Style Quantification of Scanned Multi-source Digits

Xiaoli Zhang and George Nagy
DocLab, ECSE, Rensselaer Polytechnic Institute
zhangxl@rpi.edu, nagy@ecse.rpi.edu

Abstract

The co-occurring patterns in a group carrying the traits of common origin are statistically dependent via an underlying style context. Exploiting style consistency in groups of patterns from multiple sources can increase OCR accuracy. The accuracy gains obtained by a style consistent classifier depend on the amount of style in isogenous (same-source) fields. We present mathematical models to quantify the amount of single-class and multi-class style using entropy, correlation and mutual information. We also demonstrate a method for style homogenization that allows testing our metrics on real data.

1. Introduction

We can often distinguish numerals written by Alice from numerals printed by Bob, as well as digits printed in different fonts. Forensic analysts can even tell whether two fields of digits were written with the same pen, or printed on the same printer. This shape consistency, called style, takes two forms, which we call single-class style and multi-class style.

Single-class style is the shape consistency of a single class from each source. It reveals how consistent a writer is in writing a glyph. Does Alice always cross her 7's, while Bob never does?

Multi-class style determines how much the shape of a given class reveals about the appearance of other classes from the same source. The way Alice writes 9 helps predict the way she will write 6.

In OCR, each glyph (a letter, numeral or ideograph) is usually represented by a feature vector. Style can then be characterized in terms of statistical measures collected on fields of data from different sources. We will present a measure for single-class style based on entropy, and two measures for multi-class style, one based on cross-correlation and the other on mutual information.

Table 1 illustrates the difference between the two kinds of style. In single-class style, the patterns of the same class from a given source differ only by noise. For example, the 2s from Source 1, Source 2, and Source 3 are always bold. Therefore Source 1, 2, and 3 have the same style with respect to numeral 2.

In multi-class style, if a source produces a bold 2, we can be sure that it will also produce a bold 1, and a source that favors italic 2 will also favor italic 1. These sources represent two distinct multi-class styles.

Table 1. Single-class and multi-class style

	Single-class style	Multi-class style
Source 1	**22***/07/19***27**	*22/07/1922*
Source 2	**2***5/***0***5/***19***0***5**	**05/05/1925**
Source 3	**2***1/***06/***1***943**	**21/06/1943**
Source 4	*03/***2***4/19***45**	*03/24/1945*

Either kind of dependence can be exploited in field classification. Single-class style is suitable for adaptation on long fields, while multi-class style can improve accuracy even on short, mixed-class fields [1.2]. In real data, especially hand-print, either kind of style may be difficult to detect. Our methods facilitate the cost-benefit analysis of resource-intensive field classification versus simpler singlet classification.

We first show how *style homogenization* can remove style from a dataset. We will then apply style homogenization to scanned printed and hand-printed digits, and show that (1) our measures do reveal the presence or absence of the two kinds of style, and (2) the error rate of a style-constrained field classifier is consistent with the proposed measures.

2. Style homogenization

The tables below illustrate the process. There are only two classes here, A and B, and four sources: the source of each pattern is indicated by a superscript. Each source has only 3 samples, numbered 1, 2 and 3, and indicated by a subscript. Thus B_2^3 is the second sample of class B from Source 3. The original data, Dataset (*DS*) I, before style homogenization, is indexed according to Table 2.

Table 2 - Dataset I

	Class A			Class B		
S1	A_1^1	A_2^1	A_3^1	B_1^1	B_2^1	B_3^1
S2	A_1^2	A_2^2	A_3^2	B_1^2	B_2^2	B_3^2
S3	A_1^3	A_2^3	A_3^3	B_1^3	B_2^3	B_3^3
S4	A_1^4	A_2^4	A_3^4	B_1^4	B_2^4	B_3^4

Style homogenization creates *virtual sources* by shuffling the samples. In DS II (Table 3), samples of both A and B are replaced by randomly selected samples from any source. Therefore DS II exhibits no style, regardless of the style content of DS I.

Table 3 - Dataset II

	Class A			Class B		
S1	A_2^2	A_1^3	A_1^4	B_1^2	B_3^3	B_2^4
S2	A_1^1	A_3^2	A_3^4	B_1^1	B_2^3	B_3^4
S3	A_2^1	A_1^2	A_2^4	B_3^1	B_2^2	B_1^4
S4	A_3^1	A_2^3	A_3^3	B_2^1	B_3^2	B_1^3

The statistical integrity of the data is preserved by using each sample only once. Our measures should show that the homogenized Dataset II has no single- or multi-class style except what might be expected from serendipitous configurations of a finite sample.

