CLASSIFYING ISOGENOUS FIELDS

S. VEERAMACHANENI, G. NAGY
Rensselaer Polytechinic Institute, Troy, NY, USA
E-mail: veeras@rpi.edu

C. -L. LIU, H. FUJISAWA
Hitachi Central Research Laboratory, Tokyo, Japan

In a previous report, we presented a classifier that utilizes style context in co-occurring patterns to increase recognition accuracy. Our method requires only class and source labeled data for training. It is obvious that when patterns occur as long isogenous fields, this gain is increased. We show that our method achieves higher accuracy with longer input fields because it can be trained accurately. We also present some ongoing work on simple heuristics to reduce computational complexity of the scheme.

1 Introduction

Stylistic variations in isolated handwritten digits, while not as striking as in cursive script, are significant. They are most often attributed to differing conventions of numeral formation among countries, schools and teachers. Practice turns laborious tracing and imitation of exemplars into ingrained habit. It is advantageous to include under *style* also the systematic variations due to writing instrument, position of writer and medium, constraints imposed by ruled lines or boxes, and motivation of the writer.

Researchers have devised many means of accommodating the multiplicity of forms that may correspond to a given symbol. What distinguishes style-conscious classification is the exploitation of the assumption that all patterns from a single *source* exhibit the same style. Style-consistent classification helps to eliminate errors due to inter-style confusions. Such confusions result from similar renderings, in separate styles, of two different symbols (classes).

As is typical in form-reading applications, in the NIST handwritten datasets each page was written by a single writer in a single session, and therefore presumably in the same style (several writers may of course exhibit the same style). Although styles are seldom (never ?) explicitly labeled in the training sets, sources or fields usually are.

Several recent papers presented methods to exploit style context i.e., feature dependencies between co-occurring patterns (fields of patterns) to improve recognition accuracy [1,2,3]. We have proposed a *style-conscious quadratic discriminant function* (SQDF) field classifier for isogenous fields [4]. Under the assumption that the input patterns occur in pairs, we have demonstrated that utilization of style context can yield higher accuracy than the singlet classifier which is oblivious to any style consistency.

It is evident that a style-consistent classifier should benefit from longer input fields since more information regarding the style can be gleaned. It is vital that a style-conscious classification method not require estimation of additional parameters to classify longer fields so that the gains obtained are not offset by finite-sample estimation errors. For large volume applications the classifier also should be computationally efficient.

We required style labeled data to compare our quadratic field classifier with other style-conscious classifiers. We used printed numerals in various typefaces. We present results on handwritten numerals to demonstrate that our method does not require style labeled data.

2 Problem formulation

We consider the problem of classifying an isogenous field $\mathbf{y} = (x_1^T, \ldots, x_L^T)^T$ (each x_i represents d feature measurements for one of L patterns in the field) generated by one of S sources $s_1, s_2, \ldots, s_S$ (writers, fonts, etc.).
Let $\Omega = \{\omega_1, \ldots, \omega_N\}$ be the set of singlet-class labels. Let c_i represent the identity of the i^{th} pattern of the field.

We make the following assumptions on the class and feature distributions.
1. $p(s_k|\, c_1, c_2, \ldots, c_L) = p(s_k)\ \forall k = 1, \ldots, S$. That is, any linguistic context is source independent.
2. $p(\mathbf{y}|\, c_1, c_2, \ldots, c_L, s_k) = p(\mathbf{x}_1|\, c_1, s_k)p(\mathbf{x}_2|\, c_2, s_k) \ldots\ p(\mathbf{x}_L|\, c_L, s_k)\ \forall k = 1, \ldots, S$. The features of each pattern in the field are class-conditionally independent of the features of every other pattern in the same field.

