社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

Classifiers That Improve with Use

George Nagy
Rensselaer Polytechnic Institute
Troy, NY, 12180-3590 USA
E-mail: nagy@ecse.rpi.edu

Abstract: Training on non-representative data causes any classifier to make many mistakes on new data. Retraining an OCR engine with labeled characters, obtained from routine post-editing, can reduce both the bias and the variance of the classifier, and therefore its error rate. In the absence of post-edits, the imperfect labels assigned by the classifier can be used instead. Although the theoretical foundations of decision-directed adaptation are meager, adaptation has proved successful in diverse experiments. When the operational data can be partitioned into isogenous subsets, the classifier parameters should be adapted independently on each subset. However, if the same-source subsets are small, as in postal-code or bank-check reading, it is advantageous to classify more than one character at a time. Style-constrained classification allows training the classifier on fields shorter than the classification field. Systematic methods still remain to be developed for adapting language context to the operational data stream, particularly for semi-structured business forms. Only dynamic classifiers can hope to rival human performance on imperfectly printed, written, copied, or scanned documents.

Keywords: Dynamic classifier; Semi-supervised or unsupervised learning; Decision-directed adaptation; Self-correcting classifier; Style-constrained field classification; Weakly-constrained data; Non-representative training set; Language context.

使えば使うほど賢くなる識別器

George Nagy
レンセラー工科大学
Troy, NY, 12180-3590 USA
E-mail: nagy@ecse.rpi.edu

要約：代表から外れたデータを学習させると，どんな識別器でも新データでは多くの誤認識が生じる。正解付けされた学習データを用いてOCRエンジンを再学習すれば，識別器の偏りと分散の両者、すなわち、誤認識率を減らすことが可能である。学習データを予め編集できない場合には、そのかわりに，識別器によって決められた（幾つかは誤りの可能性のある）不完全ラベルを識別器適応化に使うことになる。この判定駆動型の適応化は，それに関する理論的基礎は貧弱ではあるが、種々の実験でうまくいくことが実証されている。取り扱っているデータを一つの元から生じたサブセット（例えば、同一筆者や同一フォント種のセットなど）に分割できる場合には、識別器のパラメータをそれぞれのサブセット用に独立に適応させるべきである。しかしながら、例えば、郵便番号や小切手読取りのように、もし同じ元をもったサブセットのデータ数が少ない場合には、一度に複数文字の文字列を識別することが有利となる。スタイル制約型識別では、識別フィールドより短いフィールドで識別器をトレーニングさせることができる。特に準定型書式の読取りのためには，言語文脈情報を取扱い可能なデータストリームに適合できる体系的手法の開発がまだ残されてはいるが、この動的識別器のみが、不完全な印刷・手書き・複写文書を読取れる人間の能力に対抗できる望みとなる。

キーワード：動的識別器；準教師あり学習、教師なし学習；判定指向型適応；自己修正識別器；スタイル制約型フィールド識別；弱制約データ；非代表学習データセット；言語文脈

1. Introduction

Statistical pattern recognition is based on the doctrine that the decision boundaries can be estimated from a representative training set. However, in many practical situations, a representative training set is not available. In this essay we explore the consequences of some scenarios where the learning process of the classifier continues while it classifies unknown samples. Although some of the notions we discuss are applicable – and may indeed originate – elsewhere, we focus on Optical Character Recognition where the notion of a labeled sample, represented as a point in feature space, is relatively unambiguous.

We believe that improvements in OCR accuracy during the next several years will be due mainly to the replacement of *static* classifiers by *dynamic* classifiers that modify their decision boundaries to fit the operational data. A concomitant change in perspective is the shift from the classification of single characters to the classification of all the characters in a field, a page, or even an entire document, as in Document Image Decoding [1,2]. These notions are related, but distinct.

