Interaction for Style-constrained OCR

Sriharsha Veermachaneni[a] and George Nagy[b]

[a]SRA division, ITC-IRST, Trento, Italy
[b]RPI ECSE DocLab, Troy, NY, USA

ABSTRACT

The error rate can be considerably reduced on a style-consistent document if its style is identified and the right style-specific classifier is used. Since in some applications both machines and humans have difficulty in identifying the style, we propose a strategy to improve the accuracy of style-constrained classification by enlisting the human operator to identify the labels of some characters selected by the machine. We present an algorithm to select the set of characters that is likely to reduce the error rate on unlabeled characters by utilizing the labels to reclassify the remaining characters. We demonstrate the efficacy of our algorithm on simulated data.

Keywords: active sampling, style-constrained classification, interaction, rejection, reject criterion

1. INTRODUCTION

OCR requires a delicate balance between economy of operator intervention (labeling training samples, proofreading output, and correcting errors), acceptable throughput with ordinary desktop, laptop or pocket computers, and high accuracy.

We consider a scenario where it is known that every document to be read is in one of a finite (but possibly large) number of styles, such as any of several fonts, writers, or printing and copying machines with device-characteristic degradations. We will use the pair of terms *field* and *singlet*, interchangeably with the words *document* and *character* respectively. We know that for a single-style document, the most accurate classifier is generally the style-specific classifier trained on that particular style (i.e., the *style-aware* classifier). However, applying such a classifier requires that we know the style. If the style is not known in advance, a style classifier can first attempt to identify the style and then the appropriate style-specific classifier can be applied. We have shown elsewhere that the optimal strategy is, however, to apply a style-constrained classifier.[1] A style-constrained classifier is the Bayes optimal classifier for multi-character fields under the assumption that all the characters within the field are in the same style.

We assume that a style-specific (i.e, font, writer, degradation) classifier has already been constructed for every expected style. We further assume that we have a classifier that identifies, perhaps not very accurately, the style of the document. We show that a style-constrained classifier can be constructed as a function of the outputs of these classifiers. The style-constrained classifier thus constructed, however, may not be much more accurate than a style-blind classifier (i.e., a classifier that ignores the fact that the document is in one style) because the style classifier is error prone. Given that some human interaction is necessary to achieve a desirable classification accuracy on the document, our goal is to optimize such interaction.

Humans are not very good at recognizing styles, therefore it is desirable to enlist the machine's help in this endeavor. Humans are, however, very good at reading regardless of style, i.e., at recognizing the class of an unknown pattern. High human reading accuracy may be attributed to use of more extensive context, and to the additional information available from a rendered pattern over its internal feature representation. The features seldom capture all the discriminating information present in the pixel map representation of the characters. We therefore take advantage of human competence in reading to identify the style of each document. The key point of the paper is that labeling the appropriate patterns by class can help to identify their common style.

Further author information: (Send correspondence to George Nagy)
Sriharsha Veeramachaneni: E-mail: sriharsha@itc.it
George Nagy: E-mail: nagy@ecse.rpi.edu

We propose the following scheme. The computer processes each same-style document. It selects a few characters (say 1-3, depending on the difficulty of the document) and displays them in context. The operator keys in the alphanumeric labels of these characters. The system then makes use of the additional information to make another pass to recognize the entire document.

We present the solution to the new research problems that underlie this proposal.

- *How does the machine select the patterns to be identified by the operator?* One natural strategy is to select the patterns in the document that are most difficult to classify. We will show that that is not the best strategy.
- *How does the machine exploit the additional information for improved accuracy?* It is clear that if a human operator labels carefully selected patterns in a document, the overall error rate on the remaining samples is lower than the initial error rate on the entire document (in other words, rejection of ambiguous patterns can reduce error rate). We will show that the labeled patterns can be used to obtain a reduction in error rate beyond that obtained because of *rejection*.

2. PREVIOUS WORK

Relevant previous work includes font recognition,[2,3] style-constrained classification[1,4] and active sampling.[5–7] Both font/style recognition and style-constrained classification rely on the assumption that the style of the unlabeled test field can be identified accurately. This may be untrue in situations where the number of styles is large or the fields are short. Active learning has traditionally dealt with selecting unlabeled patterns to label in order to create a training set. We apply the principles of active learning to the problem of optimizing human interaction, during the testing phase, in order to achieve a desired classification accuracy. Although all of the above have been topics of active research, we are not aware of any previous work on combining them for intelligent interaction for the classification of style-consistent documents. Another peripherally related work is Interactive Visual Recognition,[8] where the goal is to minimize operator input for feature extraction for visual object recognition.