3. Style quantification

Our entropy measure for single-class style requires style labeled data. It compares the non-uniformity of the distribution of styles of a single class from each source to the average non-uniformity of this class in the entire dataset, using entropy as a measure of non-uniformity.

For multi-class style we propose two measures, mutual entropy and style correlation. Mutual information also requires style-labeled data. It is a non-parametric measure of the dependence between the style distributions of pairs of classes in same-source pairs of patterns.

Style correlation does not require style-labeled samples, but only pairs of samples from each source. It is therefore applicable to many pattern recognition tasks where the samples arrive in same-source fields (like addresses on postal envelopes, insurance claims, and tax forms), but where the style of these groups is not known, and many fields may share the same style.

3.1. Single-class style: entropy

We begin with N_c samples, and therefore N_c feature vectors of a single digit class c, from M sources. There are K styles, and each sample of class c has a style label. The style may be associated with a writer, a typeface, or with any other source-dependent grouping. Let there be N_m samples of this digit class from source m, N_k samples of style k, and $N_{m,k}$ samples from source m and class style k. The *Source Class-Style Probability Vector* (*SCSPV*) is

$$P_m = (p_{m,1}, p_{m,2}, \ldots, p_{m,K})$$

where $p_{m,k}=N_{m,k}/N_m$, for $m = 1,\ldots,M$, $k = 1,\ldots,K$.

We calculate the *Source Class Entropy* for each source:

$$H_m = - \sum_{k=1}^{K} p_{m,k} \log_2 p_{m,k}$$

If all the samples of a digit class for the source are assigned to the same style, then the source class entropy is its minimum value, zero. This source has maximal single-class style. In contrast, the source class entropy reaches maximum value of $\log_2 K$ when a source exhibits K equally probable variations in shaping the same digit. Such a source does not have any single-class style for the observed digit. The *Average Source Class Entropy* (*ASCE*) is defined as:

$$H_{average} = \frac{1}{M} \sum_{m=1}^{M} H_m$$

The values of H_m and $H_{average}$ depend on the fraction of samples from each style as well as on the amount of style.

With a finite number of samples, even if the sources did not exhibit any single-class style, sampling fluctuation may decrease the average entropy. To account for the finite sample size, we compute the *Expected Source Class Entropy* $E[H]$ under the multinomial sampling distribution [3].

$$P[n_1,\ldots,n_k,\ldots,n_K; p_1,\ldots,p_k,\ldots,p_K] = \frac{n!}{\prod_{k=1}^{K} n_k!} \prod_{k=1}^{K} p_k^{n_k}$$

where $p_k = N_k / N_c$, and $n = \sum_{k=1}^{K} n_k = N_c / M$ is the average number of samples from each source, and n_k is the number of samples of style k. We assume that $N_m = n$ for each source. We obtain the source class entropy of each partition by considering all ways of partitioning n samples into K styles:

$$H[n_1,\ldots,n_k,\ldots,n_K] = -\sum_{k=1}^{K} \frac{n_k}{n} \log_2 \frac{n_k}{n}$$

$E[H]$ is then derived by summing the product of the multinomial probability and the entropy for every possible source class-style partition. $E[H]$ is given by

$$\sum_{n_1=0}^{n} \cdots \sum_{n_k=0}^{n-n_1-\ldots-n_{K-1}} P[n_1,\ldots,n_k,\ldots,n_K; p_1,\ldots,p_k,\ldots,p_K] H[n_1,\ldots,n_k,\ldots,n_K]$$

E[H] approaches its normalized limiting value of unity as n increases. With Matlab we can compute E[H] up to n=170 with K=3.

The expected entropy predicts the average entropy when there is no style. We can therefore compare $H_{average}$ with $E[H]$ as a measure of single-class consistency. A large difference indicates strong single-class style.

3.2. Multi-class style: mutual information

Isogenous patterns exhibit inter-pattern style consistency. Multi-class style derives from the statistical dependence between the features of co-occurring patterns from different classes. We ignore throughout statistical dependence between the labels of adjacent patterns, known in character and speech recognition as language context. Mutual information (*MI*) quantifies multi-class style.

Each source *m* generates patterns of class *i* and style *k* according to $P_m^i = (p_{m,1}^i, \ldots, p_{m,k}^i, \ldots, p_{m,K}^i)$, and of class *j* according to P_m^j, as in Section 3.1. $P_m^i = (p_{m,1}^i, p_{m,2}^i, \ldots, p_{m,K}^i)$ is quantized to a 0/1 vector at thresholds set to $P^i = (p_1^i, p_2^i, \ldots, p_K^i)$. The resulting *Source Class-Style Assignments* are considered octal-valued random scalars G_i and G_j.