There are at least four common situations where the training set is not "representative," and where deriving additional information from operational data will be beneficial. We consider the following scenarios:

1. The training set is unbiased, but it is too small, and therefore the estimates of the classifier parameters have high *variance*.
2. The training set is *biased*, therefore any classifier based only on the parameters estimated from such a training set are suboptimal with respect to the data encountered in the field.
3. The training set is globally unbiased, but *inhomogeneous*, whereas the operational data arrives in homogenous segments (fields, pages, or documents). Therefore classifier parameters estimated from the entire training set are biased with respect to any individual segment.
4. The samples of the training set are the result of some unknown transformation of variables of a representative training set. Therefore the training set is entirely *unrepresentative*, but there exists some transformations of variables that can map the distribution of the training data to that of the field data.

We propose methods to deal with the first three cases, and speculate freely about the fourth. We also mention briefly other types of "context," and applications other than OCR and DIA where such "opportunistic" approaches have met with success. This is not a survey, but a polemic. Additional evidence that lends support to our arguments can be found in the references cited.

2. Supervised Operational Training

When the training set is too small, statistical fluctuations result in unacceptably large variance of any parameters estimated from it. The common wisdom is that at least ten samples of each class are necessary for each dimension of the feature space. With a feature set of 100 features, and the ASCII alphabet of 88 characters, this translates to a training set of about 100,000 characters, *provided* that the samples are entirely homogeneous, i.e., there are no consistent typeface, writer, or transducer variations.

The above rule-of-thumb underestimates the sample size for quadratic classifiers based on second-order statistics. With 100 features, each covariance matrix has 4950 distinct parameters in addition to the class-conditional variances. For 88 classes, the requirement of ten samples per parameter (rather than per feature) yields about four million samples, still under the assumption of homogenous classes. It is therefore not surprising that the outcomes of comparisons of the accuracy of classifier types depend mainly on the trade-off between the complexity of the classifier structure and the size and composition of the sample set (which is loosely characterized by its VC-dimension [3]). Regularization is a sound, well-established approach to avoid over-fitting the decision boundaries to the training set. In non-parametric classification, a validation set is employed for the same purpose.

With very large training sets, it is difficult to improve on the nearest neighbors classifier, because there is little reason to believe that the Bayes error in practical OCR applications is very different from zero [4]. (Classifier comparisons sometimes exclude the Nearest Neighbor, because a direct implementation is prohibitively slow. Preprocessing, which trades off storage for runtime, can yield large speed-ups, and seems eminently practicable with current computers.)

The simplest way to enlarge the training set is to make more effective use of ongoing processes. Most applications cannot tolerate many errors. Therefore either the output of the classifier is proofread and corrected, or the reject threshold is set to ensure a tolerable rate of misrecognized characters, and the rejected characters are keyed in. Either way, most of the scanned characters that have passed through the OCR system eventually gain reliable labels.

In current recognition engines, the classified character bitmaps are not retained, and the added or corrected labels are never attached to these bitmaps. We believe that future OCR systems should integrate training with post-processing, so that they will be continuously retrained with the characters actually encountered in the field. This will, of course, require even more robust training procedures than those currently used, because such training should take place during lulls in the operation, in the absence of expert personnel. As in the case of the semi-supervised training scenarios discussed below, it will be desirable to monitor the evolution of the system by rerunning periodically some standard data mix with known results. In case of egregious departures from the expected performance, the OCR system can automatically revert to a previous stable state.

This approach to increasing the size of the training set appears to be more of a question of engineering design than a research problem. Aside from increasing the size of the training set, the proposed *modus operandi* would naturally ensure a more representative training set than is likely be provided by any other method of collecting samples. We understand that in some applications, the manual corrections are physically and temporally far removed from the OCR engine itself, and may even be under a different organizational entity. Nevertheless, we believe that current computer networks can link them seamlessly.

