Style Consistent Classification of Isogenous Patterns

Prateek Sarkar, *Member*, *IEEE Computer Society*, and George Nagy, *Life Fellow*, *IEEE*

Abstract—In many applications of pattern recognition, patterns appear together in groups (*fields*) that have a common origin. For example, a printed word is usually a field of character patterns printed in the same font. A common origin induces consistency of style in features measured on patterns. The features of patterns co-occurring in a field are statistically dependent because they share the same, albeit unknown, style. *Style constrained classifiers* achieve higher classification accuracy by modeling such dependence among patterns in a field. Effects of style consistency on the distributions of field-features (concatenation of pattern features) can be modeled by hierarchical mixtures. Each field derives from a mixture of styles, while, within a field, a pattern derives from a class-style conditional mixture of Gaussians. Based on this model, an optimal style constrained classifier processes entire fields of patterns rendered in a consistent but unknown style. In a laboratory experiment, style constrained classification reduced errors on fields of printed digits by nearly 25 percent over singlet classifiers. Longer fields favor our classification method because they furnish more information about the underlying style.

Index Terms—Style, isogenous patterns, style consistency, style constrained classification, style-bound variant, style-shared variant, Optical Character Recognition, font recognition, field classification, mixture model.

1 Introduction

An accomplished typeface designer makes use of only a few well-matched typographic components—bowls, stems, bars, finials, and serifs—to configure an entire alphabet. Fine penmanship is also characterized by a certain uniformity of strokes and spacing, although the overall aspect varies from writer to writer. The subtle relationship between words and phrases that identify a particular author is yet another instance of the notion of style.

Aesthetic considerations aside, messages from the same source tend to share similar characteristics. Printers, copiers, and scanners all leave their imprint on digitized documents (Fig. 1). Even mediocre writers display a certain predictable uniformity (Fig. 2).

The goal of the work reported here is to model "style" mathematically and thereby develop a basis for more accurate classification of a group (*field*) of digitized characters from the same source. We do not assume that the source of a field is known, only that all the patterns in the same field are isogenous, i.e., they originate from the same source. The problem is not without practical relevance: the feature-space representation of any single postal address, bureaucratic form, or printed article is bound to display some measure of homogeneity due to isogeny. Human readers subconsciously make use of this phenomenon [15].

Our work is rooted in established principles of statistical minimum-error classification, as set forth in [7]. Expectation Maximization (EM), used for estimating the parameters of style consistent mixture models, is discussed and well referenced in the new edition [8].[1]

In contrast to methods based on linguistic context (letter n-grams [11], lexicons [6], [22], HMMs [14], and word-matching [10]), *style consistent classification* (alias *style constrained classification*) does not depend on the order of the patterns. The notion of exploiting spatial context was suggested in [13] and it was associated with the word style in [23], [5], [4]. Style constrained classification is closely related to font identification as practiced, for instance, in [29]. However, none of these studies presents a unified model of classification where feature distributions estimated from training samples without style labels are used to classify same-style fields. We presented such a model in [19] and [20], drawing heavily on [16]. Some closely related style models based on a single high-dimensional Gaussian distribution per class were described in [27], [26], [24].

In Section 2, we formulate the mathematical apparatus necessary to extend the optimal classification of feature vectors with given class-conditional feature probability distributions to the optimal classification of field feature vectors given style-and-class-conditional probability distributions. Although we assume that all fields are a priori equally probable, we show how linguistic context, if present, can be accommodated to advantage. We describe alternative style hypotheses that lead to different assignments of mixtures of Gaussian densities to each class and style. Because the optimal classifiers are computation intensive, we also propose a suitable approximation.

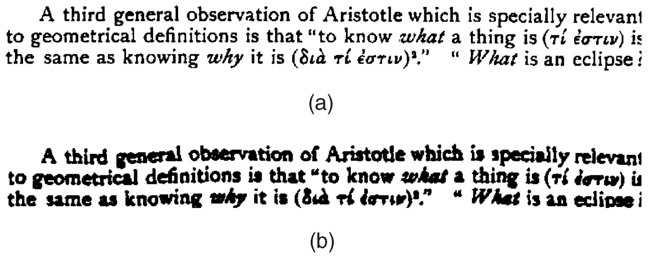
Researchers often illustrate differences between classifiers by depicting decision boundaries in a two-dimensional feature space. Style constrained classification requires at least two patterns in a field; therefore, a two-dimensional representation permits only one feature per pattern. Nevertheless, even a simple problem, with only one scalar feature,

- P. Sarkar is with the Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304. E-mail: psarkar@parc.com.
- G. Nagy is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180. E-mail: nagy@ecse.rpi.edu.

Manuscript received 4 Nov. 2002; accepted 21 May 2004.
Recommended for acceptance by L. Vincent.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 117702.

[1] The EM iteration formulae for estimating mixtures of Gaussians appeared in the first edition of [7], even before the complete EM formulation appeared in [9].

0162-8828/05/$20.00 © 2005 IEEE Published by the IEEE Computer Society



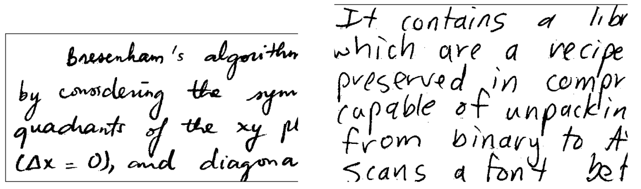
A third general observation of Aristotle which is specially relevant to geometrical definitions is that "to know *what* a thing is (τί ἐστιν) is the same as knowing *why* it is (διὰ τί ἐστιν)²." "*What* is an eclipse?

(a)

A third general observation of Aristotle which is specially relevant to geometrical definitions is that "to know *what* a thing is (τί ἐστιν) is the same as knowing *why* it is (διὰ τί ἐστιν)²." "*What* is an eclipse?

(b)

Fig. 1. Variation of style in document images can be a result of the printing and imaging processes. The text segment was scanned (a) before and (b) after multiple photocopying.



bresenham's algorithm
by considering the sym
quadrants of the xy pl
(Δx = 0), and diagona

It contains a libr
which are a recipe
preserved in compr
capable of unpackin
from binary to A
scans a font bet

Fig. 2. Handwriting styles depend on the training of the writer. Character formation is "cursive" in the left example and "blocked" in the right.

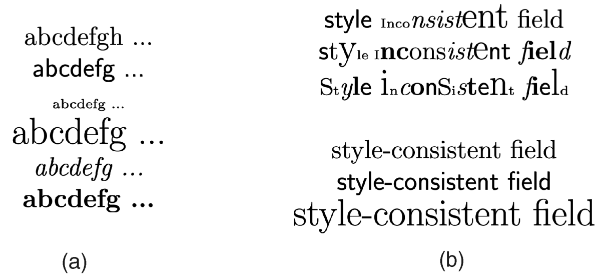
two classes, and two styles, gives considerable insight into the nature of style constrained classification. Such a problem is carefully formulated and studied in Section 3. We explore the consequences of different class and style configurations and of different hypotheses for the assignment of the Gaussian distributions, illustrate the resulting classification boundaries in 2D, and tabulate error rates. These simulation examples reveal when style consistent classification is appropriate and when it is not.

We validate our findings by comparing style constrained classification with conventional "singlet" classification on scanned alphabets of six different machine fonts. The experimental design allocates equal resources to each classifier. As we had expected, style constrained classification yields significant gains over conventional multifont classification, though it cannot quite match the error rates obtained by a style-specific classifier trained on the same font as the test field. We do not claim that this is a "real-life" test because we use well-spaced character images with the same (large) number of patterns from each class for training. A few simple feature measurements were used to represent the character patterns. Nevertheless, we believe that the results warrant investigating the benefits of style classifiers for operational products. Advances in computer speed or improved numerical techniques should also eventually render our methods applicable to speech recognition.

2 Mathematical Formulation of Style Consistency

We consider fields of L isogenous patterns, represented by feature vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L$. Each pattern, $\boldsymbol{x}_l$, belongs to one of C classes: $c_l \in 1, 2, \ldots, C$. The object of classification is to deduce the class of each pattern from the observed feature vectors.

