This file may not be identical to the final published version in:
Procs. IDAR-03, pp. 1060-1064, Edinburgh, UK August 03.

Towards a Ptolemaic Model for OCR

Sriharsha Veeramachaneni and George Nagy
Rensselaer Polytechnic Institute, Troy, NY, USA
E-mail: nagy@ecse.rpi.edu

Abstract

In style-constrained classification often there are only a few samples of each style and class, and the correspondences between styles in the training set and the test set are unknown. To avoid gross misestimates of the classifier parameters it is therefore important to model the pattern distributions accurately. We offer empirical evidence for intuitively appealing assumptions, in feature spaces appropriate for symbolic patterns, for (1) tetrahedral configurations of class means that suggests linear style-adaptive classification, (2) improved estimates of classification boundaries by taking into account the asymmetric configuration of the patterns with respect to the directions toward other classes, and (3) pattern-correlated style variability.

1. Introduction

To improve the classification of multi-source data we consider it necessary to extend statistical analysis beyond principal components, class means and class-conditional covariances. While we have not yet succeeded in incorporating all our findings into new classifiers, we hope that our conjectures and observations will stimulate the development of new models for classification of complex data sets.

In our training sets, each object is labeled by both source (or style) and class labels, but we must classify a test set where objects from the same source are grouped together, but neither source nor class labels are available. We wish to assign only class labels, not style labels, to the patterns of the test set. In our previous work, we have called such a scenario *style-consistent* or *style-constrained classification*, and showed that different classification methods are appropriate when the number of patterns from each source is small (2-6 patterns), or large (50 or more), and when the number of styles is small (2-6 styles), or large (hundreds) [7][8][9][10]

As in our earlier work, we assume that the objects of interest are intended for communicating messages. Examples are printed and hand-printed digits and letters of alphabetic scripts, glyphs designed specifically for ease of machine reading (graffiti and OCR fonts), and the phoneme repertory of various languages. We believe that different techniques may apply to the classification of glyphs of communications symbols than to pictures of "natural" patterns (flowers, chromosomes, tissue cells, fingerprints, faces) because communications symbols have either evolved, or were engineered, to maintain high separation between classes. We have no reason to believe that measurements of natural objects exhibit this property.

Given finite resources for producing each symbol (size and stroke-width limitations for print [5], limited ability to manipulate a stylus for hand print [6], energy budget and a fixed articulatory musculature for phonemes [2], we would expect the distance between any pair of classes to be approximately the same. (If it weren't, then it would be possible to change the symbols to separate neighboring classes at the cost of reducing the separation between distant pairs.) "Appropriate" features would maintain this equidistance property. The ten digits in a variety of scripts suggest that in a given alphabet most pairs are, in fact, roughly equally distinguishable (Figure 1). Exceptions may occur for high frequency symbols, such as '0','1', and '2' (according to Benford's Law, these digits account for 60% of all the leading digits in numerical fields), 'e' in written English, and 'schwa' in speech. An information-theoretic justification based on maximal entropy would, of course, also have to take into account linguistic context, but here we neglect inter-symbol class dependence in order to concentrate on modeling styles (inter-symbol feature dependence).

Multimodal class-conditional feature distributions have been modeled for at least 30 years [1]. Previous models have not, however, taken into account the correlation between the parameters of the modes arising from multiple pattern sources. The presence of such correlations reduces the uncertainty of the class-conditional feature distribution of a pattern if another pattern (of the same or of a different class) from the same source has already been observed. We attempt to model this phenomenon.

Figure 1. Numerals in different scripts.

2. Proposed constraints on style distributions

The curse of dimensionality impacts the estimation of style-constrained classifiers more than that of singlet classifiers because often only a few labeled samples of each style are available. This suggests applying as much prior knowledge as possible, including constraints on the expected configuration of the feature distributions.

We propose to restrict the configuration of the class means according to the equidistant-class hypothesis elaborated in the Introduction. The class means should be nearly uniformly distributed on a $(c-1)$ dimensional sphere, i.e., at the vertices of a $(c-1)$ dimensional simplex.