3. Weak and Strong Style

Style is induced when different subsets of the data (fields, pages, or documents) are produced by different sources (fonts, printers, scanners, writers, speakers). Subsets produced by the same source are called *isogenous* (fields, pages, or documents). We digress to explain the difference between two kinds of consistency or homogeneity that may appear in a document collection. Prateek Sarkar called them *weak style* and *strong style* [5]. The difference is illustrated in Figure 1.

	WEAK STYLE	**STRONG STYLE**
Source 1:	**22**/*07/19***25**	*25/07/1922*
Source 2:	**2***5*/**0***5*/**193***5*	**05/05/1925**
Source 3:	**2***1*/**06**/*1***943**	**02/06/1943**
Source 4:	*03/2***4***/19***4***5*	*02/25/1942*

Figure 1. Weak style and strong style. In weak style, within a field a given digit is either always bold, or always italic. In strong style, bold and italic are never mixed within a field.

In weak style, the glyphs of each class (letters, digits, phonemes) of each *style-consistent* subset are generated by the same distribution. However, the mix of generators in the various subsets is arbitrary. Therefore different subsets (e.g. pages) exhibit arbitrary mixes of the individual class styles, even though within each subset every class appears to be homogenous. We believe that weak style rarely occurs in actual OCR applications, but we give it full consideration because it is easier to exploit than strong style. Strong style implies the presence of weak style, but not vice-versa.

For a given body of data, a natural measure of the amount of weak style is the average entropy of the source-conditional style membership per class. Of course, the benefits of weak style increase with field-length.

In strong style, only some class-styles can co-occur within the same subset. Therefore the occurrence of a pattern of a given class provides information about the appearance of patterns of other classes within the same subset. We show in the sequel that style-constrained classifiers can take advantage of such statistical dependence between the features of different patterns in the same field or document.

Because the difference between weak and strong style is subtle, we give a more detailed example. Consider patterns of two classes, "A" and "B". Let there be two styles of each, bold and italic: **A**, *A*, **B**, *B*. Let each isogenous field consist of three patterns each, described by a field feature vector **x** consisting of the three individual feature vectors, ($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$). In our notation, plain capitals denote classes without specifying the style, and bold or italics specify the style. Without style, the field class AAB has *a posteriori* probability:

$$\begin{aligned} &\mathrm{p}(\mathrm{AAB}|\mathbf{x}) \\ &= \mathrm{p}(\mathbf{AAB}|\mathbf{x}) + \mathrm{p}(\mathbf{A}\mathit{A}\mathbf{B}|\mathbf{x}) + \mathrm{p}(\mathit{A}\mathbf{AB}|\mathbf{x}) + \mathrm{p}(\mathit{AA}\mathbf{B}|\mathbf{x}) \\ &\quad + \mathrm{p}(\mathbf{AA}\mathit{B}|\mathbf{x}) + \mathrm{p}(\mathbf{A}\mathit{AB}|\mathbf{x}) + \mathrm{p}(\mathit{A}\mathbf{A}\mathit{B}|\mathbf{x}) + \mathrm{p}(\mathit{AAB}|\mathbf{x}). \qquad (1) \end{aligned}$$

With weak style, the mixed "A"s are eliminated, and the distance of different field classes from one another is thereby increased:

$$\mathrm{p}(\mathrm{AAB} \mid \mathbf{x}) = \mathrm{p}(\mathbf{AAB}|\mathbf{x}) + \mathrm{p}(\mathit{AA}\mathbf{B}|\mathbf{x}) + \mathrm{p}(\mathbf{AA}\mathit{B}|\mathbf{x}) + \mathrm{p}(\mathit{AAB}|\mathbf{x}) \quad (2)$$

Strong style constraint offers further benefits. With a strong style constraint, we can assume that within a word bold **A**'s occur only with bold **B**'s, and italic *A*'s with italic *B*'s. (Of course, **A***B* and *A***B** would also constitute formally valid strong styles, but such a mix would be unusual in printed matter.) Now the field probability for AAB is:

$$\mathrm{p}(\mathrm{AAB} \mid \mathbf{x}) = \mathrm{p}(\mathbf{AAB} \mid \mathbf{x}) + \mathrm{p}(\mathit{AAB} \mid \mathbf{x}). \qquad (3)$$