For each field, we define the field feature vector, $\mathbf{x}$, as the concatenation, $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L)$, of the constituent pattern feature vectors. The concatenation of pattern classes is called the field-class or field identity, $\mathbf{c} = (c_1, c_2, \ldots, c_L)$.



abcdefgh ...
abcdefg ...
abcdefg ...
abcdefg ...
abcdefg ...
abcdefg ...

(a)

style Inconsistent field
stYle Inconsistent field
Style InconSistent field

style-consistent field
style-consistent field
style-consistent field

(b)

Fig. 3. (a) Symbols of the English alphabet in various styles—each character class has style-specific variants. (b) Notion of style consistency in fields.

The essential aspect of a style-consistency model is the statistical dependence among pattern-features in a field. In contrast, in a *singlet* model, we assume that pattern-features in a field are class-conditionally independent.

$$p(\mathbf{x}|\mathbf{c}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L | c_1, \ldots, c_L) = \prod_{l=1}^{L} p(\boldsymbol{x}_l | c_1, \ldots, c_L). \quad (1)$$

We can simplify further by assuming that the lth pattern-feature, $\boldsymbol{x}_l$, depends on the class of the lth pattern but is independent of all of the other pattern-classes.

$$p(\mathbf{x}|\mathbf{c}) = \prod_{l=1}^{L} p(\boldsymbol{x}_l | c_l). \quad (2)$$

There are known exceptions to the last assumption, such as ligatures in print and handwriting (the shape of the "i" is different after the "f" than after the "n" or "t" in the word "definitions" in Fig. 1) and coarticulation in speech. In speech recognition, context trees model feature densities in the context of the field-class [12].

When each pattern-class can be rendered in different styles, the resulting pattern-class conditional pattern-feature probability is a mixture distribution. For K styles, $1, \ldots, K$

$$p(\boldsymbol{x}_l | c_l) = \sum_{k=1}^{K} \alpha_k p(\boldsymbol{x}_l | k, c_l),$$

where α_k is the probability of occurrence of style k. The field-class conditional field-feature density is (substituting above in (2))

$$p(\mathbf{x}|\mathbf{c}) = \prod_{l=1}^{L} \sum_{k=1}^{K} \alpha_k p(\boldsymbol{x}_l | k, c_l).$$

While the above formula accounts for multiple styles of patterns, it does not model the consistency of style within a field. Thus, different patterns in a field can be randomly generated from different styles. We illustrate this in Fig. 3 (top right). Our notion of style consistency is pragmatic, induced by the observation of frequent co-occurrence. Thus, the top three examples in Fig. 3b have been labeled style inconsistent only because such combinations of font-variants are rarely seen.

In our style-consistency model, field-features have mixture distributions induced by styles, while, within a field, all patterns come from the same style.

$$p(\mathbf{x}|\mathbf{c}) = \sum_{k=1}^{K} P[k|\mathbf{c}]\, p(\mathbf{x}|k,\mathbf{c}) = \sum_{k=1}^{K} \alpha_k\, p(\mathbf{x}|k,\mathbf{c}). \qquad (3)$$

The assumption is that the style of rendering of a field is independent of the identity of the field being rendered ($P[k|\mathbf{c}] = P[k] = \alpha_k$).[2] Within each style, we assume that a pattern-feature is independent of the class-labels of other patterns in the field (compare to (2) for the singlet model). The style consistent class-conditional field-feature probability can then be written as:

$$p(\mathbf{x}|\mathbf{c}) = \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} p(\boldsymbol{x}_l|k, c_l). \qquad (4)$$

Equation (4) forms the basis of our model of style-consistency and can be applied to different kinds of feature distributions—discrete or continuous. In our implementation and experiments, we have used mixture distributions, mixtures of Gaussians in particular.

For any style k and pattern-class c, the pattern-feature probability is a mixture distribution.

$$p(\boldsymbol{x}|k,c) = \sum_{j=1}^{J} \pi_j(c,k) p(\boldsymbol{x}; \theta_j(c,k)), \qquad (5)$$

$$0 \le \pi_j(c,k) \le 1, \quad \sum_{j=1}^{J} \pi_j(c,k) = 1 \quad \forall c,k.$$

J is the number of mixture components (*variants*) in the distribution for each class and style. This number does not have to be the same for every class and style, but it makes our notation simpler. $p(\boldsymbol{x}; \theta_j(c,k))$ is the pattern-feature probability density conditioned on class c, style k, and variant j with parameters $\theta_j(c,k)$. The mixing parameters are $\pi_j(c,k)$.

In our experiments, we use the three models as explained below.

Style-bound variant (SB) model:

$$p(\boldsymbol{x}|k,c) = \sum_{j=1}^{J} \pi_j(c,k) p(\boldsymbol{x}; \theta_j(c,k)). \qquad (6)$$

There are $J \times K$ variant distributions per class, each with a different parameter set, $\theta_j(c,k)$, and weight, $\pi_j(c,k)$. Each variant distribution is bound to a style and different styles do not share parameters.

Style-shared variant (SS) model:

$$p(\boldsymbol{x}|k,c) = \sum_{j=1}^{J} \pi_j(c,k) p(\boldsymbol{x}; \theta_j(c)), \qquad (7)$$

$$\theta_j(c,k) = \theta_j(c) \text{ and } p(\boldsymbol{x}; \theta_j(c,k)) = p(\boldsymbol{x}; \theta_j(c)).$$

Here, the variant distributions (and their parameters) do not depend on the style. The parameters of the variant distributions for each class are "tied" across styles and the same set of variants are weighted differently to obtain the distributions for K styles.

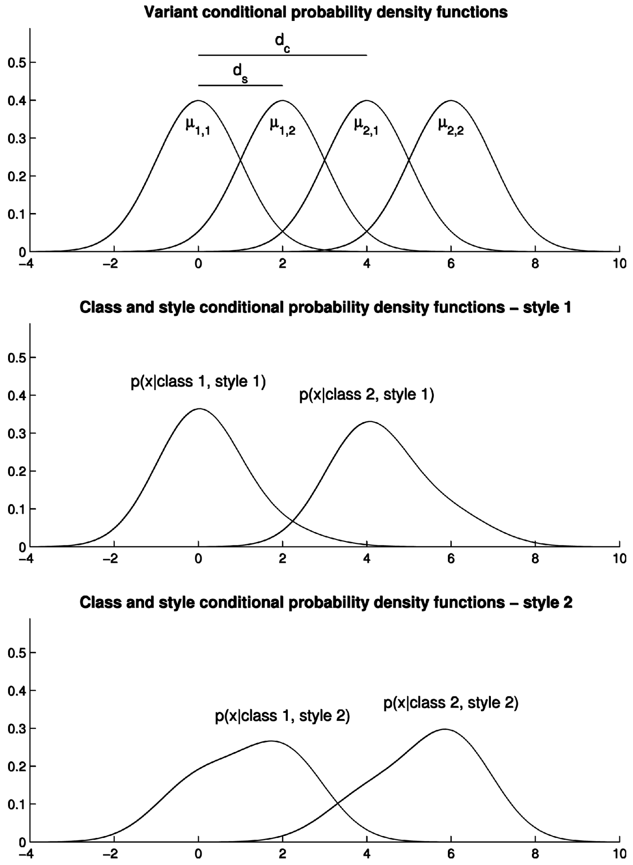


Fig. 4. Style-shared variants model. In each class, the same variant distributions are weighted differently to model non-Gaussian class-style conditional distributions.

This model is suitable for heavily overlapping feature distributions conditioned on the same class but different styles. Fig. 4 illustrates such an example. In a style-bound variants model, each variant distribution would belong exclusively to one style. In a shared variants model, overlapping non-Gaussian class-style conditional distributions are modeled as mixture Gaussians with different mixing parameters for each style and class. The style separation, d_s, and class separation, d_c, are introduced in Section 3.

Singlet (SN) model: Singlet modeling is the same as modeling with only one style, with the appropriate number of variants per class.

