An Adaptive Approach to Feature Selection as Applied to a Texture Classification Problem

H. C. Shen and R. Pilkey
Department of Systems Design Engineering
Pattern Analysis and Machine Intelligence Group
University of Waterloo
Waterloo, Ont.
Canada N2L 3G1
(Tel. 519 888-4647)
e-mail: shen@watdcs.bitnet

Abstract

In most pattern recognition problems, usually a relatively large number of features are extracted. Feature selection methods, as a practical consideration, attempt to reduce the number of features by selecting a subset representing the best features. This paper begins by outlining a ranking scheme for features based on a feature's calculated "performance potential". The performance potential is made up of a number of performance measures: extraction time, memory requirements, variance, covariance and classification success. An adaptive scheme is proposed to process a large number of initial features and arrive at the "best" subset based on their performance potential. The approach is successfully applied to a texture analysis problem. The results of the testing of the approach are presented with conclusions indicating its effectiveness.

Keywords: Feature Selection; Pattern Recognition; Adaptive Approach; Performance Potential.

1 Introduction

The goal of the pattern recognition problem is to classify an input pattern among a number of potential pattern classes. The process can be developed in three stages: feature extraction, feature selection and the final classification.

One of the area which has been identified as being critical, to the development of a successful classification, is the problem of determining an effective feature set. In our experience [12], extensive experimentation leads to the final feature set. Perhaps of more significance is the amount of interpretation of the experimental results required to arrive at an effective feature set. What is proposed is an approach which would automate the feature selection and effectively select the "n best" features. This subset should give the "best" result in classifying an input pattern to a pattern class. The fact that this subset should be more economical to maintain than the full feature set.

In this paper, we shall experiment texture classification to demonstrate the efficacy of our approach to feature selection. Texture classification problems involve a large number of potential features and a number of different pattern classes. The primary objective is to select a subset of features which shows the best performance.

2 Adaptive Feature Selection

There are various forms of feature selection [3, 4, 5, 9, 11, 14]. A major limitation of these approaches is the computational cost associated with them. This paper proposes a new approach to feature selection. The basis for the approach is the development of a performance potential which provides a measure of the relative capabilities of each feature.

A straightforward application of the performance potential is to use it to rank the full feature set. The ranking allows the selection of the "best n" features. This development is what we term the general formulation. However, it suffers similiar limitations as the other approaches in terms of computational cost. An adaptive approach is proposed which proceeds sequentially through the pattern samples of a learning set. The adaptive and iterative approach provides potential improvements in computational complexity over the general formulation.

2.1 Performance Potential

In this paper, we propose five performance measure for each feature to calculate the performance potential. (The paper looks at five specific performance measures but the methodology can incorporate any number of desired performance measures.) The five performance measures are:

1. extraction time (**ext**)
2. memory requirements (**mem**)
3. relative variance (**var**)
4. covariance (**cov**)
5. classification success (**cls**)

Each performance measure is evaluated over "K" pattern classes with a learning set of "L" sample patterns per pattern class.

Both extraction time (**ext**) and memory requirements (**mem**) can be calculated for each feature with a representative sample pattern from within the learning set. It is obvious that it is desirable to select features which exhibit short extraction times and low memory requirements.

In selecting a feature set you would like the feature set to exhibit a tight dynamic range within each pattern class. This would be reflected in a small relative variance of a feature across the learning set patterns for each class. The relative variance performance measure (**var**) can be derived from the covariance matrix. Small covariance between features would indicate lack of dependence. Thus, minimizing covariance meets the requirement of reducing the redundancy within the feature set. The covariance measure (**cov**) is the average covariance observed between the i^{th} feature and the other features of the feature set, across the "K" classes. The measure of classification success (**cls**) is calculated for each individual feature. In our experiments, we have used an intra-class distance metric to provide a classification performance measure.

Based on these performance measures, we propose a function which indicates the performance potential of a feature. A good feature should provide high classification success (**cls**), small relative variance (**var**), small covariance (**cov**), short extraction time (**ext**) and low memory requirements (**mem**). Therefore, we define the performance potential for a feature as

$$P = w_{cls} * \mathbf{cls} + \frac{w_{var}}{\mathbf{var}} + \frac{w_{cov}}{\mathbf{cov}} + \frac{w_{ext}}{\mathbf{ext}} + \frac{w_{mem}}{\mathbf{mem}} \quad (1)$$

where the w's correspond to weighting factors for the respective performance measures.