With strong style, the features are no longer class-conditionally independent:

$$\mathrm{p}(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3} \mid \omega_1, \omega_2, \omega_3) \neq \mathrm{p}(\mathbf{x_1}, \mid \omega_1)\, \mathrm{p}(\mathbf{x_2} \mid \omega_2)\, \mathrm{p}(\mathbf{x_3} \mid \omega_3).$$

A strong style constraint means that the shape of any pattern in the field provides some information about the shape of any other pattern in the same field, even if the other pattern is of a different class. We still assume, however, that "noise" is independent from pattern to pattern, i.e., that there are random variations within each class ω_i of each style *s*. Therefore the patterns are class-and-style-conditionally independent:

$$\mathrm{p}(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3} \mid \omega_1, \omega_2, , \omega_3, \mathrm{s}) = \mathrm{p}(\mathbf{x_1}, \mid \omega_1, \mathrm{s})\, \mathrm{p}(\mathbf{x_2} \mid \omega_2, \mathrm{s})\, \mathrm{p}(\mathbf{x_3} \mid \omega_3, \mathrm{s}).$$

The natural measure of strong style is the degree of statistical dependence between the features of patterns of different classes from the same source relative to the dependence between patterns from different sources.

We note that none of the above takes into account effects arising from the position or order of the patterns in the field. The feature vector for a given class is expected to have the same distribution regardless of the features of the patterns preceding or following it, or of its ordinal position in the field. The first restriction neglects the effects of ligatures. The second means that we can reduce the number of field-class computations from *sampling with replacement with order* (n^r), to *sampling with replacement without order* (${}_{n-1+r}C_r = {}_{n-1+r}C_{n-1}$), where ${}_nC_r$ is the binomial coefficient, *n* is the number of singlet classes, and *r* is the field length. Sampling is *without replacement* because each class can appear in a field any number of times.

4. Biased Training Sets and Adaptation

Decision-directed adaptation is a simple approach to character recognition under the weak style constraint. It is easy to implement with any trainable classifier. It consists of two iterated steps. In the first step, a classifier trained on whatever training data was available classifies an isogenous (and therefore presumably homogenous) batch of data – say a page. In the second step, the classifier is retrained with this same batch of data, using the labels assigned by the classifier. Then we return to the first step, and reclassify the data with the retrained classifier. Surprisingly, the error rate will now generally be lower than in the first classification cycle!

Fifty years ago, Goldstein et al. at MIT successfully applied a similar idea to the transcription of hand-sent Morse signals [6]. Here the length of the dots, dashes and three kinds of spaces varied from operator to operator, but for a given operator was relatively constant over a stretch of time. Another application, of much greater impact, was Robert Lucky's development of adaptive equalization at Bell Laboratories in the sixties [7,8]. When telephone lines were first used for digital transmission, tapped delay-line filters decoded the waveforms into digital signals. But the waveforms were affected by slowly (compared to the transmission rate) varying cross talk, and electro-magnetic disturbances. Lucky constructed an ingenious circuit that kept the "eye diagram" open by taking advantage of the fact that most of the waveforms were correctly decoded to adjust the filter weights. In the decades since, "unsupervised" adaptation has been widely used in speech recognition [9] and in multispectral classification in remote sensing [10]. We used it recently in interactive flower recognition.

Table I. Character error rates on handwritten numerals before and after adaptation. 100 Hitachi blurred directional features. Training covariance matrices not regularized.