Under these assumptions the mean $\mu_{i,j,\ldots,k}$ and covariance matrix $K_{i,j,\ldots,k}$ for the field-class $\mathbf{c} = (c_1, c_2, \ldots, c_L)^T = (\omega_i, \omega_j, \ldots, \omega_k)^T$ are given by

$$\mu_{i,j,\ldots,k} = \begin{pmatrix} \mu_i \\ \mu_j \\ \vdots \\ \mu_k \end{pmatrix};\ K_{i,j,\ldots,k} = \begin{pmatrix} C_i & C_{ij} & \ldots & C_{ik} \\ C_{ji} & C_j & \ldots & C_{jk} \\ \vdots & \vdots & \ddots & \vdots \\ C_{ki} & C_{kj} & \ldots & C_k \end{pmatrix}$$

$$\begin{aligned} \text{where } \mu_i &= E\{\mathbf{x}_1|\, c_1 = \omega_i\} \\ C_i &= E\{\mathbf{x}_1\mathbf{x}_1^T|\, c_1 = \omega_i\} - E\{\mathbf{x}_1|\, c_1 = \omega_i\}E\{\mathbf{x}_1^T|\, c_1 = \omega_i\} \\ C_{ij} &= E\{\mathbf{x}_1\mathbf{x}_2^T|\, c_1 = \omega_i, c_2 = \omega_j\} - E\{\mathbf{x}_1|\, c_1 = \omega_i\}E\{\mathbf{x}_2^T|\, c_2 = \omega_j\} \\ &= E_{\mathbf{s}}\{E\{\mathbf{x}_1\mathbf{x}_2^T|\, c_1 = \omega_i, c_2 = \omega_j, s_k\}\} - \mu_i\mu_j^T \qquad (1) \end{aligned}$$

$$= E_{\mathbf{s}}\{E\{\mathbf{x}_1 | c_1 = \omega_i, s_k\} E\{\mathbf{x}_2 | c_2 = \omega_j, s_k\}\} - \mu_i \mu_j^T \qquad (2)$$

and similarly,

$$C_{ii} = E\{\mathbf{x}_1 \mathbf{x}_2^T | c_1 = \omega_i, c_2 = \omega_i\} - E\{\mathbf{x}_1 | c_1 = \omega_i\} E\{\mathbf{x}_2^T | c_2 = \omega_i\}$$
$$= E_{\mathbf{s}}\{E\{\mathbf{x}_1 | c_1 = \omega_i, s_k\} E\{\mathbf{x}_2 | c_2 = \omega_i, s_k\}\} - \mu_i \mu_i^T \qquad (3)$$

The class-pair-conditional Expectations are subject to $\mathbf{x}_1$ and $\mathbf{x}_2$ being from one source. Therefore, in Equations 1, 2 and 3, the inner Expectations are source conditional and the outer Expectations are over all sources. C_{ij} captures the dependence between patterns from different classes in the same field, whereas C_{ii} captures the dependence between patterns from the same class. (C_{ii} and C_{ij} are the off-diagonal blocks in the covariance matrices, C_{ii} for same-class pairs and C_{ij} for different-class pairs.)

3 Training the field classifier

According to the above expressions, the means and covariance matrices for any field-class can be estimated from class and source labeled training data from correlations between pairs of patterns. Hence the *style conscious quadratic classifier* (SQDF) [4], derived in Section 2, can be constructed for fields of any length without any training beyond that required for pairs.

Let us assume for simplicity that each source corresponds to a distinct style and that no linguistic context is present. Then an *optimal style-conscious classifier* (OPT) [1,5] would assign the label $c^* = (\omega_{i_1}, \ldots, \omega_{i_L})$ to the input field $\mathbf{y}$ with the maximum a posteriori probability $p(c^* | \mathbf{y}) \propto \sum_{k=1}^{S} p(\mathbf{x}_1 | \omega_{i_1}, s_k) \ldots p(\mathbf{x}_L | \omega_{i_L}, s_k) p(s_k)$. Under the assumption that style-conditional feature distributions, i.e., $p(x | \omega_i, s_k)$, are normal for each singlet class, the OPT classifier can be constructed when source and class labeled training data is available.