The exponentially high value of labeled samples compared to unlabeled samples was pointed out, in the context of style-free classification, by Castelli and Cover.[9] The value is relatively even higher when, as in the present context, they are selected purposefully rather than randomly. However, Castelli and Cover used labeled samples to identify class-conditional feature distributions, whereas we use them directly to classify the remaining patterns in the document.

3. STYLE-CONSTRAINED CLASSIFICATION

For completeness we present in brief the basic assumptions and results about style-constrained classification. We consider the problem of classifying a field-feature vector $\boldsymbol{y} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_L^T)^T$ (each $\boldsymbol{x}_i$ represents d feature measurements for one of L singlet patterns in the field) produced in one style $s \in 1, \ldots, S$. The field feature vector is an instance of the random vector $\mathbf{y} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_L^T)^T$. Let $\mathcal{C}$ be the set of singlet-class labels. Let $\mathbf{c}^i$ represent the class of the i^{th} pattern of the field*. We make the following assumptions on the class and feature distributions.

1. $p(\mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^L) = p(\mathbf{c}^1)p(\mathbf{c}^2)\ldots p(\mathbf{c}^L)$. That is, there is no higher order linguistic dependence than the prior class probabilities.
2. $p(s|\,\mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^L) = p(s)\ \forall s \in 1, \ldots, S$. That is the prior class probabilities are style-independent.
3. $p(\mathbf{y}|\,\mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^L, s) = p(\mathbf{x}_1|\,\mathbf{c}^1, s)p(\mathbf{x}_2|\,\mathbf{c}^2, s)\ldots p(\mathbf{x}_L|\,\mathbf{c}^L, s)\ \forall s \in 1, \ldots, S$. The features of each pattern in the field (i.e., given the style) are class-conditionally independent of the features of every other pattern in the same field.

*If the fifth pattern of the field is a B, then it is denoted $\mathbf{c}^5 = B$.

For later use, we can show that for any subset $\{i_1, \ldots, i_P\} \subset \{1, \ldots, L\}$

$$p(\mathbf{c}^{i_1}, \ldots, \mathbf{c}^{i_P} | s, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) = p(\mathbf{c}^{i_1} | s, \boldsymbol{x}_{i_1}) \ldots p(\mathbf{c}^{i_P} | s, \boldsymbol{x}_{i_P}) \tag{1}$$

Under our assumptions the classification rule (called $SOPT$) that minimizes the singlet error rate assigns the label $\hat{c}_l$ to $\boldsymbol{x}_l$, the l^{th} pattern in the field $\boldsymbol{y} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_L^T)^T$, where

$$\hat{c}_l = \underset{c \in \mathcal{C}}{\operatorname{argmax}}\; p(\mathbf{c}^l = c | \,\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \sum_s p(\mathbf{c}^l = c | \,\boldsymbol{x}_l, s) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \tag{2}$$

and

$$p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) = \frac{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L | s) p(s)}{\sum_s p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L | s) p(s)} = \frac{p(s) \prod_{l=1}^{L} p(\boldsymbol{x}_l | s)}{\sum_s p(s) \prod_{l=1}^{L} p(\boldsymbol{x}_l | s)} \tag{3}$$

We note from Equation 2 that we can build the $SOPT$ classifier if we have access to style-specific classifiers for each style as well as a style classifier. (The classifiers are required to output posterior probabilities and not just a classification decision.) For the sequel we postulate that these classifiers are available (i.e., we have estimates for $p(s|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$ and $p(c_i|s, \boldsymbol{x}_i)$ for all s and i). The classifiers are pictorially depicted in Figure 1.