If, for instance, there are three equiprobable styles of classes 5 and 6 (i.e., $P^5 = P^6 = (1/3, 1/3, 1/3)$), and Source 13 has generated (5,4,1) and (0,8,2) samples of each style of classes 5 and 6 respectively, then the corresponding values for the 13th source sample will be G_5 = (1,1,0) = 6_8 and G_6 = (0,1,0) = 2_8. Since the total number of samples per source is fixed, some values of *G* (here 0_8 and 7_8) cannot occur.

The extent of multi-class style is determined by the departure from statistical independence of the joint sample distribution of G_i and G_j ($i \neq j$), which is quantified by *MI*

$$I(G_i, G_j) = \sum_{G_i, G_j} P(G_i, G_j) \log \frac{P(G_i, G_j)}{P(G_i) P(G_j)}$$

$P(G_i)$ is the fraction of sources with source class-style label G_i. $P(G_i, G_j)$ is the number of (G_i, G_j) combinations over all sources divided by the number of sources *M*. The number of possible combinations increases rapidly with finer quantization, resulting in too few samples for accurate estimates of the joint probabilities.

If there is strong multi-class style, that is, if the Source Class-Style Assignments of two classes are always and only associated with each other, then the difference between the values of *MI* before and after style homogenization will reach its maximum value.

3.3. Multi-class style: style correlation

Style correlation is based on the cross-covariance matrices which capture the class-conditional dependence between the singlet patterns in a field [4].

A singlet pattern is denoted by $\mathbf{x} \in \mathrm{R}^d$ with class label $\mathtt{c} \in \{c_1, c_2, \ldots, c_N\}$, where *N* is the number of singlet classes. An isogenous field, with *L* singlet patterns, is represented by $\mathbf{y} = [\mathbf{x}_1^T \ \mathbf{x}_2^T \ldots \mathbf{x}_L^T]^T$ and its field class label is the concatenation of its singlet class labels. The covariance matrices for field-classes of arbitrary length can be constructed from the *N* singlet-class-conditional covariance matrices $C_1, C_2, \ldots, C_N$ and the $N(N-1)/2$ *cross-covariance* matrices $C_{11}, C_{12}, \ldots, C_{NN}$ [2]. We consider the field-class-conditional covariance matrices for field patterns of length *L*=2, $\mathbf{y} = [\mathbf{x}_1^T \ \mathbf{x}_2^T]^T$, with class labels $\mathbf{c} = [c_i \ c_j]^T$, $i, j = 1, 2, \ldots, N$.

$$\mathbf{K}_{ij} = \begin{bmatrix} C_i & C_{ij} \\ C_{ji} & C_j \end{bmatrix} \qquad \forall i, j = 1, 2, \ldots, N$$

where $C_i = E[\mathbf{x}_1 \mathbf{x}_1^T \mid c_i] - \mu_i \mu_i^T$, $C_{ij} = E[\mathbf{x}_1 \mathbf{x}_2^T \mid c_i c_j] - \mu_i \mu_j^T$, and μ_i and μ_j are the means of class *i* and *j* [1].

The average field-class covariance matrix (over all field classes except those with repeated singlet class) is:

$$\overline{\mathbf{K}} = \frac{1}{N(N-1)} \sum_{i,j=1, i \neq j}^{N} \mathbf{K}_{ij}$$

The $d \times d$ off-diagonal blocks of $\overline{\mathbf{K}}$ are average cross-covariance matrices over class pairs, which represent the average style dependence in a dataset. In order to facilitate comparison across different datasets, we normalize the covariances in the matrix $\overline{\mathbf{K}}$ by the standard deviations of the features. The pq^{th} element of the matrix $\overline{\mathbf{K}}$, i.e., $\overline{\mathbf{K}}_{pq}$ is normalized as

$$\widehat{\mathbf{K}}_{pq} = \frac{\overline{\mathbf{K}}_{pq}}{\sqrt{\overline{\mathbf{K}}_{pp} \overline{\mathbf{K}}_{qq}}} \qquad \forall p, q = 1, 2, \ldots, 2d$$

The elements of the $d \times d$ off-diagonal blocks of the normalized matrix $\widehat{\mathbf{K}}$ are the net correlation coefficients between features of pairs of distinct patterns. We use the average of the absolute net correlation coefficients as a scalar measure to quantify multi-class style in a dataset, i.e.,

$$R = \frac{1}{d^2} \sum_{p=1}^{d} \sum_{q=d+1}^{2d} \left| \widehat{\mathbf{K}}_{pq} \right|$$

Style correlation can be extended to an arbitrary number of patterns in the fields. Here again the difference in the values of R before and after style homogenization predicts the gain of field classification.