Another approach to utilizing style context is to require a training set with isogenous fields of length L and construct a *quadratic discriminant function field classifier* (F) to recognize the input fields. Table 1 compares the number of parameters to be estimated for each of the above classifiers. Clearly the OPT and the SQDF are the better choices because the number of parameters is independent of the field length. The F classifier is unsuitable for classifying long fields because of the exponential increase in the number of parameters with field length. The OPT classifier requires style labeled data for training, in the absence of which methods such as the EM algorithm are required to estimate the style and class conditional feature distributions.

Classifier	Number of parameters
Quadratic discriminant field classifier (F)	$O(N^L d^2)$
Optimal style-conscious classifier (OPT)	$O(NSd^2)$
Style-conscious quadratic field classifier (SQDF)	$O(N^2 d^2)$

Table 1: Number of parameters to be estimated for various style-conscious classifiers. N is the number of classes, L is the field length, S is the number of styles and d is the dimensionality of the singlet feature space.

4 Classifying longer fields

The style-conscious classifiers are computationally expensive due to the exponential increase in the number of field-classes with L as well as due to the complex discriminant function computation per field-class. We have tested some heuristics to reduce computation while utilizing style-context present in longer fields. Since most of the style information in the field can be extracted from looking at the patterns one pair at a time, we conjectured that it is possible to classify a field by computing the style dependecies between all pairs of patterns.

Given an input field $\mathbf{y} = (x_1^T, \ldots, x_L^T)^T$ we attempt to approximate its similarity $s(\mathbf{y}; \mathbf{c})$ to class $\mathbf{c} = (\omega_{i_1}, \ldots, \omega_{i_L})$ from the probabilities $p(x_j, x_k | \omega_{i_j}, \omega_{i_k})$ $j = 1, \ldots, L-1, k = j+1, \ldots, L$. The other pair densities are not required because, under our assumptions,
$p(\mathbf{x}_1 = x_1, \mathbf{x}_2 = x_2 | c_1 = \omega_i, c_2 = \omega_j) = p(\mathbf{x}_1 = x_2, \mathbf{x}_2 = x_1 | c_1 = \omega_j, c_2 = \omega_i)$. This will enable us to classify a field of length L by performing $L(L-1)/2$ pair classifications. Heuristics H1 and H2 are based on the conjecture that $p(x_1, x_2 | x_3, c_1, c_2, c_3) \approx p(x_1 | x_3, c_1, c_3) p(x_2 | x_3, c_2, c_3)$ for high ranking field classes given the feature vector.

Pairs & singlets heuristic (H1): We compute the similarity of $\mathbf{y}$ to field-class $\mathbf{c} = (\omega_{i_1}, \ldots, \omega_{i_L})$ according to
$s(\mathbf{y}; \mathbf{c}) = (\prod_{j=1}^{L-1} \prod_{k=j+1}^{L} p(x_j, x_k | \omega_{i_j}, \omega_{i_k}))^2 / (\prod_{j=1}^{L} p(x_j | \omega_{i_j}))^{L-2}$.

All-pairs heuristic (H2): We compute the similarity of $\mathbf{y}$ to field-class $\mathbf{c} = (\omega_{i_1}, \ldots, \omega_{i_L})$ according to
$s(\mathbf{y}; \mathbf{c}) = \prod_{j=1}^{L-1} \prod_{k=j+1}^{L} p(x_j, x_k | \omega_{i_j}, \omega_{i_k})$.

Best-pair heuristic (H3): We classify each singlet x_j in the field by assigning to it the class label obtained from the pair label
$(c_j, c_k)^* = \underset{(\omega_m, \omega_n)}{\operatorname{argmax}} \{p(x_j, x_k | \omega_m, \omega_n),\ k = 1, \ldots, L,\ k \neq j,\ m, n = 1, \ldots, N\ \}$.