2.2 Iterative and Adaptive Approach

The proposed adaptive approach is iterative in nature. It is outlined in the following algorithm:

STEP 1 Initialization

1.1 Select the size of the selected feature set, "n".

1.2 Select the number of sample patterns (per class), to be evaluated for each iteration – "S".

1.3 Select the number of iterations "M" in each monitoring stage.

1.4 Calculate the performance potential, P, for all "N" features over "$S * K$" sample patterns. Rank the features, such that

$$P_i \geq P_r \quad \text{where} \quad i = 1, \ldots, n \qquad r = n+1, \ldots, N$$

1.5 Form the initial selected feature set, $\{F_i\}$, and the unselected feature set, $\{F_r\}$.

STEP 2 For each iteration, t, and S_t sample patterns

2.1 Calculate the performance potential of the "n" selected features, over S_t sample patterns, denoted by P_{it} Eq.(1).

2.2 Calculate the average performance potential, over "n" features, denoted by

$$\overline{P}_t = \frac{1}{n} \sum_{i=1}^{n} P_{it}$$

STEP 3 For every "M" iterations

3.1 Compile the average performance potential history,

$$\{\overline{P}_t, \overline{P}_{t+1}, \ldots, \overline{P}_{t+M-1}\}$$

3.2 Determine the slope of the average performance potential history, D.

3.5 IF D < 0 then

3.5.1 For each feature, i, determine the slope – d_i, of the feature's performance potential history,

$$\{P_{it}, P_{i(t+1)}, \ldots, P_{i(t+M-1)}\}$$

3.5.2 Calculate the trend measure tm_i, as

$$tm_i = P_{i(t+M-1)} + d_i$$

3.5.3 Rank tm_i and remove feature F_x from $\{F_i\}$ and concatenate to the end of $\{F_r\}$, such that

$$tm_x = \text{ min. } \{tm_i \mid \text{i} = 1, \ldots, \text{n}\}$$

3.5.4 Select a new feature to $\{F_i\}$ from the front of $\{F_r\}$

STEP 4 IF the selected feature set remains stable for a desired number of monitoring stages then

- The final "n" features, $\{F_i\}$, represent the "best" features as determined by the process.
- otherwise, return to STEP 2.

3 Experiements

Numerous texture analysis methods have been proposed in the past [2, 6, 7, 8, 10, 13]. Each of these methods extracts a number of texture measurements for classification. We have chosen the co-occurrence matrix method in our research. A total of 150 features from each texture sample are extracted. Experiments were performed to select the "best 4" features for the texture classification. A set of 8 texture images (Figure 1) was taken from the photographic album of Brodatz [1]. A total of 210 sample patterns of size 64x64 pixels were formed from each image.

The texture problem outlined provides us with an opportunity to use the adpative process. To evaluate the effectiveness of the adpative approach we will compare it the general formulation. The results are presented in the following sections. To begin, we shall briefly describe the co-occurrence matrix method.

3.1 Texture Feature Extraction

For an indepth description of the co-occurrence matrix refer to Haralick et al. [6]. The method first constructs a gray-level co-occurrence matrix. From the co-occurrence matrix a number of texture features can be extracted.

The co-occurrence method is based on the estimated second order joint conditional probability density functions P(i,j | d, θ). Each P(i,j | d, θ) denotes the probability of occurrence of a pair of grey levels (i,j) at distance d and angle θ. For our work, we consider 4 different directions: $0^0, 135^0, 45^0 and 90^0$. We also combined the directional components to extract the absolute range and the average. Five different displacement magnitudes were used, 1, 2, 3, 4 and 8 pixels. This gives a total of 30 different forms of the co-occurrence matrices. From each co-occurrence matrix, 5 standard feature measures are extracted for each pair (d,θ).