We further note that the class-style means in each class are dispersed about the class means in a manner similar to the dispersion of individual patterns about the class-style means. More specifically, the class-mean-centered style-mean vectors are significantly correlated. Therefore knowing how Ann writes "4" provides measurable information about how she writes "5". This phenomenon is called *strong style*. It allows classifying several patterns of the same source (in a single field) more accurately than classifying them individually [8][9]. We emphasize that this phenomenon is different from the more commonly exploited *weak style*, where patterns of the same class from a given source are similar (i.e., Ann always crosses her 7s), but may co-occur in arbitrary combinations (i.e., crossed 7s are as likely to be found in fields with open 4s as with closed 4s).

We conjecture that the correlation coefficients among features depend more on the specific feature set than on the classes and styles. This suggests pooling samples of different styles and class to estimate correlation coefficient, then rescaling the correlations into the class-specific covariance matrices according to individual variance estimates (which are more stable than covariance estimates).

The last model constraint suggested by our observations is that patterns are not distributed symmetrically about their class means. Their dispersion in the direction of the other classes is smaller than away from other classes.

We attempt to illustrate the above notions by visualizing the distributions in feature space (Figure 2). There are two features, three classes, and four styles. The class means are at the vertices of an equilateral triangle (2-D simplex). The correlation among the style means is indicated by their identical placement about their class means. The class means are not at the center of their equiprobability contours, because the feature distributions are skewed away from each other. We don't attempt to indicate the putative similarity of the correlation matrices.

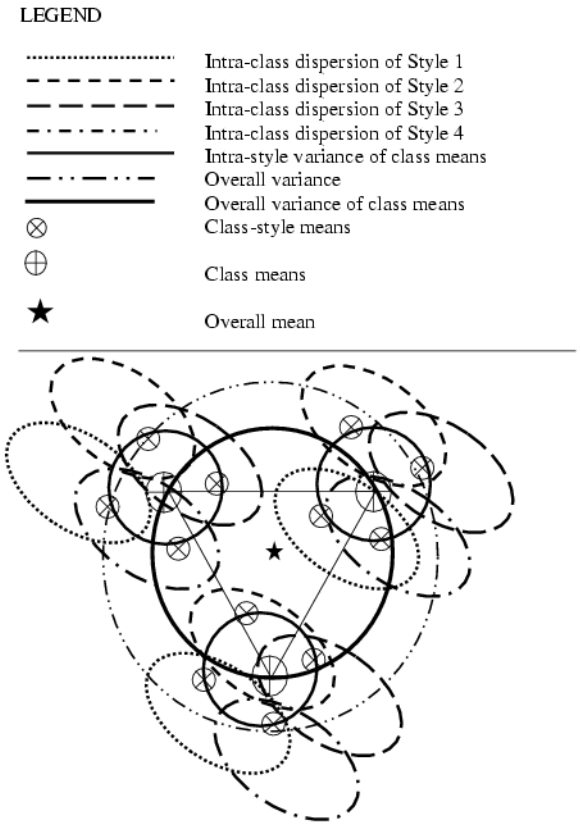

Figure 2. Model for class, style and feature distribution.

3. Database for experimentation

Databases SD3 and SD7, which are part of the NIST Special Database SD19 [3] contain handwritten numeral samples labeled by writer and class (but not of course by style). We constructed four datasets (with different writers in the training and test sets), two from each of SD3 and SD7, as shown in Table 1. We extracted 100 blurred directional (chaincode) features from each sample [4]. We then computed the principal components of the SD3-Train+SD7-Train data onto which the features of all samples were projected to obtain 100 principal-component (PCA) features for each sample.

Table 1. Handwritten numeral datasets.

	No. of Writers	No. of samples
SD3-Train	395	42698
SD7-Train	99	11495
SD3-Test	399	42821
SD7-Test	100	11660

4. Observations and justification of the model

Means of classes are indeed almost equally spaced from one another, at the vertices of a $(c-1)$-simplex (a tetrahedron in 3-D) The distribution of the Euclidean distance between class means is more uniform than the distribution of their "average" Mahalanobis distance (Figure 3). The distribution cannot be uniform in a space of lower than $c-1$ dimensions (Figure 4).