[1] To simplify the notation, we let $E[\cdot]$ denote estimates of the corresponding sample averages.

4. Experiments

We conducted experiments on machine-printed (MP) and handwritten (HP) digits data (Fig. 1), each with ten digit classes 0-9, to evaluate our measures. MP digits are rendered in Arial, Avant Garde and Bookman Old Style. HP digits are from the SD3 dataset of NIST special database SD19. The writers of SD3 are census bureau field personnel.

Both datasets in Table 4, and the feature extraction methods, are described in [1, 2]. For comparability with HP, we divided the MP dataset into 75 virtual sources with 10 samples per class each.

Table 4. MP & HP training sets

Dataset	Sources	Number of samples
SD3-Train	0-399	42969
Machine-printed	1-75	7500

We partitioned the printed digits into 3 styles by typeface, and clustered the hand-print into 3 styles with the *K*-means algorithm. The number of Source Class-Style Assignments increases quadratically with the number of styles, which causes small sample effects with finite samples. Three styles are used to as a compromise between small sample effects and adequate representation of style variation. We applied style homogenization, as described in Section 2, to each data set, and computed all the style measures on the resulting homogenized training set (Table 5).

Table 5. Style measures for MP & HP training sets

Data Set	MP (E[H]=0.90)			HP (E[H]=0.87)		
	$H_{average}$	MI	R	$H_{average}$	MI	R
DS I	0.00	1.60	0.026	0.50	0.37	0.019
DS II	0.90	0.27	0.001	0.86	0.04	0.004

Both $H_{average}$ and $E[H]$ in Table 5 are averaged over all classes and normalized by the maximum entropy $\log_2 3$. We can observe that DS I has more single-class style than DS II since the difference between the $H_{average}$ and $E[H]$ in DS I is much higher than in DS II. Furthermore, MP exhibits more single-class style than HP, as expected. Entropy is a logarithmic unit, so a difference of 0.1 is significant. DS II, with single-class style removed, verified the hypothesis that without single-class style, the difference between the average and expected entropy will be close to zero.

To compute *MI*, we generated 36 possible combinations of Source Class-style Assignments for each digit-pair by quantizing the Source Style Probability Vectors per class to 6 possible style assignments. We observe that DS II with multi-class style removed has much lower values of *MI* than DS I. This is consistent with our hypothesis that high *MI* means strong multi-class style.

Style correlation *R* also showed high difference between DS I and II, implying the absence of style consistency (multi-class style) in DS II due to style homogenization.

We ran the SQDF style-constrained field classifier [2] on the four test sets with different field lengths *L* (Table 6). DS I has indeed fewer classification errors for $L>1$ than DS II. The results are consistent with the amount of style information, that is, the more style, the greater the gain from field classification.

Table 6. Singlet errors on 7,500 digits of MP-test and 42,821 digits of SD3 test vs. field length L

Data Set	MP			HP		
	L=1	L=2	L=3	L=1	L=2	L=3
DS I	41	33	28	657	625	607
DS II	41	41	41	664	667	663

5. Conclusion

The accuracy gains obtained by a style-consistent classifier depend on the amount of style in a dataset. We investigated and applied three measures to machine-printed and handwritten digit data. Entropy is an effective measure of single-class style. For multi-class style, *R* is a more sensitive measure than *MI* because it is based on the amount of inter-class statistical dependence of the features used for classification, rather than on what is reflected by the style labels. Furthermore, *R* is not biased by small-sample fluctuations, whereas they inevitably *increase MI*. We also demonstrated the validity of the three proposed metrics by comparing the amount of single-class and multi-class style between the datasets before and after style homogenization, and through their correspondence with the number of errors obtained by field classification.

6. References

[1] P. Sarkar and G. Nagy, "Style consistent classification of isogenous patterns," *IEEE Trans. PAMI*, 27(1), 14-22, 2005.

[2] S. Veeramachaneni, G. Nagy, "Style context with second order statistics," *IEEE Trans. PAMI*, 27(1), 88-98, 2005.

[3] G. Nagy and X. Zhang, "Simple statistics for complex feature space," in *Data Complexity in Pattern Recognition* (M. Basu and T. K. Ho, eds.), Springer Verlag, in press 2006.

[4] X. Zhang and S. Andra, "Towards quantifying the amount of style in a dataset," *SPIE/IS&T Conf. On Document Recognition & Retrieval XIII*, vol. 6067-07, 2006.