Energy (Angular Second Moment)

$$E_{d,\theta} = \sum_{i=0}^{g-1} \sum_{j=0}^{g-1} [P(i,j \mid d,\theta)]^2 \qquad (2)$$

Contrast (Inertia)

$$C_{d,\theta} = \sum_{i=0}^{g-1} \sum_{j=0}^{g-1} (i-j)^2 P(i,j \mid d,\theta) \qquad (3)$$

Correlation

$$R_{d,\theta} = \sum_{i=0}^{g-1} \sum_{j=0}^{g-1} \frac{(i-\mu_x)(j-\mu_y)P(i,j \mid d,\theta)}{\sigma_x \sigma_y} \qquad (4)$$

where

μ_x, σ_x are the mean and the standard deviation of the row sums of matrix P(i,j | d, θ),

μ_y σ_y are the mean and the standard deviation of the column sums of matrix P(i,j | d, θ)

Entropy

$$H_{d,\theta} = -\sum_{i=0}^{g-1} \sum_{j=0}^{g-1} P(i,j \mid d,\theta) \lg P(i,j \mid d,\theta) \qquad (5)$$

Local Homogeneity (Inverse Difference Moment)

$$L_{d,\theta} = \sum_{i=0}^{g-1} \sum_{j=0}^{g-1} \frac{P(i,j \mid d,\theta)}{1+(i-j)^2} \qquad (6)$$

Thus, we have a total of 150 feature values extracted for each texture sample.

3.2 Performance Measures

One of the considerations in implementing the performance potential is the relative weightings assigned to the different performance measures. For our experiments, we have assigned heavy weighting (value of 1) to relative variance, covariance and classification success and a low weighting of 0.1 to extraction time and memory requirements. The intent of the low weighting is to reduce the influence on the performance potential but still retain an indication of the effect.

Both the extraction time and memory requirements can be approximated, for the various features, before running the experiments. The remaining performance measures are calculated during the execution of the process. Both the relative variance and covariance are derived from the average covariance matrix. The last performance measure – the classification success, requires a calculation which is dependent on the classifier selected. For our work, we used a minimum intra-class distance metric. For the two class case, the decision rule employed is a ratio of the distance metric calculated for each pattern class. The decision rule is evaluated for each feature across the sample patterns under consideration. An exponential function measures the classification success of each feature. The final classification measure, for each feature, is the sum of the exponential function for all the sample patterns. Extending the calculation to the "K" pattern classes, involves taking the average of pairwise applications of the classifier over the "K" classes.

3.3 Results

In our experiments a total of 8 texture classes were selected. The different texture images are taken from a photographic album by Brodatz [1]. The eight classes are: reptile skin (image D3), pressed cork (image D4), woven aluminum wire (image D6), herringbone weave (image D16), French canvas (image D20), French canvas (image D21), reptile skin (image D22) and wood grain (image D68). The various textures are illustrated in Figure 1. Each texture image is divided into 210 sample patterns or subimages, consisting of 64x64 pixels. The system used to run the adaptive process is a 10 MHz. IBM PC AT, without a math coprocessor. The program was developed with Turbo Pascal V5.0.

The initial step of the process is to select the various parameters which will control the adaptive process. To begin the adaptive process we require the size of the selected feature set "n", the number of sample patterns per iteration "S" and the number of iterataions for each monitoring stage "M". In our experiments, we have chosen n to be 4; S to be 1 (with 8 classes, we have 8 samples per iteration) and M to be 3 [1]. The final stage of the initialization is to calculate the performance potential for the full feature set over a set of sample patterns. We have chosen 5 samples from each class to ensure good stability when introducing the features.

A further consideration is to specify a stopping rule for the process. In the experiments, we stop the process after the composition of the "best" n features has remained constant for $2*N$ monitoring stages. The simple rule of $2 * N$ is an attempt to allow the selected feature set to "view" the full feature set at least two times before the feature set is considered stable.

With the 30 different co-occurrence matrices and the corresponding 5 feature measures, we generated 150 different features for each texture sample. Rather than applying the algorithm directly to the entire 150 feature set, we orga-