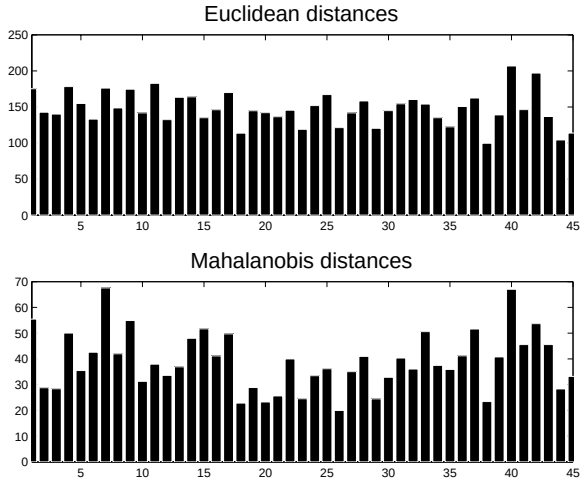

Figure 3. Bar chart of distances between all pairs of class means in 100-dimensional feature space.

The tetrahedral arrangement, which means that no class mean is a convex combination of any other class means (i.e., that the class means are on the "outside" of the feature space), coupled with the similar dispersion of the patterns about the class means, suggests that there exist "external regions" that contain mostly patterns of a single class. (These external regions are defined by hyperplanes through each class mean and normal to the vector pointing to the centroid of the other classes. The external region for each class is on the side of its own hyperplane away from the centroid, and on the side of the other hyperplanes towards the centroid.) Our experiments show that the average impurity of the external regions is only 2%.

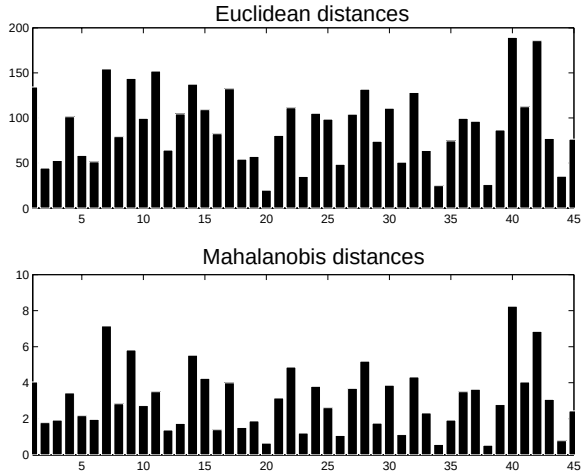

Figure 4. Bar chart of distances between all pairs of class means in top 2 principal component feature space.

If each class mean can be separated from all the other class means by means of a single hyperplane, and the dispersion of the patterns about the class means is relatively small, then a linear discriminant will be adequate. In statistical pattern recognition this is achieved in practice by using the same "average" covariance matrix for every class. In style-constrained classification the number of patterns from a single source tends to be small, therefore it is advantageous to have to estimate fewer parameters. The variance of the features does, however, change significantly from class to class, therefore we estimate the style variances, then rescale the pooled correlation coefficients (Table 2, from [9]). Class or class-and-style conditional correlations may improve classification, but only if there are enough samples to estimate them accurately.

Table 2. Character error rates on handwritten data for quadratic singlet classification and style-adaptive linear classification.

		Character error rate (%)	
Training set	Test set	Quadratic singlet	Style-adaptive linear
SD3-Train	SD3-Test	2.2	0.8
	SD7-Test	8.0	3.0
SD7-Train	SD3-Test	3.7	1.1
	SD7-Test	3.6	1.8
SD3-Train	SD3-Test	1.7	0.6
+SD7-Train	SD7-Test	4.7	1.9

The largest source of variation among patterns is the separation between class means. The minimum distance between classes increases rapidly with dimensionality in either the space of principal components of all the patterns or the principal components of the class means (Figure 5). The feature subspace spanned by the class means affords better classification than the corresponding subspace of the largest $(c-1)$ principal components.

