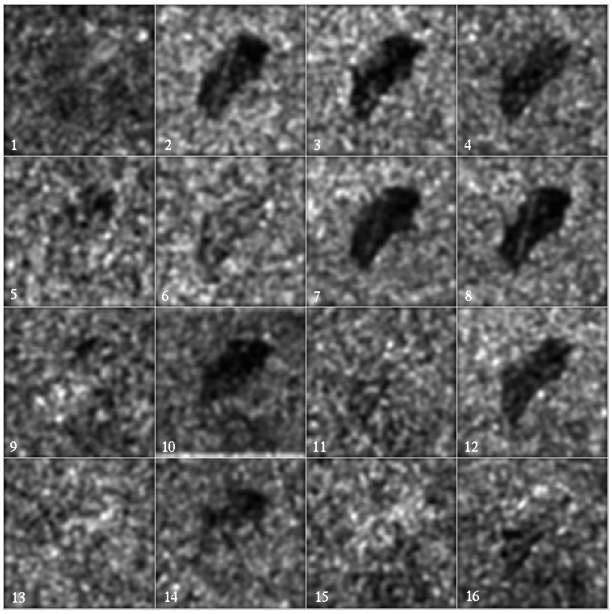


Figure 2: Wilson-spann and FWP subband decompositions of example texture image D15D9f (straw/grass) in Figure 1(b). (a) Three levels of the Wilson-Spann subband decomposition using a centre-surround arrangement of subbands across the Fourier domain. (b) Two levels of a Full Wavelet Packet decomposition. Note how in both cases certain subbands clearly contrast the two textures.