[1] These values can be determined by experienced users

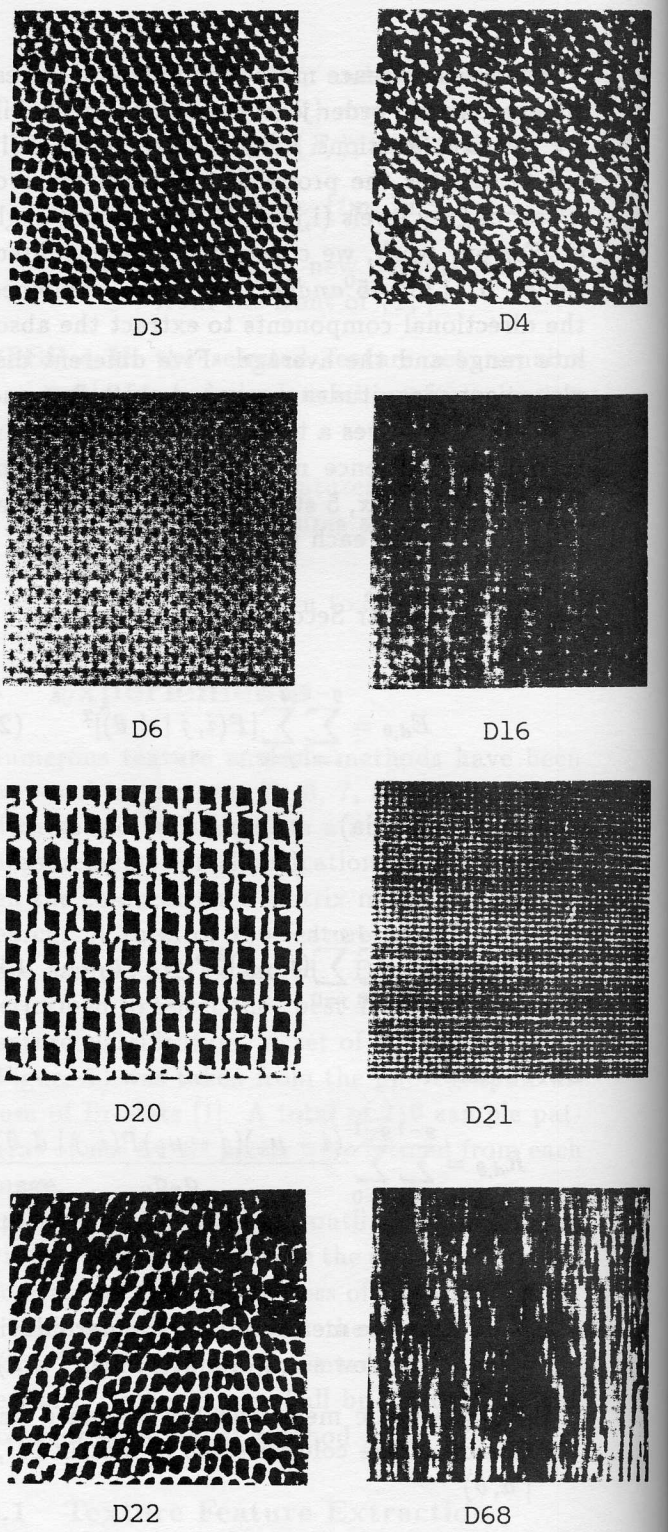

Figure 1: Eight Different Textures

nized the features into 5 sets, corresponding to the 5 feature measures extracted from each co-ocurrence matrix. Therefore, each set has 30 feature values. The algorithm is applied to each set separately. For each set, four "best" features will be selected and the results are combined to form a set of 20 features. The algorithm is applied to this final set to produce the four "best" features of the entire 150 features. There are two main reasons in doing this.

First of all, parallel computation can be applied on the 5 sets. This is a saving of execution time. In our experiments, the set on the energy feature measure took the longest (576 minutes on the IBM PC/AT). Comparing to the straightforward application of the performance potential to the full set of features which took 1285 minutes, the adaptive approach is much superior. Secondly, it is also reasonable to determine the "best" distance (d) and direction (θ) pair for each of the five feature measures in the co-ocurrence matrix. The final selection will indicate which of the feature measures paired with the distance and direction are the "best" features.

Table 1 presents the summary of the execution times of the adaptive approach and the straightforward ranking of the performance potential of the full feature set. Note that the experiments were done on an IBM PC/AT system which accounts for the long execution times. However, a saving in the execution time is reflected.

From our results, two feature measure sets (contrast and correlation) showed selected features different from the other three. For the energy, entropy and local homogeneity feature measure sets, the features extracted from the co-ocurrence matrices of 4 distances at the average direction are selected. This compares well with the work of Haralick et al. [6]. The results for the contrast and correlation feature measure sets can be explained by examining the actual function of the features.