Character error rate (%),,				
Training set	Test set	Before adaptation	After mean adaptation	After mean & covariance adaptation
SD3	SD3	2.2	1.9	0.8
	SD7	8.0	6.6	3.0
SD7	SD3	3.7	2.7	1.1
	SD7	3.6	3.1	1.8
SD3+SD7	SD3	1.7	1.5	0.6
	SD7	4.7	3.8	1.9

In 1966, we showed that a "self-corrective" classifier could reduce the error rate of a multifont classifier by a factor of five when retrained to twelve individual fonts without *any* labeled samples of these fonts [11,12]. A skeptical Henry Baird replicated the results in 1994 on a much larger data set of 100 fonts of four type sizes produced by his pseudo-random defect model [13]. In these experiments, only the mean class vectors were adapted. Recently, Veeramachaneni devised a scheme to adapt not only the means, but also the variances [14,15]. He demonstrated smaller, but significant, reduction of the error rate on two NIS hand-printed data sets represented by local directional features [16] (Table I).

We have also tried decision-directed adaptation with InkLink, a recognition system for on-line cursive writing [17,18]. InkLink is based on constrained localized polygram matching of one unknown word against many reference words, using a lexicon of valid words. InkLink avoids the explicit character segmentation required by character-based systems, and does not require a sample of each unknown word, as do word-based systems. The recognition is based on segments at least two letters long, because the feature representations of polygrams are much more distinctive than those of unigrams. The recognition of an unknown cursive word consists of the following steps:

1. The expected location where the unknown matches the reference words is pre-computed (the number of features in every letter of each lexicon and reference word is estimated by least squares).
2. The feature matches of the unknown against the reference words are found by string matching.
3. The unknown is hypothesized as each lexicon word in turn.
4. The hypothesis that corresponds best to the expected length and location of the matches is chosen.

The improvement obtained by adding recognized words to the reference list is shown in Figure 2. The average error rate on three 100-word test sets dropped from 28% to 7%, even though many of the added words were evidently mislabeled. Equivalently, with adaptation we obtained the same error rate with an initial set of 100 reference words as without adaptation with 500 reference words.

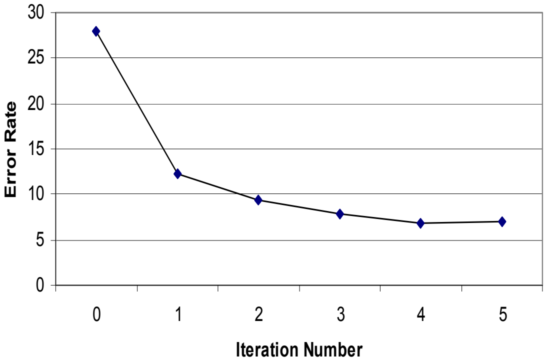

Figure 2. Results of adaptation in InkLink.

It is not clear that such a scheme can be successful with classifiers that attempt to estimate decision boundaries directly, such as Support Vector Machines [19], or that make use of discriminative training [20,21,22,23]. An obvious idea is to improve retraining of the classifier by feeding back only the patterns accepted with high confidence. Obvious as it is, this idea has failed whenever it was tried. Patterns near the boundary, that are only slightly more likely than not to be correctly classified, are obviously important. This observation suggests that we will be able to extend adaptation to high-performance classifiers and training methods.

5. Inhomogeneous Data and Style-Based Classification

As we have already noted, in most OCR applications, the input data is naturally divided into quasi-homogenous subsets. In hand-printed *forms processing* – where we include postal addresses, bank checks, and gyro forms – it is safe to presume that each form originates from a single writer. Machine-printed pages generally contain only one, or a few, typefaces and sizes. Less obviously, the data may be partitioned by the printer, scanner (including analog fax), or copier that played a role in the production of any single page or document. Even linguistic aspects may be source-dependent. It therefore makes sense to consider classification methods designed for *isogenous* documents, i.e., for applications where the input stream is partitioned into single-source documents. Different techniques may be applicable depending on the average number of patterns per document: in transcribing an archival journal, we can count on thousands of samples of each font and size, but in postal zip-code reading, we can hardly expect multiple occurrences of each class on any one envelope.