We now present the expanded equations for the three models.

$$\text{SB}: \quad p(\mathbf{x}|\mathbf{c}) = \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} \sum_{j=1}^{J} \pi_j(c_l,k)\ \ p(\boldsymbol{x}_l; \theta_j(c_l,k)) \qquad (8)$$

$$\text{SS}: \quad p(\mathbf{x}|\mathbf{c}) = \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} \sum_{j=1}^{J} \pi_j(c_l,k)\ \ p(\boldsymbol{x}_l; \theta_j(c_l)) \qquad (9)$$

$$\text{SN}: \quad p(\mathbf{x}|\mathbf{c}) = \prod_{l=1}^{L} \sum_{j=1}^{J} \pi_j(c_l)\ \ p(\boldsymbol{x}_l; \theta_j(c_l)). \qquad (10)$$

[2] For simplicity of notation, we have omitted the random variables in probability terms throughout. This should not perpetuate ambiguity since we use different notations for the "values" of the random variables. Thus, $P[c]$ is the probability of class c, while $P[k]$ is the probability of style k.

TABLE 1
Comparison of Number of Parameters in Style-Bound Variants, Style-Shared Variants, and Singlet Models with the Same Number of Variants per Class

	SB(K,J)	SS(K,KJ)	SN(KJ)
Number of variants per class	KJ	KJ	KJ
Number of style probabilities α	$K-1$	$K-1$	0
Number of variant weights π	$K(J-1)$	$K(KJ-1)$	$KJ-1$
$K=2, J=1$			
Number of variants per class	2	2	2
Number of style probabilities α	1	1	0
Number of variant weights π	0	2	1
$K=2, J=2$			
Number of variants per class	4	4	4
Number of style probabilities α	1	1	0
Number of variant weights π	2	6	3

We shall use the abbreviation $\mathrm{SB}(K,J)$ to denote a model with K styles and J style-bound variants per class per style ($K \times J$ variants per class). Let $\mathrm{SS}(K,J)$ denote a style-shared variants model with K styles and J variants per class and $\mathrm{SN}(J)$ denote a singlet mixture model with J variants per class.

Since different models have different numbers of parameters, they are not directly comparable. In practice, since the variant distributions account for most of the parameters, we compare different models that have the same number of variant distributions. Table 1 shows that the style-bound variants model is more economical in terms of parameters than the style-shared variants model when both use the same number of variants, KJ, per class. However, the style-shared variants model can cover a wider range of probability distributions. In particular, an $\mathrm{SS}(K,KJ)$ model can reduce to any $\mathrm{SB}(K,J)$ model as well as any $\mathrm{SN}(KJ)$ model with an appropriate setting of parameter values. Consequently, classification performance with the best $\mathrm{SS}(K,KJ)$ model should be at par or better than that with the best of all $\mathrm{SB}(K,J)$ and $\mathrm{SN}(KJ)$ models.

Maximum likelihood (ML) classification with the singlet model (2) selects the field-class that maximizes the objective function:

$$f_{\mathrm{ML,SN}}(\mathbf{x},\mathbf{c}) = \prod_{l=1}^{L} p(\boldsymbol{x}_l|c_l).$$

This function is the product of L terms with no shared variables. Each term can therefore be maximized independently. The ML field-classifier function then becomes

$$\begin{aligned}\Psi_{\mathrm{ML,SN,field}}(\mathbf{x}) &= \arg\max_{(c_1,\ldots,c_L)} \prod_{l=1}^{L} p(\boldsymbol{x}_l|c_l)\\ &= (c_1^*,\ldots,c_L^*) \text{ where } c_l^* = \arg\max_{c_l} p(\boldsymbol{x}_l|c_l).\end{aligned} \tag{8}$$

The process is thus equivalent to ML classification of the patterns, one at a time, and juxtaposition of the assigned pattern-classes to obtain a field-class. This is why we call this model the singlet model.

2.1 Maximum-Likelihood Style Constrained Classifier

A maximum-likelihood style constrained classifier is obtained from a style consistency model for the field-feature probability.

$$\Psi_{\mathrm{ML,SB,LO}}(\mathbf{x}) = \arg\max_{(c_1,\ldots,c_L)} \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} \sum_{j=1}^{J} \pi_j(c_l,k)\;\; p(\boldsymbol{x}_l;\theta_j(c_l,k)) \tag{12}$$

$$\Psi_{\mathrm{ML,SS,LO}}(\mathbf{x}) = \arg\max_{(c_1,\ldots,c_L)} \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} \sum_{j=1}^{J} \pi_j(c_l,k)\;\; p(\boldsymbol{x}_l;\theta_j(c_l)). \tag{13}$$

We call the above classifiers *label only* (LO) or *top-label* classifiers since they identify the top (most probable) label of the field. This is to distinguish them from their suboptimal approximations which identify the top field-label and style, which we shall present in Section 2.2. Note that the ML classifiers are easily transformed to the respective MAP classifiers since the field-class probability provided by a linguistic model is assumed to be independent of the style of rendition of the field.

2.2 A Suboptimal Approximation for a Style Constrained Classifier

A field of patterns collectively can furnish more information regarding the style of rendition than a single pattern. The longer the field, the better the resolution between different styles and the less the chances of interstyle class confusions. However, the number of field labels and, hence, the computational cost of classification with label-only classifiers, grows exponentially with the length of the field.

For a long field, we can assume that in (12) or (13) the term corresponding to the true style of the field $k = k^*$ will dominate the outer summation which is over all styles. This leads to an approximation of the label-only classifier, where this outer summation is replaced by a *maximum*. This approximation is, of course, suboptimal, but it leads to a dramatic reduction in computation, because the maximization over field-labels $(c_1,\ldots,c_L)$ can now be promoted in order ahead of the maximization over styles, as well as the product over pattern indices.

$$\begin{aligned}&\max_{(c_1,\ldots,c_L)} \sum_{k=1}^{K} \alpha_k \prod_{l=1}^{L} \sum_{j=1}^{J} \pi_j(c_l,k)\;\; p(\boldsymbol{x}_l;\theta_j(c_l,k))\\ &\approxeq \max_{(c_1,\ldots,c_L)} \max_{k=1\ldots K} \alpha_k \prod_{l=1}^{L} \sum_{j=1}^{J} \pi_j(c_l,k)\;\; p(\boldsymbol{x}_l;\theta_j(c_l,k))\\ &= \max_{k=1\ldots K} \alpha_k \times \max_{(c_1,\ldots,c_L)} \prod_{l=1}^{L} \sum_{j=1}^{J} \pi_j(c_l,k)\;\; p(\boldsymbol{x}_l;\theta_j(c_l,k))\\ &= \max_{k=1\ldots K} \alpha_k \times \prod_{l=1}^{L} \max_{c_l} \sum_{j=1}^{J} \pi_j(c_l,k)\;\; p(\boldsymbol{x}_l;\theta_j(c_l,k)).\end{aligned} \tag{14}$$

The approximation is thus equivalent to running K style-specific singlet pattern-classifiers, and choosing the output of the one that yields maximum field-feature likelihood (weighted by the a priori style probability α_k). We call such

TABLE 2
Error Rates (%) for Different Relative Positions of Gaussians

	d_c						
d_s	0	σ	2σ	3σ	4σ	5σ	6σ
0	73.9	51.7	29.5	13.4	5.1	1.5	0.3
	73.9	51.7	29.5	13.4	5.1	1.5	0.3
σ	74.4	54.3	33.6	17.0	7.1	2.5	0.7
	74.7	54.7	34.5	18.2	7.7	2.8	0.7
2σ	74.7	55.9	38.0	22.0	10.7	4.4	1.6
	74.6	61.6	45.7	29.7	16.1	7.0	2.6
3σ	74.9	53.3	35.1	24.7	13.8	6.3	2.6
	74.6	60.0	50.5	44.1	29.1	15.9	6.9
4σ	75.6	52.1	31.4	19.1	17.1	9.1	3.7
	75.4	55.6	41.7	38.5	43.9	29.0	15.9

Each cell shows singlet error rates below LO error rates. Each estimate is based on 10,000 simulated fields of length 2.