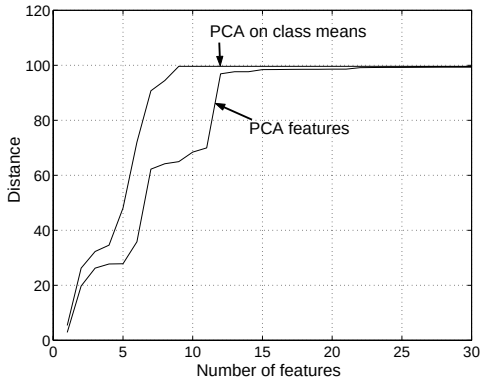

Figure 5. Minimum distance between class pairs as function of the number of features for regular PCA features and PCA features computed on the means.

The patterns are not distributed symmetrically about their class means. The variance in the direction of the other classes is smaller than away from other classes. In Table 3 "Outside" refers to patterns on the external side of the hyperplane through the class mean, and "Inside" refers to the patterns towards the centroid of the other classes. The covariances are estimated by reflecting either half of the patterns about the class mean. The error rate of a quadratic discriminant varies by a factor of five depending on which half of the patterns of each class is used for estimating the class-conditional covariance matrices. This observation may eventually lead to improved methods of regularization for estimating covariance matrices.

Table 3. Quadratic discrimination error rates (in %) with different covariance estimates.

Training set	Test set	Full covariance	Inside reflected	Outside reflected
SD3-Train	SD3-Test	2.2	1.6	8.2
SD7-Train	SD7-Test	3.6	2.8	12.3
SD3-Train+ SD7-Train	SD3-Test SD7-Test	1.7 4.7	1.4 3.4	7.0 16.2

The class-style means within each style are correlated, giving rise to "strong style". This implies that the mean vector of one class is predictable from the mean vector of another class from the same style. Therefore classifying several patterns of the same source as a single field yields a lower error rate than classifying them individually [8][9].

Even though the patterns do not obey a Gaussian distribution, various covariance matrices provide useful information for classification. The covariance of the patterns centered about their class-and-style means describes the "within-style variation." The covariance of the class-and-style means about their class mean represents the "style-to-style" variation. Each overall class-conditional covariance matrix is the sum of the style covariance matrix and the within-style noise covariance matrix (Figure 6).

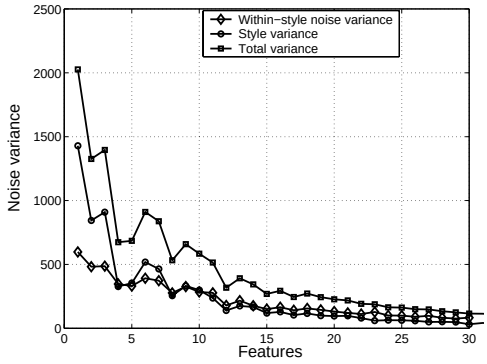

Figure 6. Within-style noise, style and total variance for the top 30 PCA features.

The ratios of the within-style to between-style variances are close to unity in the original feature space (Figure 7). The principal components that discriminate best between classes are also most effective for separating styles.

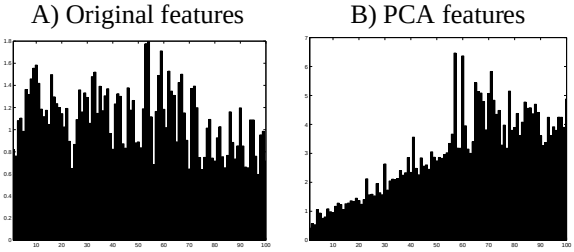

Figure 7. Ratio of the within-style noise variance to the style variance (variances averaged over all classes) for each feature.

The style means of the patterns of different styles are distributed in a roughly Gaussian fashion about overall class

Table 4. Correlations between means of class-pairs. The correlation coefficients with the maximum absolute value over all feature pairs are shown.