Adaptation works well when there are many instances of each class. When there are only a few, style-constrained classification is the only way to go. Style-constrained classification is a type of field classification. Conventional field classifiers are trained on samples of each field class, therefore the number of classes necessary for training increases combinatorially with field length. To avoid this exponential growth of the training set, we can exploit the unique property of Gaussian random variables that their joint behavior is specified entirely by their second order statistics. For example, the covariance matrix of three random variables (x, y, z) can be constructed entirely from the covariance matrices of (x, y), (x, z), and (y, z), because the elements of the triple covariance matrix are only the covariances of the variables taken in pairs. This holds even if x, y, and z are vectors rather than scalars. The significance of this observation is that a Gaussian field classifier for fields of arbitrary length can be trained with only pairs of samples. This decreases the size of the required training set enormously. It can also be shown that style-constrained classification makes more efficient use of the data than font recognition followed by single-font classification.

Depending on the application, we can postulate either a discrete number of strong styles, or an infinite number of continuous strong styles. In the discrete case, each style is defined by its mean vector and its covariance matrix. In the continuous case, the mean vector itself has a Gaussian distribution. Sarkar has derived the optimal classifier for the discrete case [24,25, 26, 27], and Veeramachaneni for the continuous case [28]. Because the optimal classifiers are slow, we have also devised heuristics that rely on the style consistency of overlapping pairs of characters [29].

Table II Field error rate (%) on NIST hand-printed digits.

Field length	L=2		L=5	
Test data	w/o style	with style	w/o style	with style
SD3	**1.4**	**1.3**	**3.0**	**2.5**
SD7	**2.7**	**2.4**	**5.3**	**4.5**

Table II shows an example of the reduction of error rate due to style-constrained classification. The writers in the training set and the tests are mutually exclusive.

To summarize, style-conscious classification improves the accuracy on isogenous fields compared to style-free classification (analogous to Equation 1), by exploiting either weak style constraints (Equation 2), or strong style constraints (Equation 3).

6. Weakly-Constrained Data

It is not necessary for the training data to be statistically directly representative of the test data. We may, instead, consider a labeled training set $\{\mathbf{x},z\}$ with "marginal" probability density $p_{\mathbf{X}}(\mathbf{x})$, and a test set $\{\mathbf{y}\}$ with probability density $p_{\mathbf{Y}}(\mathbf{y})$, where $\mathbf{y} = g(\mathbf{x})$. The deterministic transformation $g(\bullet)$ could be a rotation or scaling of the feature space, or even some nonlinear transformation. If $g(\bullet)$ is a global transformation, that has the same effect on all the classes, then it can be specified by relatively few parameters. Furthermore, if the transformation is independent of the class structure, then it can be discovered from a relatively few *unlabeled* test samples. It can be shown that very few labeled samples suffice to identify well-separated distributions, i.e., to determine whether a particular distribution represents A or B [30].

What kind of transformations do we expect? We expect that even after the transformation, samples of a given class will be closer to other samples of the same class (when measured with an appropriate metric) than to samples of other classes. We will be surprised if the mean of any class in feature space (with origin at the grand mean of all the samples) is a linear combination of the means of any other classes. There are both theoretical and experimental reasons to believe the classes are distributed on the surface of a hypersphere [31 , 32] and that their means are approximately equidistant, forming a regular simplex. We also suspect that the covariance matrices are determined much more by the feature set than by the nature of the patterns, and therefore the individual covariance matrices are only perturbations of the grand covariance matrix.

Nevertheless, it is certainly true that a 180 degree rotation takes *6* into *9*, or *p* into *d*. And *b* and *d,* or *p* and *q,* are nearly reflections of each other. Therefore given only a pair of such distributions, it would be impossible to decide which is which. Such symmetries disappear, however, when we consider all ten digits, or all the letters of the Latin alphabet.