a classifier a top *label-style* (LS) classifier because it picks out the maximum-likelihood combination of field-label and style. The resulting style-bound and style-shared variant maximum-likelihood classifiers are:

$$\Psi_{\mathrm{ML,SB,LS}}(\mathbf{x}) = (c_1^{k^*}, \ldots, c_L^{k^*}), \text{ where}$$
$$k^* = \arg\max_{k=1\ldots K} \alpha_k \times \prod_{l=1}^{L}\sum_{j=1}^{J} \pi_j(c_l^k, k)\;\; p(\boldsymbol{x}_l; \theta_j(c_l^k, k))$$
$$c_l^k = \arg\max_{c_l=1\ldots C}\sum_{j=1}^{J} \pi_j(c_l, k)\;\; p(\boldsymbol{x}_l; \theta_j(c_l, k)), \tag{15}$$

$$\Psi_{\mathrm{ML,SS,LS}}(\mathbf{x}) = (c_1^{k^*}, \ldots, c_L^{k^*}), \text{where}$$
$$k^* = \arg\max_{k=1\ldots K} \alpha_k \times \prod_{l=1}^{L}\sum_{j=1}^{J} \pi_j(c_l^k, k)\;\; p(\boldsymbol{x}_l; \theta_j(c_l^k))$$
$$c_l^k = \arg\max_{c_l=1\ldots C}\sum_{j=1}^{J} \pi_j(c_l, k)\;\; p(\boldsymbol{x}_l; \theta_j(c_l)). \tag{16}$$

3 Simulation Experiments

We present some simple examples of the style-bound-variant model with scalar (one-dimensional) pattern features, two classes, two styles, and one Gaussian variant per class per style. For a field of length L:

$$p(\mathbf{x}|\mathbf{c}) = \sum_{k=1}^{2} \alpha_k \prod_{l=1}^{L} p(\boldsymbol{x}_l | c_l; \mu_{c_l,k}), \tag{17}$$

where each density function $p()$ is Gaussian with unit variance and mean $\mu_{c_l,k}$. Each $\boldsymbol{x}_l$ is a scalar pattern-feature, although we continue to use the bold $\boldsymbol{x}$ notation for consistency. With equiprobable styles ($\alpha_1 = \alpha_2 = 0.5$), for a field of length $L = 2$, (17) expands to:

$$\begin{aligned} p(\mathbf{x}|\mathbf{c}) = & \frac{1}{2} p(\boldsymbol{x}_1|c_1, \mu_{c_1,1}) p(\boldsymbol{x}_2|c_2, \mu_{c_2,1}) + \\ & \frac{1}{2} p(\boldsymbol{x}_1|c_1, \mu_{c_1,2}) p(\boldsymbol{x}_2|c_2, \mu_{c_2,2}). \end{aligned} \tag{18}$$

TABLE 3
Error Rates (%) for Different Relative Positions of Gaussians (with Inversion)

	d_c						
d_s	0	σ	2σ	3σ	4σ	5σ	6σ
0	73.9	51.7	29.5	13.4	5.1	1.5	0.3
	73.9	51.7	29.5	13.4	5.1	1.5	0.3
σ	70.9	52.0	32.8	17.4	7.6	2.8	0.7
	74.3	52.0	32.8	17.4	7.6	2.8	0.7
2σ	63.3	49.9	39.9	27.0	15.1	6.8	2.6
	74.2	52.0	39.9	27.0	15.1	6.8	2.6
3σ	56.9	39.2	34.6	37.5	26.2	14.9	6.8
	75.5	59.5	47.5	37.5	26.2	14.9	6.8
4σ	53.2	31.0	23.2	28.5	37.4	26.2	14.9
	74.9	55.9	41.5	36.9	37.4	26.2	14.9

Each cell shows singlet error rates below LO error rates. Each estimate is based on 10,000 simulated fields of length 2.

Notice that the style consistency models avoid mixed-style terms that appears in a similar expansion of the singlet model:

$$\begin{aligned} p(\mathbf{x}|\mathbf{c}) = & \frac{1}{4} p(\boldsymbol{x}_1|c_1, \mu_{c_1,1}) p(\boldsymbol{x}_2|c_2, \mu_{c_2,1}) + \\ & \frac{1}{4} p(\boldsymbol{x}_1|c_1, \mu_{c_1,2}) p(\boldsymbol{x}_2|c_2, \mu_{c_2,2}) + \\ & \frac{1}{4} p(\boldsymbol{x}_1|c_1, \mu_{c_1,1}) p(\boldsymbol{x}_2|c_2, \mu_{c_2,2}) + \\ & \frac{1}{4} p(\boldsymbol{x}_1|c_1, \mu_{c_1,2}) p(\boldsymbol{x}_2|c_2, \mu_{c_2,1}). \end{aligned} \tag{19}$$

Our objective is to measure the gains of modeling style consistency for different configurations of the four variant distributions. If two styles are identical in feature distributions or if, within each style, the classes have identical distributions, then we expect no gains by modeling style consistency.

We parameterize the means of the distributions in the following way:

$$\mu_{1,1} = 0; \mu_{1,2} = d_s; \mu_{2,1} = d_c; \mu_{2,2} = d_c + d_s; \tag{20}$$

d_c denotes the within-style separation between classes, while d_s is the within-class separation between style-variants (Fig. 4). Tables 2 and 3 present results of classifying simulated fields of length 2.

When the interclass distance is 0, the two classes are impossible to tell apart (Table 2) and none of the classifiers do better than chance.[3] When interstyle distance is 0, modeling two styles is useless since the styles are the same and the LO classifier is at par with the singlet classifier. The gains due to the LO classifier are maximum when $d_c = d_s$. This causes distributions for (class 1, style 2) and (class 2, style 1) to be identical, resulting in cross-style class confusions. Such a condition is illustrated in Fig. 5, where the seven in the left-hand style is identical to the one in the right-hand style. Most of these confusions can be resolved with style information furnished by the other pattern in the field.

Table 3 shows the results of the same experiment but with the relative positions of $\mu_{2,1}$ and $\mu_{2,2}$ reversed.

[3] Classification by chance (random class assignment to patterns) would yield 75 percent field error for fields of length 2. The error rates reported here are estimated by simulation.

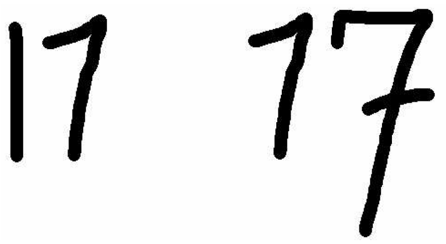


Fig. 5. A situation where style consistency alone can help resolve the dichotomy between "7" of the left-hand style and "1" of the right-hand style.

$$\mu_{1,1} = 0; \mu_{1,2} = d_s; \mu_{2,1} = d_c + d_s; \mu_{2,2} = d_c. \qquad (21)$$

Here, the gains are maximum when $d_c = 0$, which causes (class 1, style 2) to be identical to (class 2, style 1), and (class 1, style 1) to be identical to (class 2, style 2) in distribution. Consequently, the pattern-class conditional pattern feature distributions are identical for the two pattern classes. Therefore, the singlet classifier classifies by the "toss of a coin." The LO classifier does better only because it does not admit mixed style fields.

Fig. 6 plots the absolute gain (over singlet error-rate) and relative gain (absolute gain as percent of singlet error-rate) in error rate for LO and LS classifiers, with varying interclass distance, while the interstyle separation is fixed at 2σ (no inversion). The gain rate (absolute gain in error rate) is highest when d_c equals $d_s = 2\sigma$. The relative gain, however, keeps increasing with the interclass distance, indicating that a high fraction of singlet errors can be corrected by style constrained classification if the inherent separation between classes is good. In practice, the benefit will, of course, depend on how well we can model the tails of data distributions. The figure suggests that the LS classifier is a good approximation to the optimal LO classifier.

3.1 Performance on Same-Class and Mixed-Class Fields

For Gaussian variants with small interclass or interstyle separation, style constrained classifiers tend to perform worse on same-class fields than singlet classifiers. But, they gain more on mixed-class fields than they lose on same-class fields, thus improving the overall error-rate. When class separation is high, style constrained classifiers yield gains for both same-class and mixed-class fields.

For the two-class problem, with fields of length two, erroneous classification of same-class fields lowers the overall gain due to style constrained classification. Both more classes, and longer fields, result in diminishing probabilities of same-class fields and, therefore, favor style constrained classification.