Consistency heuristic (H4): We classify each pair in the field using a style-conscious field classifier. *Consistent pair classification* of the field signifies that each singlet in the field is assigned the same label by all of its $L-1$ pair classifiers. If a field is consistently classified, we assign to it the labels obtained from pair classification, otherwise, a N^L-class quadratic classifier is used to classify the field. For a classifier with low error rate, most fields are consistently classified, which means that the expensive long-field classifier will be seldom invoked.

5 Experiments on handwritten numerals

We used the databases SD3 and SD7, which are contained in the NIST Special Database SD19 [6]. The database contains handwritten numeral samples labeled by writer and class (but not of course by style). SD3 was the training data released for the First Census OCR Systems Conference and SD7 was used as the test data. We constructed four datasets, two from each of SD3 and SD7, as shown in Table 2. Since we compute the field class-conditional covariance matrices from source-specific class-conditional covariances we require that each writer have at least two samples of each singlet class. We therefore deleted all writers not satisfying this criterion from the training sets.

	Writers	Number of samples
SD3-Train	0-399 (395)	42698
SD7-Train	2100-2199 (99)	11495
SD3-Test	400-799 (399)	42821
SD7-Test	2200-2299 (100)	11660

Table 2: Handwritten numeral datasets

We extracted 100 blurred directional (chaincode) features from each sample [7]. 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 features for each sample. The samples of each writer in the test sets were randomly permuted to simulate fields without any linguistic context.

The classification results for the SQDF classifier are presented in Table 3. The performance of various heuristics to classify fields of length $L=3$ are also shown. The actual counts are shown instead of the percentages, because the values are close.

Test set	Features	SQDF				
		$L = 1$	$L = 2$	$L = 3$		
				H1	H2	H3
SD3-Test (42821)	All 100	746	712	713	715	725
SD7-Test (11660)	All 100	551	534	527	531	538
SD3-Test+SD7-Test	All 100	1297	1246	1240	1246	1263

Table 3: Number of character errors for various experiments
(Training set = SD3-Train + SD7-Train)

The OPT classifier requires style labeled data. In order to compare the gains in accuracy obtained by the SQDF classifier with those obtained by the OPT classifier we experimented with machine-printed numerals with known style labels.

6 Experiments on machine-printed numerals

A database of multi-font machine-printed numerals was generated as follows[1]. Six pages, containing the ten digits 0-9 spaced evenly and replicated 50 times, were prepared using Microsoft Word 6.0. Each page was rendered in a different typeface, namely Arial (6 pt), Avant Garde (6 pt), Bookman Old Style (6 pt), Helvetica (6 pt), Times New Roman (6 pt), and Verdana (6 pt), and printed on a 600 dpi Apple LaserWriterSelect. Each page was scanned 10 times at 200 dpi into 10 bilevel bitmaps using an HP flatbed scanner. This yielded a total of 30,000 samples. Since the typeface Arial was unavailable on the computer on which the MS Word files were generated, it was substituted with Helvetica. Therefore there are only five typefaces in the data. This allows us to verify the property of our method to correctly identify style-context when multiple sources share the same style. A few of the samples are shown in Figure 1. Some possibly ambiguous patterns are shown in Figure 2. The resulting scanned images were segmented and for each digit sample 64 blurred directional (chaincode) features were extracted and stored [7]. For each typeface 2500 samples were included in the training set, while the remaining 2500 samples were randomly permuted to simulate fields and used for testing. That is, in all 15,000 samples were used for training and 15,000 for testing.

If the typeface labels for the test patterns are known, we can classify them by using a *typeface-specific classifiers* (TS) (i.e., six quadratic classifiers are trained separately on different typefaces and the appropriate classifier is used to classify each typeface). We extracted principal component features, of which the top 32 features were chosen for experimentation. The error rate

Figure 1: Samples of the machine-printed digits, reproduced at approximately actual size

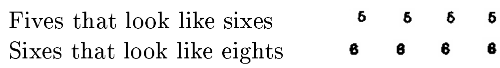


Figure 2: Some ambiguous patterns

obtained with the top 32 features is so small that the advantage of modeling style consistency is obscured. Hence we present results using the top 8 principal component features as well.