	0	1	2	3	4	5	6	7	8	9
0	1.00	-0.79	0.61	0.71	0.74	0.65	0.80	0.75	0.68	0.80
1	-0.79	1.00	0.64	-0.70	0.84	0.71	-0.75	-0.84	0.75	-0.85
2	0.61	0.64	1.00	0.67	-0.55	0.55	0.55	0.59	-0.53	0.61
3	0.71	-0.70	0.67	1.00	0.65	0.63	0.64	0.66	0.66	0.74
4	0.74	0.84	-0.55	0.65	1.00	0.68	0.72	0.81	0.70	0.79
5	0.65	0.71	0.55	0.63	0.68	1.00	0.63	0.67	0.63	0.64
6	0.80	-0.75	0.55	0.64	0.72	0.63	1.00	0.71	0.64	0.70
7	0.75	-0.84	0.59	0.66	0.81	0.67	0.71	1.00	-0.69	0.81
8	0.68	0.75	-0.53	0.66	0.70	0.63	0.64	-0.69	1.00	0.72
9	0.80	-0.85	0.61	0.74	0.79	0.64	0.70	0.81	0.72	1.00

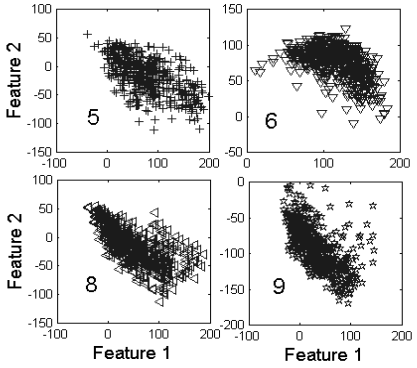

Figure 8. Scatter plot of the top two PCA features of writer-specific class means.

means (Figure 8). Correlation coefficients between style means are shown in Table 4 . If there were no strong style, all the off-diagonal elements would be zero.

The concept of style can be extended to a deeper hierarchy. For instance, we can consider NIST SD3 and SD7, which exhibit significantly different appearances, class means, covariance matrices, and error rates, as two super-styles. In print, we could consider serif and sans serif fonts as different super-styles.

5. Conclusions

We are working towards accurate, widely applicable prior information that can be used as a sanity check by partially supervised style-adaptive classifiers. We argued that appropriate feature representation of symbolic patterns intended for communication between individuals must maximize the minimum separation between classes. We have shown experimentally that this is preserved by a tetrahedral configuration of the class means in feature space. We presented evidence for the asymmetric distribution of patterns about their means and showed how this can be exploited for improved classification boundaries. We demonstrated that the style means are closely grouped about their respective style means, and that at least some feature values are highly correlated between patterns of the same style, which is useful for style classification.

We thank Dr. Hiromichi Fujisawa and Dr. Cheng-Lin Liu for their valuable suggestions and advice.

References

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

[2] J. Greenberg. *Universals of Human Language*, volume 2. Stanford University Press, 1978.

[3] P. Grother. Handprinted forms and character database, NIST special database 19, March 1995. Tech. Rpt. and CDROM.

[4] C. L. Liu, H. Sako, and H. Fujisawa. Performance evaluation of pattern classifiers for handwritten character recognition. *IJDAR*, 4(3):191–204, 2002.

[5] R. McLean. *Typography*. Thames & Hudson, London, 1980.

[6] R. Plamondon. A kinematic theory of rapid human movements. *Parts I, II, and III, in Biological Cybernetics*, pages 72(4) 320, 72(4) 295–307 (1995), 78 133–135 (1999).

[7] P. Sarkar and G. Nagy. Classification of style-constrained pattern fields. In *Proceedings of ICPR-15*, pages 859–862, 2000.

[8] P. Sarkar and G. Nagy. Style consistency in isogenous patterns. In *Proceedings of ICDAR-6*, pages 1169–1174, 2001.

[9] S. Veeramachaneni. *Style constrained quadratic field classifiers*. PhD thesis, Rensselaer Polytechnic Institute, Troy, New York, 2002.

[10] S. Veeramachaneni, H. Fujisawa, C. L. Liu, and G. Nagy. Style-conscious quadratic field classifier. In *Proceedings of ICPR-16*, pages 72–75, August 2002.