3.2 Decision Boundaries

Some of the differences between different classifiers can be visualized by observing the differences in classification boundaries that they induce. In Fig. 7, we present the decision boundaries of different classifiers for the same simple model as in the previous simulations. The pattern-feature is scalar and the field length is 2. Therefore, the field-feature is bivariate, allowing us to plot field-feature distributions and the classification boundaries obtained by simulation. We show the decision boundaries only for the LO and LS classifiers. A singlet classifier always classifies each pattern independently of the other and produces decision boundaries that are parallel to the axes. In all five examples, the singlet decision boundaries are identical, splitting the illustrated feature space into four equal quadrants.

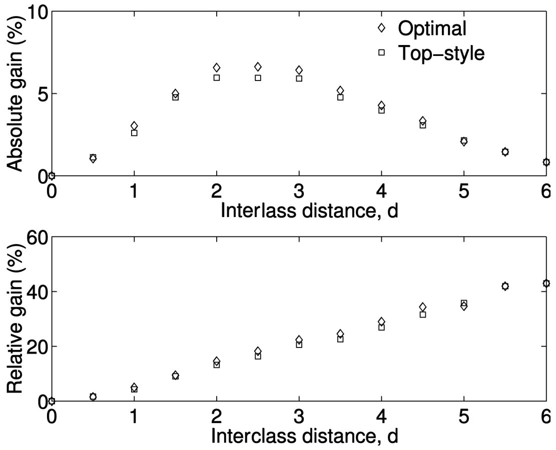


Fig. 6. Absolute and relative improvements in field error-rate as functions of interclass distance d_c, for fields of length 2.

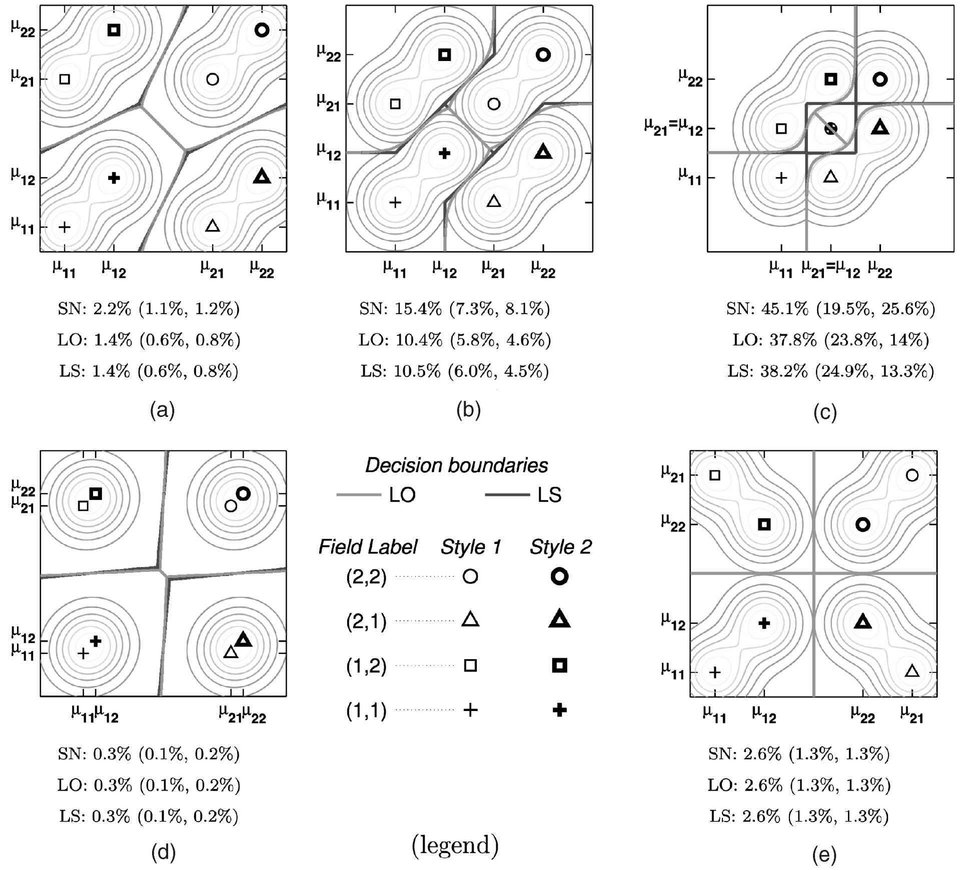
In each subfigure the locations of the means of the class-style conditional Gaussian distributions are marked along the axes. The locations of style and field-class conditional means are plotted (see the legend) for each of the four field-labels and two styles. Within each optimal decision region, we also plot the iso-probability contours for the probability density conditioned on the corresponding field label. These contours illustrate how mixtures of independent distributions model the statistical dependence in field-feature vectors. The numbers in percentage under each subfigure indicate the aggregate error rate and, within parentheses, the breakdown for same-class fields and mixed-class fields, respectively. The singlet (SN) error rates are the benchmarks for evaluating our classifiers.

When classes and styles are well separated, as in Fig. 7a, the LS approximation closely resembles the LO classifier, as we would expect. Both classifiers are quite different from what we would expect from a singlet model. As the interclass distance shrinks, as in Figs. 7b and 7c, differences between the two emerge. In Fig. 7c, the distributions for (class 1, style 2) and (class 2, style 1) are identical. When styles are not very pronounced because the style-separation is small, the LO and LS boundaries approach singlet boundaries, as in Fig. 7d.

When different classes have similar distributions within the same style, style constrained classification is not very effective. Fig. 7e corresponds to $d_c = 4$ and $d_s = 2$, but with inversion. Most singlet errors result from the confusion between the two classes within style 2. These errors cannot be corrected by style-consistency modeling and the LO, LS boundaries match the singlet boundaries—parallel and perpendicular to the axes.

3.3 Degree of Style Consistency

If there are no significant differences in style between fields of data, the singlet classifier can perform nearly as well as a style constrained classifier. To analyze the effect of the degree of style consistency, we generated style-consistent data according to the following distributions, where $G_{c,j}(\boldsymbol{x})$ is the unit-variance Gaussian density function with mean $\mu_{c,j}$:



SN: 2.2% (1.1%, 1.2%)
LO: 1.4% (0.6%, 0.8%)
LS: 1.4% (0.6%, 0.8%)
(a)

SN: 15.4% (7.3%, 8.1%)
LO: 10.4% (5.8%, 4.6%)
LS: 10.5% (6.0%, 4.5%)
(b)

SN: 45.1% (19.5%, 25.6%)
LO: 37.8% (23.8%, 14%)
LS: 38.2% (24.9%, 13.3%)
(c)

SN: 0.3% (0.1%, 0.2%)
LO: 0.3% (0.1%, 0.2%)
LS: 0.3% (0.1%, 0.2%)
(d)

(legend)

SN: 2.6% (1.3%, 1.3%)
LO: 2.6% (1.3%, 1.3%)
LS: 2.6% (1.3%, 1.3%)
(e)

Fig. 7. Classification boundaries (between field-classes) resulting from different classifiers, with different within-class style separation, d_s, and within-style class separation, d_c. (a) $d_s = 2$, $d_c = 6$. (b) $d_s = 2$, $d_c = 4$. (c) $d_s = 2$, $d_c = 2$. (d) $d_s = .5$, $d_c = 6$. (e) $d_s = 2$, $d_c = 6$ with inversion.

$$\begin{aligned}
p(\boldsymbol{x}|class1, style1) &= \pi G_{1,1}(\boldsymbol{x}) + (1-\pi)G_{1,2}(\boldsymbol{x})\\
p(\boldsymbol{x}|class1, style2) &= (1-\pi)G_{1,1}(\boldsymbol{x}) + \pi G_{1,2}(\boldsymbol{x})\\
p(\boldsymbol{x}|class2, style1) &= \pi G_{2,1}(\boldsymbol{x}) + (1-\pi)G_{2,2}(\boldsymbol{x})\\
p(\boldsymbol{x}|class2, style2) &= (1-\pi)G_{2,1}(\boldsymbol{x}) + \pi G_{2,2}(\boldsymbol{x}).
\end{aligned}$$

The two styles were set to be equiprobable ($\alpha_1 = \alpha_2 = 0.5$). The above corresponds to the SS($K = 2, J = 2$) model. The variant means were assigned as follows:

$$\mu_{1,1} = -4 \quad \mu_{1,2} = -2 \quad \mu_{2,1} = 2 \quad \mu_{2,2} = 4.$$

When the parameter π is set to 0, the two styles are distinct because only variants $G_{1,1}$ and $G_{2,1}$ occur together in fields of style 1, while only variants $G_{1,2}$ and $G_{2,2}$ occur together in fields of style 2. When $\pi = 0.5$, however, the two styles become identical. The four variants of two classes can randomly intermix with each other in a field. As π goes from 0 to 0.5, we have a continuum from two styles to a single style, the latter having an equivalent singlet model.