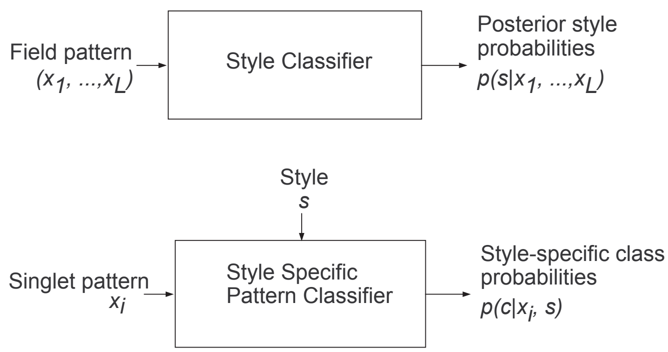

Figure 1. We require the availability of these classifiers. Note that it is not necessary that the style classifier operate on the same feature space as the style-specific classifiers.

We have recently shown that the singlet error rate of the $SOPT$ classifier decreases with field length to the average intra-style error rate.[10] The rate of decrease is determined by the style-identifiability, that is, the extent to which the style-conditional pattern distributions differ in the singlet feature space. In situations where the styles are confusable, a human operator that provides the labels of carefully selected singlets in the test field can help the system to reduce style uncertainty, thereby reducing the error rate on the unlabeled part of the field.

4. STYLE-CONSTRAINED CLASSIFICATION OF PARTIALLY LABELED FIELDS

Before we reason about the selection of the singlets in the field to be labeled, we will present the optimal way to utilize the labels once they are obtained. Let us assume that the human operator has provided the labels for the subset $\{i_1, \ldots, i_P\} \subset \{1, \ldots, L\}$. Let the labels for the singlets in the subset be $(c_{i_1}, \ldots, c_{i_P})$. Let the l^{th} singlet in the field be unlabeled. The assignment to this singlet, optimized for singlet error rate, is

$$\begin{aligned}
\hat{c}_l &= \underset{c \in \mathcal{C}}{\operatorname{argmax}}\; p(\mathbf{c}^l = c | \,\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_P} = c_{i_P}) \\
&= \underset{c \in \mathcal{C}}{\operatorname{argmax}}\; \frac{p(\mathbf{c}^l = c, \mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_P} = c_{i_P} | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)}{p(\mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_P} = c_{i_P} | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)} \\
&= \underset{c \in \mathcal{C}}{\operatorname{argmax}}\; \frac{\sum_s p(\mathbf{c}^l = c, \mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_P} = c_{i_P} | s, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)}{\sum_s p(\mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_P} = c_{i_P} | s, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)}
\end{aligned}$$

$$= \operatorname*{argmax}_{c \in \mathcal{C}} \frac{\sum_s p(\mathbf{c}^l = c | \boldsymbol{x}_l, s) p(\mathbf{c}^{i_1} = c_{i_1} | \boldsymbol{x}_{i_1}, s) \ldots p(\mathbf{c}^{i_P} = c_{i_P} | \boldsymbol{x}_{i_P}, s) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)}{\sum_s p(\mathbf{c}^{i_1} = c_{i_1} | \boldsymbol{x}_{i_1}, s) \ldots p(\mathbf{c}^{i_P} = c_{i_P} | \boldsymbol{x}_{i_P}, s) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)} \tag{4}$$

Note that all the quantities required in Equation 4 can be obtained from the style and style-specific classifiers shown in Figure 1. Of course only the numerator of Equation 4 needs to be computed, since the denominator is independent of c. Since the posterior class probabilities of the l^{th} singlet can also be written as

$$p(\mathbf{c}^l = c | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_P} = c_{i_P}) = \sum_s p(\mathbf{c}^l = c | \boldsymbol{x}_l, s) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_P} = c_{i_P}) \tag{5}$$

we effectively alter the posterior style probabilities by labeling a subset of the singlets in the field, thereby reducing the error rate on the unlabeled singlets. If the resulting distribution over styles is highly peaked, i.e., if we are almost certain of the style of this document, we may proceed to classify all the characters with the appropriate style-specific classifier. If, on the other hand, the entropy of the style distribution is relatively high, then a style-constrained classifier with the new style probabilities will be better.