The contrast feature is a measure of the local variation within the image. Few of the textures considered show distinct local regions. The contrast that does exist within the texture patterns is relatively fine grain. It would be expected that a small pixel displacement magnitude should provide the best discrimination of the texture patterns. We see this clearly in the results, with the selection of the displacement magnitude of 1 pixel across all the directional components.

Feature Set	Exec. Time (mins)
(in parallel)	
Energy	576
Contrast	324
Correlation	239
Entropy	218
Homogeneity	168
Final	279
Total	855
Full	1285

Table 1: Summary of the Execution Times

The other co-occurrence feature – correlation, is a measure of the gray level linear dependence. It would be expected that if the images have a lot of linear structure then the directional components could show better performance than their average. The selected feature set provided by the correlation feature includes two features capturing the range over the directional components and directional components from an angle of 0^0 and 135^0. The range over the directional components was shown by Haralick [6] to provide good classification success. The remaining selected features are more difficult to explain. If we examine the performance potential for each feature, under the correlation measure, we find a small range for the performance potential. This suggests that the corresponding ranking does not signify a significant improvement in performance for the selected feature set. Returning to the images, it is also difficult to discriminate the various texture patterns solely on linear structure. This may in turn lead to the difficulty in providing a significant selected feature set.

4 Conclusion and Future Work

In this paper, we have proposed an adaptive approach to feature selection. The efficacy of this approach has been demonstrated by the experiments performed in a texture classification problem. We admit that the computer system used did not show fantastic gain in the execution time of the adaptive approach over the general formulation. However, we believe that the proposed methodology has great potential in practical applications. First of all, researchers can be encouraged to explore new feature measures and combine different feature extraction methods to obtain as much information as possible. Utilizing the proposed approach, they will not be inhibited by the large number of features to be considered in their respective problems. Secondly, application dependent potential measures can be used to redefine the performance potential (Equation 1). For example, if memory requirement is not a constraint, then it can be deleted from Equation (1). Finally, parallel computation can be incorporated in this approach. Analysis and development of parallelizing the algorithm applied to image analysis is currently being considered.

References

[1] Brodatz, P., *Textures.* New York: Dover, 1966.

[2] Conners, R.W. and Harlow, C.A., "A Theoretical Comparison of Texture Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 3, May 1980, pp.204-222.

[3] De Jong, K., "Adaptive System Design: A Genetic Approach," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-10, no. 9, September 1980, pp.566-574.

[4] Duda, R.O. and Hart, P.E., *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[5] Fu, K.S., "Statistical Pattern Recognition," in *Adaptive, Learning and Pattern Recognition Systems,* J.M. Mendel and K.S. Fu, Eds. New York: Academic Press, 1970.

[6] Haralick, R.M., Shanmugam, K. and Dinstein, I., "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, Nov. 1973, pp.610-621.

[7] Haralick, R.M., "Statistical and Structure Approaches to Texture," *Proceedings of IEEE*, vol. 67, no. 5, May 1979, pp.786-804.

[8] Laws, K.I., "Texture Energy Measures," *Proceeds of Image Understanding Workshop*, Nov. 1979, pp.47-51.

[9] McMurtry G.J., "Adaptive Optimization Procedures," in *Adaptive, Learning and Pattern Recognition Systems,* J.M. Mendel and K.S. Fu, Eds. New York: Academic Press, 1970.

[10] Pietikainen, M., Rosenfeld, A. and Davis, L.S., "Experiments with Texture Classification using Averges of Local Pattern Matches," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 3, May/June 1983, pp.421-426.

[11] Samuel, A., "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Develop.,* vol.3, 1959, pp.221-229.

[12] Shen, H.C. and Pilkey, R.M., "Classification of Impedance Traces," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-17, no. 4, July/Aug 1987, pp.707-713.

[13] Shen, H.C. and Wong, A.K.C., "Generalized Texture Representation and Metric,"

Computer Graphics and Image Processing, vol. 23, 1983, pp.187-206.

[14] Viglione, G.J., "Applications of Pattern Recognitions Technology," in *Adaptive, Learning and Pattern Recognition Systems,* J.M. Mendel and K.S. Fu, Eds. New York: Academic Press, 1970.