In Table 4, we show the effect of style-consistent modeling on classifier performance along this continuum. We list the field-error rates (field length = 2) of the singlet classifier SN($J = 2$) with known parameters, the style constrained classifier SS($K = 2, J = 2$)-LO with known parameters, and the style constrained classifier SB($K = 2, J = 1$)-LO with parameters estimated by Expectation-Maximization. Note that the SB($K = 2, J = 1$) model has only one (Gaussian) variant per class per style and is therefore appropriate only when $\pi = 0$ and is otherwise handicapped in comparison to the style-shared-variant model. The class-pattern-conditional marginal distributions remain the same for all values of π and, therefore, the singlet error rate remains the same. Variations are due only to estimation error.[4]

TABLE 4
Style Constrained Classification along a Continuum from Two Styles to One Style

π	Error rates (%)		
	SN	SS-LO	SB-LO†
0.00	2.3	1.3	1.3
0.05	2.3	1.7	1.8
0.10	2.2	2.0	2.0
0.15	2.3	2.1	2.2
0.20	2.2	2.3	2.2
0.30	2.3	2.2	2.3
0.40	2.3	2.3	2.3
0.50	2.3	2.3	2.3

† *Computed with the best EM estimate for $SB(K = 2, J = 1)$ model.*

[4] In all simulation experiments, error rates have been estimated by classifying pseudorandomly generated data. Though all experiments are run on large data sets, error rates deviate slightly from expected values or from other seemingly identical/equivalent experiments. These should be viewed as estimation errors, rather than anomalies.

TABLE 5
Variation of Field-Error Rates with Field Length

Field-length	Field-error rates (%)		
(L)	Singlet	LO	LS
1	1.1	1.1	1.1
2	2.3	1.4	1.4
3	3.5	1.3	1.3
4	4.5	1.2	1.2
5	5.5	1.2	1.2

3.4 Field Length

Applications such as OCR and speech recognition require classification of longer fields. This is an advantage because longer fields help to resolve styles and, therefore, favor style constrained classification. In Table 5, we compare the field-error rates of singlet and style constrained classifiers for different field lengths for Gaussian distributions. The distribution used is $\mathrm{SB}(K = 2, J = 1)$, with d_c set to 6 and d_s set to 2, according to (20).

Within-style confusions cannot be corrected by style-consistency modeling. Since in such cases, feature distributions are similar for two or more classes within the same style, one has to look for other forms of context (e.g., linguistic context) to resolve such confusions.

3.5 Comparison of LO Classifier and the LS Approximation

The LS classifier is a suboptimal approximation to the optimal LO field-classifier, as presented in Section 2.2. With large separation between styles, it is a good approximation and yields almost the same gains as LO classification. Even otherwise, information regarding style can be augmented by longer fields. In Table 6, we show how the LS classifier approaches the LO error rate with increasing field length.

The experiments were run with our $\mathrm{SB}(K = 2, J = 1)$ model with unit-variance Gaussians. The means are placed according to (20) with $d_s = 2$ and $d_c = 4$. Each error rate was estimated by classifying 100,000 fields.

The processing time for both singlet and LS classifiers grows linearly with the number of patterns in the field. The runtime of the LO classifier increases exponentially over the fixed overhead (for reading data, parameters, and writing output) because, for each field, it explores all 2^L possible field-classes to pick one. A smart search through this exponential search space is reported in [17].

TABLE 6
Performance Comparison of LO and LS Classifiers

d_c	Field-length (L)	Field error rates (%)		
		SN	LO	LS
4	1	7.67	7.67	7.67
	2	15.88	10.86	10.91
	3	22.57	11.77	11.85
	4	28.56	12.63	12.70
	5	34.03	13.61	13.63
	6	39.05	14.90	14.94
	7	43.93	16.41	16.43

3.6 Modeling Broad Styles

In our style-consistency model, we assume that there is statistical dependence between co-occurring variants of different classes, i.e., the variant of class "A" appearing in a field depends on which variant of class "B" appears in the same field and vice versa. We call this a *strong style consistency* assumption. If, in the aggregate of many styles, most or all combinations of different variants of classes occur, then the benefits of strong style consistency modeling are less. A classifier can still take advantage of consistent rendering, within a field, of samples of the same class (*weak style consistency* assumption). This can be achieved by adapting the classifier to the specific style of the test field, provided the field is long enough [3], [28], [2], [25], [18].

In the presence of many styles, it may be impractical to model each style because of requirements on the training sample size and because of computational complexity. On the other hand, if the field-feature distributions show broad clusters of styles, style-modeling can prove useful. For example, broad styles may be induced by gender (pitch) for speech, nationality (training) of writers, or type-style of fonts.

Increasing the number of styles (parameters) often leads to progressively better approximation of style-consistent distributions. We report an experiment where data was generated according to the $\mathrm{SB}(K = 10, J = 1)$ model with unit variance Gaussian variants.

The means of the Gaussians were $\mu_{c,k} = (c - 1)d_c + k$ for class $c = 1, 2$ and style $k = 1 \dots 10$. Thus, for $d_c = 10$, the means were at $1, 2, \dots 10$ for class 1 and, correspondingly, at $11, 12, \dots 20$ for class 2. Table 7 reports the results of approximating the distribution with our EM estimates for 1, 2, 3, 4, 5, and 10 styles. Label-only and singlet classifiers with an equal number of Gaussian variants per class are compared

TABLE 7
Field Error Rates (in Percent) for Fields of Length 2, when a Style-Consistent Distribution Is Approximated with Fewer Styles

K	$d_c = 10$		$d_c = 8$		$d_c = 6$		$d_c = 4$	
	Styles	Singlet	Styles	Singlet	Styles	Singlet	Styles	Singlet
1	7.1	7.1	18.5	18.5	33.5	33.5	47.6	47.6
2	1.2	7.1	5.4	18.6	14.5	33.4	32.2	51.4
3	1.2	6.9	5.3	18.6	12.3	33.5	25.3	47.8
4	1.2	7.1	5.3	18.5	12.0	33.5	23.9	47.8
5	1.2	7.0	5.5	18.6	12.0	33.5	23.5	48.0
10	1.2	7.0	5.2	18.6	11.9	33.5	23.5	48.0

An underlying distribution with 10 styles and one variant per class per style is approximated 1) by style-consistent distributions with K styles (one variant per class per style) and 2) by K variants-per-class singlet distributions.

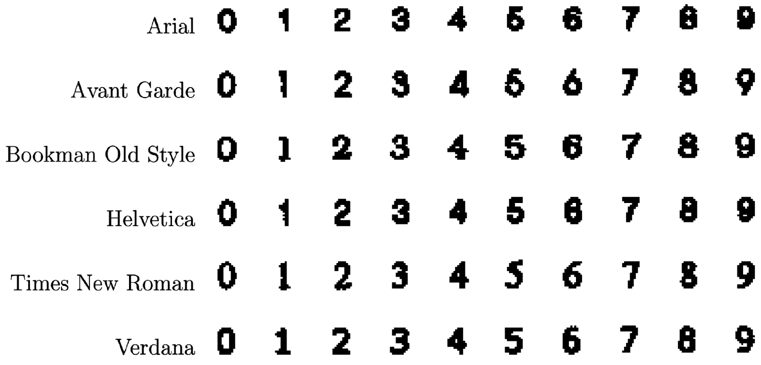


Fig. 8. Examples of bitmaps of the 10 digits from six different fonts.

by the respective field error rates (in percent) on classifying fields of two patterns. The results indicate diminishing returns with increasing K.

4 Experiments on Machine Printed Digits

We performed controlled laboratory experiments to demonstrate the application of style-consistency models in classifying machine-printed digits.