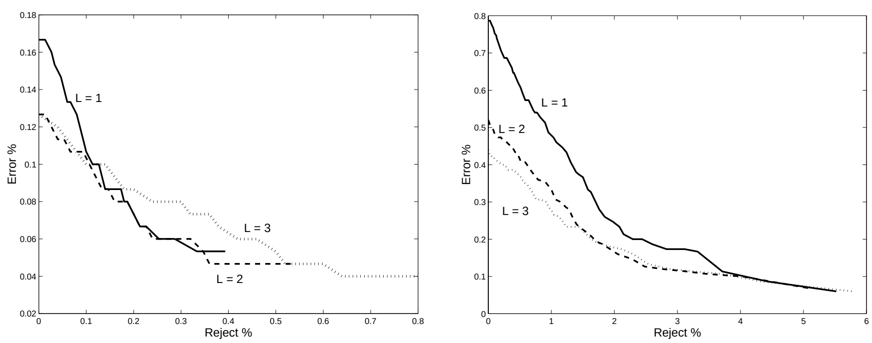
The OPT classifier was constructed using one Gaussian per typeface per class (i.e., a total of 6 Gaussians). The main difference between the TS and the OPT classifier is that the OPT classifier does not require style labeled test data.

Table 4 shows the character classification error rates of the TS, OPT and the SQDF classifiers on the machine print data for increasing field lengths with both the top 32 and top 8 principal component features. It is evident that, ideally, no style-conscious classifier can outperform the TS classifiers. However, the results indicate that the OPT classifier is better than the TS classifiers for $L > 1$ when 32 features are used. This counter-intuitive result can be attributed to the small-sample estimation errors (each TS classifier uses a 32-feature quadratic discriminant function trained on only 250 samples/class/typeface). We believe that even with infinite training data, the error-rate of the style-conscious classifiers approaches a lower limit (the intra-style error rate) with increasing field length. The gain achieved from $L = 2$ to $L = 3$ is lower than that from $L = 1$ to $L = 2$, because most of the inter-style errors have already been corrected.

The character reject-error curves for the SQDF classifier for $L = 1$, $L = 2$ and $L = 3$ are shown in Figure 3. The reject criterion for fields of all lengths is based on thresholding the *a posteriori* probability of the assigned label. We observe that the initial slope of the curves is smaller for longer field lengths. Typically two fields (i.e., $2L$ characters) must be rejected to eliminate one character error [8]. Fields of length 6 were constructed and classified by the $L = 1$, $L = 2$ and $L = 3$ SQDF classifiers operating on L patterns at a time. Figure 4 shows corresponding *field* reject-error curves. Classification with increasing field length yields increasingly dominant field reject-error characteristics.

Classifier	L	Number of errors	
		Top 32 pca	Top 8 pca
TS	1	11 (0.07%)	37 (0.25%)
OPT	1	11 (0.07%)	50 (0.33%)
	2	7 (0.05%)	37 (0.25%)
	3	6 (0.04%)	36 (0.24%)
SQDF	1	25 (0.17%)	118 (0.79%)
	2	19 (0.13%)	78 (0.52%)
	3	19 (0.13%)	66 (0.44%)

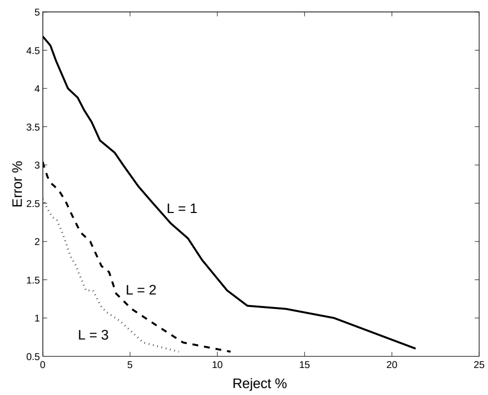
Table 4: Classification results for various classifiers and field lengths