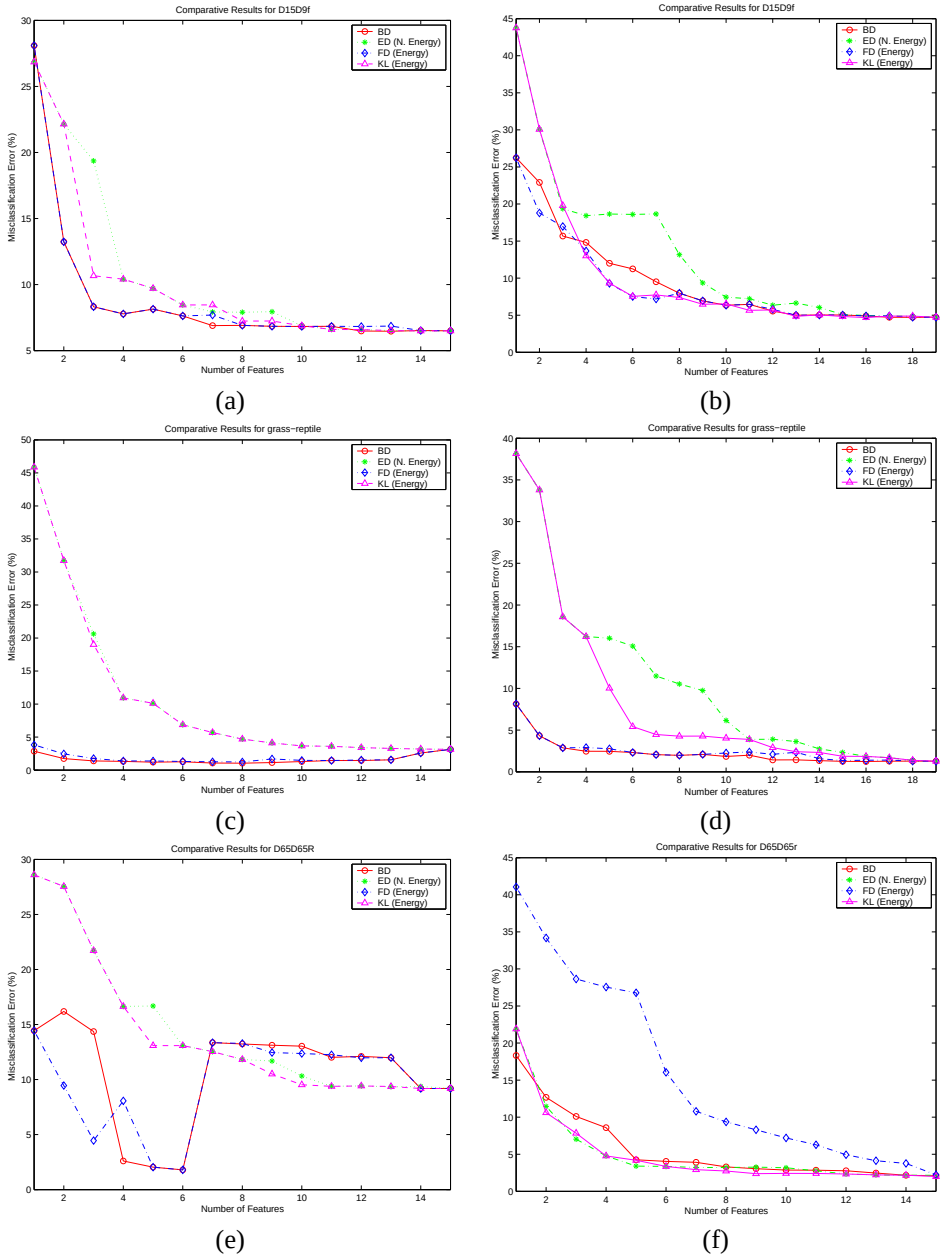
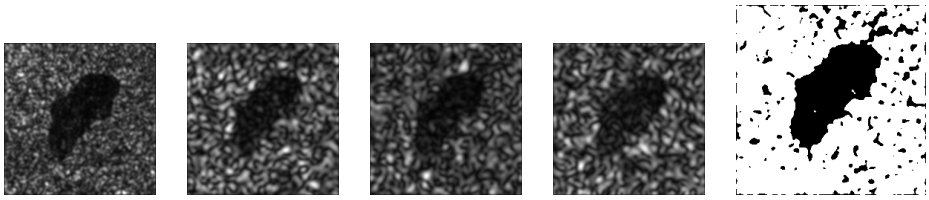
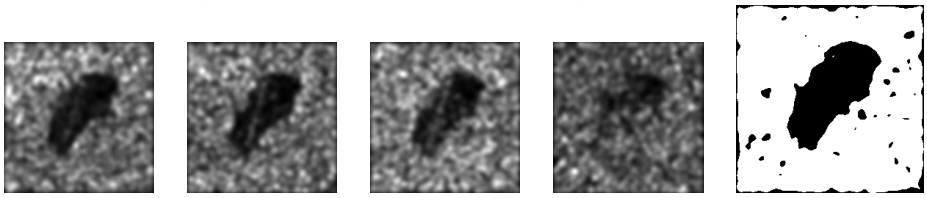


Figure 3: Classification results for FWP and Wilson-Spann subbands.



(a) Feature 6 (b) Feature 12 (c) Feature 11 (d) Feature 10 (e) Classified



(f) Feature 7 (g) Feature 8 (h) Feature 2 (i) Feature 14 (j) Classified

Figure 4: (a)-(d) The first 4 most discriminant subband feature images from 3rd order Wilson-Spann decomposition of grass/straw image D15D9f. (e) k-means classification result using first 4 features only. (f)-(i) The first 4 most discriminant subband feature images from 2 levels of FWP on grass/straw image D15D9f. (j) k-means classification result using first 4 features only.

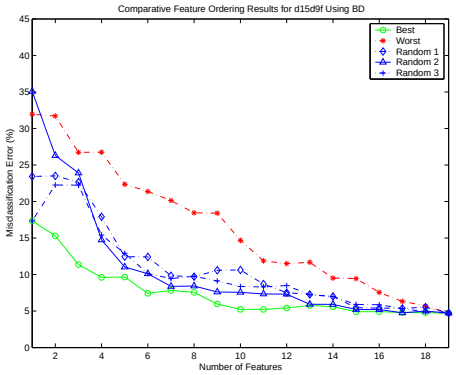


Figure 5: Comparing different discriminant rank orderings to assess the optimality of the decreasing rank order ('Best') against increasing rank order ('Worst') and three random orderings. The best and worst orderings are the boundaries the distribution of all $n!$ rank orderings and the random orderings appear to cluster near the centroid of this distribution.