4.1 Data

The 10 digit classes "0" through "9" were printed in six different 6 point fonts on a 600 dpi laser printer and bilevel image samples were obtained by scanning at 200 dpi. The printer toner and scanner settings were unchanged during sample generation so that it was reasonable to presume that the major stylistic differences were due to fonts. Examples of enlarged digit images in the six fonts are shown in Fig. 8. We obtained 500 images per font per digit class or 30,000 images in all.

Four central moments—M_{00}, M_{20}, M_{02}, M_{11}—were computed as features for each digit image.

$$M_{mn} \triangleq \sum_{x=1}^{W} \sum_{y=1}^{H} b(x,y)(x-x_0)^m (y-y_0)^n. \qquad (22)$$

$b(x,y)$ is the value of the pixel at column x, row y of the digit bitmap, whose width and height are W and H pixels, respectively. The "value" of a pixel is considered 1 if the pixel is black (foreground), 0 otherwise (background). (x_0, y_0) is the centroid of the bitmap.

4.2 Experimental Design

The printed digit data was divided into two equal samples (250 patterns per digit class per font). The first sample was used for training and the second for testing in each of the experiments.

Font-specific training. The training sample was subdivided according to font and 250 feature vectors per font per class were used to estimate the parameters of font and class conditional distributions. Each such distribution was modeled as Gaussian with a diagonal covariance matrix.

Multifont training and testing. For the purpose of these experiments, we used known font information only to construct isofont training and test fields. Font-labels of these fields were discarded. Singlet models with G Gaussians per class, SN($J = G$), were compared to style-models, SB($K = G, J = 1$), that had the same number of Gaussians per class. The Expectation-Maximization (EM) algorithm was used for style-unsupervised training. The details of the algorithms are explained in [16].

Training fields. Training fields of length 13 were formed with samples from the same font. The field classes of training fields were chosen by cycling through the 10 classes. Thus, training field-classes were 9876543210987, 6543210987654, etc. Only 14,430 of the 15,000 patterns in the training sample were actually used for training (due to the method used to scramble digit patterns to form fields). The choice of 13 as the field-length guaranteed the co-occurrence of every class in each field. Further, since field length (13) and the number of classes (10) are relatively prime, the cycle of field-class labels was long. The object of this scheme was to take advantage of the large number of samples per font (style) and test the estimation algorithm with an assortment of field-classes.

Test fields. The test patterns from each font were permuted randomly and then sliced into fields of length 2 or 4. All 15,000 digits in the test sample were used for testing, classification being performed twice (7,500 fields of length 2 or 3,750 fields of length 4).

Although we have used fixed length fields for training and testing for simplicity, our model and algorithms for classification and estimation apply directly to fields of varying length.

4.3 Results

Table 8 presents a summary of the results of our experiments. All error rates refer to pattern-error rates rather than field-error rates, i.e., we count the number of digits, not digit-fields that are mislabeled. Each monofont classifier (1 Gaussian per class) was applied to all the test data, and the best of them yielded 43.5 percent errors. When the single Gaussian per class was trained on samples of all fonts, the error rate dropped to 34.2 percent—an effect

TABLE 8
Error Rates (%) of Different Classifiers for Printed Digit Data

Best of 6 monofont classifiers	43.5
Multifont singlet classifier (1 Gaussian per class)	34.2
Multifont singlet classifier (6 Gaussians per class)	19.8
Multifont style conscious LO classifier (fields of length 2)	16.5
Multifont style conscious LO classifier (fields of length 4)	14.9
Font-specific classifier (1 Gaussian per class per font)	14.2

called *generalization*. On allowing six Gaussians per class to model the six fonts in a singlet classifier, the error rate decreased further to 19.8 percent.

If, instead of ignoring the font label of each test field, we apply the appropriate monofont classifier, we achieve an error rate of 14.2 percent. This represents a lower bound on the error rate achievable with our features and classifiers. With a style constrained classifier, where we do not assume knowledge of the parent font, we expect to achieve a performance better than singlet classification, but worse than the above font-specific classifiers. This is confirmed in our experiments as we obtain error rates of 16.5 percent and 14.9 percent for style conscious classification of fields of length 2 and 4, respectively. Longer fields favor style constrained classification because more patterns in a field yield stronger evidence of the true parent style of the field.

Note that the high error rates (in absolute terms) can be attributed to the simplistic feature set used for classification (the first four central moments). Nevertheless, the experiments demonstrate the concepts and utility of style constrained classification. While we concentrate on classification accuracy, we would like to point out that modeling style consistency leads to a better description of the data in general that may be useful for other tasks, such as compression. In support of that we report, in Table 9, the overall data-likelihood at the end of EM iterations. Both the singlet and style-bound variant models compared in the table use exactly the same number of parameters in each case. The style-bound variants model always achieves better cumulative data-likelihood, indicating a better fit to data than the singlets model. For a more formal analysis of model selection, let us consider the case for $G = 6$. A hypothesis that style modeling adds unnecessary complexity can be rejected by examining the log likelihood-ratio of $(1.32 - 1.20) \times 10^5 = 0.12 \times 10^5$ on the basis of either the generalized likelihood ratio test (GLRT) ([1, p. 229]) or Bayes Information Criterion (BIC) [21]. The likelihood ratio can also be used as a measure of style consistency in data.[5]

TABLE 9
Log-Likelihood of Training Data under Different Models

	Data log-likelihood ($\times 10^5$)	
	G styles, 1 Gaussian per class-style	Singlet, G Gaussians per class
G	SB(K=G,J=1)	SN(J=G)
1	-1.56	-1.56
2	-1.48	-1.45
3	-1.36	-1.39
4	-1.29	-1.36
5	-1.20	-1.33
6	-1.20	-1.32

5 Summary and Conclusion

When patterns are isogenous, the common origin leaves its style imprint on the patterns, resulting in style consistency. An appropriate style specific classifier yields a higher classification accuracy than an omnifont classifier. Even when the parent style of a field of patterns is not known, the knowledge of style consistency in the field can be used to improve classification accuracy over singlet classification. We present a system of notations, a hierarchical mixture model of style consistency, and develop the formulae for field classification.

We model style consistency expressly on the basis of pattern co-occurrence in the space of classifier features. Since no special style-indicator features have to be designed, we can estimate style models directly from pattern co-occurrence data. This is especially useful for styles that are difficult to enumerate, as in hand-print or image degradation. Also, distinct styles are relevant only if they induce significant differences in the distributions of the features that are actually used in classifying the patterns.

We model strong style consistency, allowing us to improve classification accuracy of short fields by enforcing consistency across classes of patterns; adaptation methods presented in pattern recognition literature require long fields to adapt to the parent style.

Through experiments and simulations we demonstrate several properties of style constrained classification. Our probabilistic model of style consistency reduces the error rate in machine-printed digit classification. Longer fields favor style constrained classification because they furnish more information about the parent style. When errors are dominated by within-style confusions, style constrained classification yields less improvement. Though we focus on classification, applications such as compression may benefit from style consistency modeling because it enables better statistical characterization of data (higher cumulative data likelihood), while using the same number of parameters as a singlet model.

The machine-printed digit recognition experiments presented in this paper were designed to be simple (only four moments used as pattern features) and carefully controlled (font was the source of stylistic differences) to highlight the salient features of style constrained classification and our model of style consistency. The error rates do not reflect the state-of-the-art in isolated digit recognition; they help us illustrate our points. Similar experiments on hand-printed digits (with style-unsupervised training) yielded a 25 percent reduction of errors [16], but, because of the lack of style labels, we could not compare the error rate of the style constrained classifier with that of a style-specific classifier.

A practical problem associated with optimal (label-only) classification is that computational complexity grows exponentially with field length. Label style (LS) classification is a suboptimal approximation that has worked quite well in our experiments. This approximation provides a bridge between optimal style constrained classification (good for short fields) and adaptation to the parent style (good for long fields). Sarkar [17] presents an algorithm that enables optimal classification with exponential worst-case complexity but excellent empirical average complexity.