In spite of the nearly universal use of the Gaussian distribution for parametric multivariate classifiers, we do not believe that it is a good representation of most symbolic glyphs. Even when within-class variances are very different, we have never been able to find any 1-D projection of feature space where patterns of the class with the larger mean appear on either side of the mean of the other class. Our experiments indicate that projecting the samples of one class on the 1-D subspace that contains the mean of that class and the grand-mean of all the classes yields a very asymmetric distribution. The tail of the distribution is much longer away from the grand mean. This is intuitively pleasing, because both the human reader and a classifier ought to be able to tolerate large variations in the shape of a character, *provided* that such variation does not make it resemble patterns of other classes.

The expected tight constellations of classes in feature space do not take into account outliers that don't belong to any class. Such patterns can be generated not only by writers, printers, scanners and copiers, but also by mis-segmentation before the classification stage. While readily tolerate outliers, it is only recently that classifier resistance to outlier tolerance has been seriously investigated in OCR [20-23]. A small number of outliers should not affect the global estimate of the transformation that maps the training set into the test set, or vice-versa.

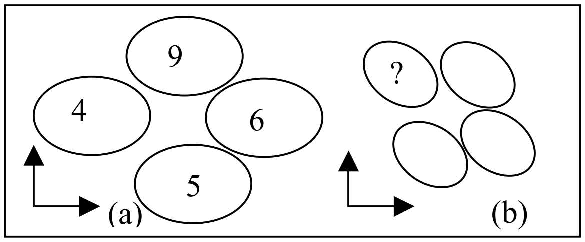

Figure 3. Four classes in two different 2-D feature spaces

Consider the example in Figure 3, which is a schematic representation of four numeral classes in two different feature spaces. The two features spaces are based on two different sets of features: for example, one could be Fourier coefficients, and the other could be pixel n-tuples [33]. There is little if any overlap between the training and test samples of any class. Is it inconceivable that we can train a classifier on samples in Feature Space (a) and modify it to classify samples in Feature Space (b)?

7. Language Context

Language context is based on the observation that only a tiny fraction of all possible sequences of symbols ever appear. Of the over 300 million six-letter combinations of the 26 English letters, the most common 100 words (*stop words*) account for over 50% of ordinary text. Furthermore, fewer than 100 sequences of any six words would make sense. Therefore in principle most errors can be readily detected, and many can be corrected.

We have demonstrated that if the bitmaps of scanned text are clustered, and each letter is replaced by a unique cluster number, then the resulting substitution cryptogram can be readily solved with a language model. In the early days, our language model was a table of bigram frequencies, and we needed several hundred words to decode the cryptogram [34,35]. By the late seventies, we were able to store a few hundred common words, and we could decode texts of less than one hundred words [36,37]. Anybody's aunt can decode a substitution cipher of a dozen words, and so can George Hart's algorithm [38]. But these algorithms are brittle. We have only demonstrated the principle: even our recent algorithms cannot tolerate upper-lower case confusions, and have trouble with clusters that split or merge classes [39]. Yet a human reader would have only a little difficulty with text where all the '*e*'s (about 10% of the letters) were replaced by '*c*'s. Beyond morphological (valid letter n-grams), lexical (valid words), and syntactic (valid combinations of words), we must also exploit semantic (does it make sense?) and pragmatic (goal dependent) constraints [40,41,42].

We have noticed that in the ISRI (Information Science Research Institute, University of Nevada, Las Vegas) competitions, OCR systems with aggressive lexical context correction often introduced egregious mistakes [43]. It is well known that using too large a lexicon, without word frequency data, is as bad as using no lexicon at all. There is no question that adapting the language model to the current input stream could result in significant reduction of the error rate on plain text. However, we anticipate the most interesting developments not in reading plain text, but in forms reading (including postal addresses and financial forms). The context here often resides in some specialized and perhaps private database, not in a public dictionary. The problem that must be solved is how to extract specialized syntactic rules from a *dynamic* database, and how to weight items of varying relevance to the *current* document. Solutions must exist, because very few forms are ambiguous to knowledgeable readers.