Top 32 pca features　　　Top 8 pca features

Figure 3: Character reject-error curves for the SQDF classifier

We compared the heuristics described previously on the machine-print data for $L = 3$. For H1 and H2 we compared similarity values for only a few of the field classes obtained from the top candidates for each pair in the field. The error rates for each of the heuristics are presented in Table 5. We observe that all but H3 have a lower error rate than the 0.52% obtained by the SQDF pair classifier ($L = 2$). On this data set the percentage of fields with inconsistent pair labels was 0.98% (on which the expensive triple classification was performed for heuristic H4). H4 also requires more storage because triple inverse covariance matrices need to be stored. It can be shown that due to the nature of the covariance matrices, there is no need to store different inverse covariance matrices for field-classes that are permutations of one another. This



Top 8 pca features

Figure 4: Field reject-error curves for the SQDF classifier (for fields 6 characters long)

greatly reduces the amount of storage required.

Heuristic	Number of errors
H1	68 (0.45%)
H2	70 (0.47%)
H3	83 (0.55%)
H4	66 (0.44%)

Table 5: Classification results for various heuristics for $L = 3$ using top 8 pca features

7 Discussion and future work

Exploiting style context in longer fields reduces the error rate of both the *optimal style-conscious* (OPT) classifier and the *style-conscious quadratic discriminant* (SQDF) classifier. The SQDF pair classifier gains over the singlet classifier on both the machine-printed and handwritten data. On multi-font machine-printed data, the error rate of the SQDF classifier decreases faster with increasing field length than that of the almost error-free OPT classifier. The SQDF classifier proffers no advantage over a singlet classifer blind to style for character classification with high reject rate, but is significantly better for *field rejection*.

The parameters of the SQDF classifier are easy to estimate from source-specific data. No prior information about the distribution of styles is required.

Classifying only inconsistent fields with the more expensive triple classifier is effective on the machine-printed data, while the methods that we have tried so far to combine multiple pair classifiers to classify longer fields are only moderately effective on both datasets. This suggests that better heuristics and algorithms need to be explored so that the SQDF classifier can benefit from long input fields.

We observe that the reduction in error rate is more significant going from $L = 1$ to $L = 2$ than from $L = 2$ to $L = 3$. One cause for this effect is that most of the style-context in a field can be obtained from pairs, i.e., most inter-style confusions are corrected by the pair classifier. Also, the triple classifier is more sensitive to estimation errors. Although the SQDF classifier is constructed using pair correlations, estimation errors are likely to increase the classifier bias and variance with field length.

References

1. P. Sarkar and G. Nagy. Style consistency in isogenous patterns. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 1169–1174, 2001.
2. T. Kawatani. Character recognition performance improvement using personal handwriting characteristics. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, volume 1, pages 98–103, 1995.
3. T. Koshinaka, D. Nishikawa, and K. Yamada. A stochastic model for handwritten word recognition using context dependency between character patterns. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 154–158, 2001.
4. S. Veeramachaneni, H. Fujisawa, C.-L. Liu, and G. Nagy. Style-conscious quadratic field classifier, 2002. Submitted to the Sixteenth International Conference on Pattern Recognition.
5. P. Sarkar. *Style consistency in pattern fields.* PhD thesis, Rensselaer Polytechnic Institute, Troy, NY, 2000.
6. P. Grother. Handprinted forms and character database, NIST special database 19, March 1995. Technical Report and CDROM.
7. C.-L. Liu, H. Sako, and H. Fujisawa. Performance evaluation of pattern classifiers for handwritten character recognition. *International Journal on Document Analysis and Recognition*, in press 2002.
8. C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, IT-16(1):41–46, 1970.