"Ideal" features are style insensitive but discriminate among classes. Style indicator features would be at the other extreme, namely, class insensitive. In practice, features are between these extremes: Adding more features can improve recognition accuracy so that further improvement with

[5] The style-bound variants model does not admit the singlet model as a special case - a requirement for the GLRT as stated in Bickel and Doksum. However, the class of shared-variants models SS(6,6) (6 styles, 6 Gaussians per class) admits both the style bound and singlet models as special cases. Such a model of style, even after penalizing it for 30 extra parameters that are trivially set to match the style-bound model, passes, with high confidence, both the GLRT and BIC tests for selection over a singlet model.

style-conscious classification is less. More features do not help with our pathological example of ones and sevens where style is the only discriminator. In applications where training (and test) data come in isogenous groups, modeling style consistency can help us exploit co-occurrence information that would otherwise be wasted.

Acknowledgments

The authors would like to thank Dr. H. Fujisawa and his group at the Hitachi Central Research Laboratories, Japan, for inspiring and supporting this project. This work has been conducted in the New York State Center for Automation Technologies (CAT), which is partially supported by a block grant from the New York State Science and Technology Foundation. We are grateful for the comments and support of Henry Baird and of other colleagues at PARC.

References

[1] P.J. Bickel and K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics.* Englewood Cliffs, N.J.: Prentice Hall 1977.

[2] T. Breuel and C. Mathis, "Classification Using a Hierarchical Bayesian Approach," *Proc. 16th Int'l Conf. Pattern Recognition,* pp. 40103-40106, Aug. 2002.

[3] H.S. Baird and G. Nagy, "A Self-Correcting 100-Font Classifier," *Document Recognition, Proc. SPIE,* L. Vincent and T. Pavlidis, eds., vol. 2181, pp. 106-115, 1994.

[4] I. Bazzi, R. Schwartz, and J. Makhoul, "An Omnifont Open-Vocabulary OCR System for English and Arabic," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 6, pp. 495-504, June 1999.

[5] V. Bouletreau, N. Vincent, R. Sabourin, and H. Emptoz, "Synthetic Parameters for Handwriting Classification," *Proc. Fourth Int'l Conf. Document Analysis and Recognition,* vol. 1, pp. 102-106, 1997.

[6] R.G. Casey, "Text OCR by Solving a Cryptogram," *Proc. Eighth Int'l Conf. Pattern Recognition,* pp. 349-351, 1986.

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

[8] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification.* New York: John Wiley and Sons, 2001.

[9] A.P. Dempster, M.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc.,* vol. 39, no. 1, pp. 1-38, 1977.

[10] T.K. Ho, J.J. Hull, and S.N. Srihari, "Word Recognition with Multi-Level Contextual Knowledge," *Proc. First Int'l Conf. Document Analysis and Recognition,* pp. 905-915, Oct. 1991.

[11] J.J. Hull and S.N. Srihari, "Experiments in Text Recognition with Binary N-Gram and Viterbi Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 4, no. 5, pp. 520-530, Sept. 1982.

[12] F. Jelinek, *Statistical Methods in Speech Recognition.* Cambridge, Mass.: MIT Press, 1997.

[13] G. Nagy, "Teaching a Computer to Read," *Proc. 11th Int'l Conf. Pattern Recognition,* vol. 2, pp. 225-229, Sept. 1992.

[14] K.A. Nathan, J.R. Bellegarda, D. Nahamoo, and E.J. Bellegarda, "On-Line Handwriting Recognition Using Continuous Parameter Hidden Markov Models," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 5, pp. 121-124, 1993.

[15] R. Plamondon, D.P. Lopresti, L.R.B. Schomaker, and R. Srihari, "On-Line Handwriting Recognition," *Wiley Encyclopedia of Electrical and Electronics Eng.,* J.G. Webster, ed. pp. 123-146, New York: John Wiley & Sons, 1999.

[16] P. Sarkar, "Style Consistency in Pattern Fields," PhD thesis, Rensselaer Polytechnic Inst., 2000.

[17] P. Sarkar, "An Iterative Algorithm for Optimal Style-Conscious Field Classification," *Proc. 16th Int'l Conf. Pattern Recognition,* vol. IV, pp. 243-246, Aug. 2002.

[18] P. Sarkar, H.S. Baird, and X. Zhang, "Training on Severely Degraded Text-Line Images," *Proc. Seventh Int'l Conf. Document Analysis and Recognition,* pp. 38-43, Aug. 2003.

[19] P. Sarkar and G. Nagy, "Classification of Style-Constrained Pattern-Fields," *Proc. 15th Int'l Conf. Pattern Recognition,* pp. 859-862, 2000.

[20] P. Sarkar and G. Nagy, "Style Consistency in Isogenous Patterns," *Proc. Sixth Int'l Conf. Document Analysis and Recognition,* pp. 1169-1174, Sept. 2001.

[21] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics,* vol. 6, no. 2, pp. 461-464, 1978.

[22] R.M.K. Sinha and B. Prasada, "Visual Text Recognition through Contextual Processing," *Pattern Recognition,* vol. 20, no. 5, pp. 463-479, 1988.

[23] J.B. Tenenbaum and W.T. Freeman, "Separating Style and Content," *Advances in Neural Information Processing Systems 9,* M. C. Mozer, M.I. Jordan, and T. Petsche, eds., San Mateo, Calif.: Morgan Kaufmann, 1997.

[24] S. Veeramachaneni and G. Nagy, "Style Context with Second-Order Statistics," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 1, Jan. 2005.

[25] S. Veeramachaneni and G. Nagy, "Adaptive Classifiers for Multisource OCR," *Int'l J. Document Analysis and Recognition,* vol. 6, no. 3, pp. 154-166, Aug. 2004.

[26] S. Veeramachaneni and G. Nagy, "Style-Conscious Quadratic Classifier," *Proc. 16th Int'l Conf. Pattern Recognition,* vol. II, pp. 72-75, Aug. 2002.

[27] S. Veeramachaneni, G. Nagy, C.-L. Liu, and H. Fujisawa, "Classifying Isogenous Fields," *Proc. Eighth Int'l Workshop Frontiers of Handwriting Recognition,* pp. 41-46, Aug. 2002.

[28] Y. Xu and G. Nagy, "Prototype Extraction and Adaptive OCR," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 12, pp. 1280-1296, Dec. 1999.

[29] A. Zramdini and R. Ingold, "Optical Font Recognition Using Typographical Features," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 877-882, Aug. 1998.

Prateek Sarkar received the Bachelor of Technology from the Indian Institute of Technology, Kharagpur (http://www.iitkgp.ernet.in, 1993), and the MS and PhD degrees in computer and systems engineering from Rensselaer (http://www.rpi.edu, 1994, 2000). He has interned at Panasonic Laboratories, Princeton, and in the Human Language Technologies division of the IBM T.J. Watson Research Center, where he helped build speech recognition engines for automobiles. He is a researcher at the Palo Alto Research Center (PARC) where he works on statistical pattern recognition, document image analysis, and machine learning. His research includes modeling spatial sampling effects in images, statistical models for quality control in large-scale scan conversion, unsupervised estimation and exploitation of style-consistency in pattern recognition, and statistical models for perceptual organization. He is a member of the IEEE Computer Society.

George Nagy received the BEng and MEng degrees from McGill University and the PhD degree in electrical engineering from Cornell University in 1962 (on neural networks). For the next 10 years, he studied pattern recognition at the IBM T.J. Watson Research Center in Yorktown Heights. From 1972 to 1985, he was a professor of computer science at the University of Nebraska-Lincoln (nine years as chair), and worked on geographic information systems, remote sensing applications, and human-computer computer interfaces. Since 1985, he has been a professor of computer engineering at Rensselaer Polytechnic Institute, where he established the ECSE DocLab. In addition to document image analysis, OCR, geographic information systems, and computational geometry, his students have engaged in solid modeling, finite-precision spatial computation, and interactive computer vision, often with a focus on systems that improve with use. He has benefited from visiting appointments at the Stanford Research Institute, Cornell, the University of Montreal, the National Scientific Research Institute of Quebec, the University of Genoa and the Italian National Research Council in Naples and Genoa, AT&T and Lucent Bell Laboratories, IBM Almaden, McGill University, the Institute for Information Science Research at the University of Nevada, and the Center for Image Analysis in Uppsala. He is a life fellow of the IEEE and the IEEE Computer Society.