5. SELECTION OF SINGLETS TO LABEL

Let us first consider the selection of the single most useful character in the field to be labeled. Clearly the best choice is the one that minimizes the error rate on the remaining singlets in the field in the expected sense (the expectation is over all the possible labels of the chosen singlet). The probability that the j^{th} singlet in the field is misclassified given that the i^{th} singlet has label c_i is given by

$$p(e_j | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^i = c_i) = 1 - \max_{c_j \in \mathcal{C}} p(\mathbf{c}^j = c_j | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^i = c_i) \tag{6}$$

Therefore the expected probability of error in the j^{th} singlet when the i^{th} singlet is labeled is given by

$$p(e_j | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^i) = 1 - \sum_{c_i \in \mathcal{C}} \max_{c_j \in \mathcal{C}} p(\mathbf{c}^j = c_j | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^i = c_i) * p(\mathbf{c}^i = c_i | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \tag{7}$$

where

$$\begin{aligned} p(\mathbf{c}^j = c_j | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^i = c_i) &= \frac{\sum_s p(\mathbf{c}^i = c_i, \mathbf{c}^j = c_j | s, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)}{p(\mathbf{c}^i = c_i | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)} \\ &= \frac{\sum_s p(c_j | s, \boldsymbol{x}_j) p(c_i | s, \boldsymbol{x}_i) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)}{p(\mathbf{c}^i = c_i | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)} \end{aligned} \tag{8}$$

Therefore Equation 7 can be rewritten as

$$p(e_j | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^i) = 1 - \sum_{c_i \in \mathcal{C}} \max_{c_j \in \mathcal{C}} \sum_s p(c_j | s, \boldsymbol{x}_j) p(c_i | s, \boldsymbol{x}_i) p(s | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \tag{9}$$

Note that all the quantities required for the calculation of the expected error rate can be obtained from the classifiers in Figure 1. Viewing the singlet errors as Bernoulli random variables with the probabilities given by Equation 9, we can calculate the expected number of errors in the field given that the i^{th} singlet has been labeled. According to Le Cam's theorem†, the expected number of errors is well approximated by

$$R(i) = \sum_{j \neq i} p(e_j | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^i) \tag{10}$$

Therefore the best singlet to label is the one that minimizes $R(i)$.

Now we wish to find the P most useful characters in the field to label. Four alternative methods present themselves immediately:

†Le Cam's theorem states that the distribution of the sum of independent but not necessarily identically distributed Bernoulli random variables is 'well' approximated by a Poisson distribution, whose mean is equal to the sum of the means of the Bernoulli variables.

1. Choose the P characters with minimum $R(i)$ values.
2. Choose the P characters $(i_1, ..., i_P)$ such that $R(i_1, ..., i_P)$ is minimum. ($R(i_1, ..., i_P) = \sum_{j \notin \{i_1,...,i_P\}} p(e_j|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^{i_1}, \ldots, \mathbf{c}^{i_P})$ is the expected error rate on the remaining singlets in the field when the subset $(i_1, ..., i_P)$ is labeled. It can be computed by extending the above formulas straightforwardly.)
3. Iteratively choose the P singlets, by minimizing at each step the expected error at the next iteration, given all the labels obtained thus far. This is the so-called *greedy* or the *single step lookahead* approach.
4. Iteratively choose the P singlets, by minimizing at each step the expected error after all P singlets have been labeled, given all the labels obtained thus far. This is called the *multi step lookahead* approach.

It is easy to find examples to show that Alternative 1 is suboptimal. Alternatives 2 and 4 are computationally expensive because of the combinatorial explosion of number of choices to compare. We demonstrate the efficacy of Alternative 3 on simulated data in the section on experimental results. We call the sampling strategy based in Alternative 3 the *Greedy Minimum Error (GME)* sampling scheme. Given a style-consistent document, our algorithm to sample the labels of P characters, and to subsequently classify the field, is shown as pseudocode below.

```
Algorithm : SAMPLEANDCLASSIFY((x_1, ..., x_L), P)

  comment: Arguments are the field (x_1, ..., x_L) and the number of singlets to label P
  PartialLabels = ( )
  do P times
      for each (i ∈ {1, ..., L}) such that (x_i not yet labeled)
          R(i) = 0
          for each (j ∈ {1, ..., L}) such that (j ≠ i) & (x_j not yet labeled)
              R(i) = R(i) + p(e_j | x_1, ..., x_L, PartialLabels, c^i)
                  comment: Calculate the expected conditional error in singlet j if singlet i
                           was to be labeled, given all the other labels
          end
          î = argmin_i R(i)    comment: Select the best singlet to label
          c_î = Label the singlet î    comment: The human operator enters the label
          Append {c^î = c_î} to PartialLabels
      end
  end
  ClassificationLabel = SOPT((x_1, ..., x_L), PartialLabels)
      comment: Classify the field given the partial labels, cf. Equation 4
 return (ClassificationLabel)
```

In the pseudocode above the only calculation we have not yet shown explicitly is that of $p(e_j|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, PartialLabels, \mathbf{c}^i)$. It is calculated as follows. Assume that $PartialLabels = (\mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_p} = c_{i_p})$, then

$$p(e_j|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L, \mathbf{c}^{i_1} = c_{i_1}, \ldots, \mathbf{c}^{i_p} = c_{i_p}, \mathbf{c}^i) = \\ 1 - \sum_{c_i \in \mathcal{C}} \max_{c_j \in \mathcal{C}} \frac{\sum_s p(c_j|s, \boldsymbol{x}_j) p(c_i|s, \boldsymbol{x}_i) p(c_{i_1}|s, \boldsymbol{x}_{i}) \ldots p(c_{i_p}|s, \boldsymbol{x}_{i_p}) p(s|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)}{\sum_s p(c_{i_1}|s, \boldsymbol{x}_{i_1}) \ldots p(c_{i_p}|s, \boldsymbol{x}_{i_p}) p(s|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)} \tag{11}$$

Note that, as before, all the quantities we need to compute the above expected conditional error rate can be obtained from the classifiers in Figure 1.

6. EXPERIMENTS

For evaluating our sampling strategy we pseudorandomly generated $50,000$ test fields of length $L = 5$. The style of each test field is drawn independently from the set $\{1, 2\}$ with equal likelihood (i.e., $p(s = 1) = p(s = 2) = 1/2$). For each pattern in a field, its singlet-class label is drawn independently from the set $\mathcal{C} = \{A, B\}$ with equal likelihood (i.e., $p(A) = p(B) = 1/2$).

The feature value for each singlet is drawn from unit variance Gaussian style-and-class-conditional feature distributions that are configured as shown in Figure 2. The feature distributions are $(\mathsf{x}|\, A, s_1) \sim N(0, 1)$, $(\mathsf{x}|\, A, s_2) \sim N(3, 1)$, $(\mathsf{x}|\, B, s_1) \sim N(2, 1)$ and $(\mathsf{x}|\, B, s_2) \sim N(1, 1)$.

Since we know the true joint style-class-and-feature distribution for our problem, we can easily construct the required style and style-specific classifiers shown in Figure 1.

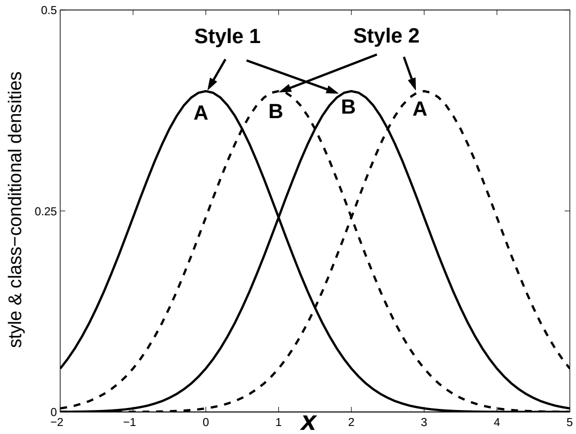

Figure 2. Style-and-class conditional distributions for the simulation experiments.

For each field in the dataset we select P singlets to label according to one of three sampling schemes, a)Random: randomly select the P singlets from the $L = 5$ singlets in the field, b)Difficult: select the P most difficult singlets in the field (i.e., the P singlets whose $SOPT$ posterior class probabilities for the two classes are most similar), and c)GME: using the incremental minimum expected error method described in the pseudocode in the previous section. After the P singlets have been selected, we assign them their true labels simulating an infallible human operator.

Once the P singlets have been labeled, we can either a)Reject: classify the entire field using $SOPT$, then ignore the "rejected" P singlets and leave the labels assigned to the remaining singlets in the field unaltered or b)Use: use the labels of the P singlets to reclassify the remaining field (cf. Equation 4).

The resulting singlet error rates on the unlabeled part of the dataset for each sampling scheme, when P singlets of each field are labeled, is shown in Table 1 for increasing values of P.

7. DISCUSSION

We observe from Table 1 that the GME sampling algorithm picks a 'better' set of singlets to label than the more common strategy that picks the most difficult patterns in the field. The results also show that the labels obtained from the human can be utilized for more accurate classification of the remaining field. In fact, even when the P singlets are chosen randomly from the field, the operator-assigned labels can be used to reduce the error rate on the remaining portion of the field, whereas if the P singlets were just ignored, the error rate would be unchanged. We do not have the error rates for GME for *Reject* when $P > 1$ because GME uses the labels of previously chosen singlets to pick the next singlet (i.e., the previously picked singlets are not really ignored).

Table 1. Singlet error rates for various sampling methods, both when the selected samples are rejected and when the labels are used to reclassify the remaining field. The different rows correspond to different number of labeled singlets. Compare the error rates to that of the *style-aware* classifier (i.e., when the style is known) which is equal to $Q(1) = \frac{1}{2}(1 - \text{erf}(\frac{1}{\sqrt{2}})) = 0.159$.

Sampling scheme	**Random**		**Difficult**		**GME**	
P	Use	Reject	Use	Reject	Use	Reject
0	0.267	0.267	0.267	0.267	0.267	0.267
1	0.220	0.267	0.200	0.235	0.189	0.240
2	0.195	0.267	0.162	0.208	0.137	
3	0.180	0.266	0.138	0.189	0.091	
4	0.171	0.266	0.126	0.174	0.057	

Although all our discussion assumed that we are dealing with one style-consistent document from which we need to select singlets to label, our results can be extended easily to a situation where several short fields have to be processed and we need to select P singlets to label not per field but over all fields. This is a more reasonable approach in problems like postal code reading, where high rejection rates are common but interaction for every field is impractical.

Another important point to note is that operators can enter contiguous character strings much faster than isolated characters. In a practical application, it may be desirable to take this into account in designing the interaction algorithm and interface. There is clearly a balance between the human time required to label the selected style-discriminative characters, the machine-time, and the cost of any residual errors. A formal model for reaching the optimal trade-off among these cost factors was presented at DIAL04. A better stopping criterion, instead of just selecting a predetermined number of singlets to label, could be obtained when all the different costs are specified.

8. SUMMARY

On the premise that a style-aware classifier yields the highest accuracy for a single-style document, an effective interaction (that attempts to minimize human effort for the greatest gain in classification accuracy) is derived for an operational OCR system consisting of several style-specific classifiers. One way to utilize operator intervention is to query the style of the document. Because only highly trained operators are able to make sound determination of the style of a document, we proposed an indirect approach to error minimization. We presented a way to select the characters whose labels are most likely to reduce the uncertainty of the remaining patterns in the field. The operator then labels these characters. The final classification exploits the resulting estimates of style probabilities for accurate classification of all the characters in the document.

We note that all the quantities necessary for style-constrained classification as well for selecting the singlets to label are obtained by combining the results of the style and style-specific classifiers.

The proposed method is most appropriate for short documents with large alphabets (like Chinese advertisements or forms) in one of several styles. It should also apply to holistic large-vocabulary word classification with only a few words per message from each font or writer. It is inappropriate for long documents with few classes, because then algorithmic style, font or writer recognition font recognition will do well without any human help.

REFERENCES

1. S. Veeramachaneni and G. Nagy, "Style context with second-order statistics," *IEEE Trans. PAMI* **27**(1), pp. 88–98, 2005.
2. H. Shi and T. Pavlidis, "Font recognition and contextual processing for more accurate text recognition," in *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, **1**, pp. 39–44, 1997.
3. A. Zramdini and R. Ingold, "Optical font recogition using typographical features," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8), pp. 877–882, 1998.

4. P. Sarkar and G. Nagy, "Style consistent classification of isogenous patterns," *IEEE Trans. PAMI* **27**(1), pp. 14–22, 2005.
5. D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," in *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, eds., **7**, pp. 705–712, MIT Press, 1995.
6. D. J. C. MacKay, "Information-based objective functions for active data selection," *Neural Computation* **4**(4), pp. 590–604, 1992.
7. H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.
8. J. Zou and G. Nagy, "Evaluation of model-based interactive flower recognition.," in *ICPR (2)*, pp. 311–314, 2004.
9. V. Castelli and T. M. Cover, "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *IEEE Transactions on Information Theory* **42**, pp. 2102–2117, November 1996.
10. S. Veeramachaneni and G. Nagy, "Analytical results on style-constrained Bayesian classification of pattern fields." Accepted to IEEE Trans. PAMI.