8. Conclusions

We believe that the future belongs to pattern recognition systems that continue learning after they leave the factory. At the very least, OCR systems must be informed of every error they make, so that they can adjust their decision boundaries. But it is not enough to let them know when they make errors: they also need to know when their decisions are right. When classifiers are idle, they should put their time to good use by reviewing the millions of patterns that they have already encountered.

Decision-directed adaptation is a powerful methodology. Although we have no solid theory to explain when it works and when it does not, we can be sure that any improvement in the basic recognition engine will be amplified through adaptation. Its success is simply a reflection of the fact that most classifiers are trained according to statistical principles; therefore the increase in the number of correctly labeled patterns will generally outweigh the much smaller increase in the number of mislabeled patterns. We have observed that more than 50% error in some classes can be reduced significantly by adaptation, *provided* that confusions between two classes are not symmetrical. Decision-directed adaptation can be applied with any classifier that is built automatically from a set of training samples. However, it has only been demonstrated experimentally to work well only on parametric classifiers with relatively few parameters. The possibility of a run-away classifier can be limited by close monitoring of recognition errors on a standard data set on which acceptable results have been previously obtained.

When the number of samples in each isogenous field is too small to allow adaptation, we can take advantage of the fact that how a particular pattern is printed, written or spoken reveals something about the appearance or sound of every other pattern produced by the same source. Under reasonable assumptions of either a discrete number or a continuum of styles, style-constrained recognition yields optimal classification. Its benefit compared to field-classification is that it does not require training samples of every possible field label. It is also provably superior to font or writer identification. In contrast to adaptation, the cost of style-constrained classification is high. It should therefore be used only on fields where simpler classifiers cannot produce dependable results.

The benefits of using linguistic context at the morphological, lexical and functional (as in forms recognition) levels are well established. We have shown that even low-level linguistic context is powerful enough to recognize text without any restriction on character shape other than that the same letter will be represented by the same shape, and different letters will be represented by different shapes. So far context has been exploited only statically, with stored letter n-gram frequencies, lexica, and grammars. We expect further benefits from systematic adaptation of syntactic and semantic context to documents and to collections of documents.

Some OCR systems have reached accuracies compared to those of an *indifferent* reader. But they are not nearly as accurate as an *interested* reader with a stake in the message and a background of relevant information. We must strive to endow our OCR machines with motivation and knowledge comparable to those of the most assiduous reader.

Acknowledgment

I am grateful to Hitachi CRL for technical and financial support since 1992. Special thanks to Drs. Fujisawa, Liu, Sako, and Shima, and to Messrs. Koga, Marukawa, and Mine. I also learned much about constellations in feature space from my former and present students, especially Jung, Sarkar,Veeramachaneni, & El-Nasan. But I take full responsibility for any misconceptions, errors, omissions, and exaggerations presented here.

References

[1] H. Baird, K. Popat, T. Breuel, P. Sarkar, and D. Lopresti, Assuring high-accuracy document understanding: retargeting, scaling up, and adapting, in *Proc. Symposium on Document Image Understanding Technology* (SDIUT03), Greenbelt, MD, pp.17-30, 2003.

[2] G. Kopec, Multilevel character templates for image decoding, in *Proceedings of Document Recognition IV*, SPIE Vol. 3027, 1997.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[4] T. Ho and H. Baird, Asymptotic accuracy of two-class discrimination, *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval* (SDAIR04), Las Vegas, pp. 275-288, April 1994.

[5] P. Sarkar, *Style Consistency in Pattern Fields*, PhD thesis, Rensselaer Polytechnic Institute, Troy, New York, 2000.

[6] B. Gold, Machine recognition of hand-sent Morse code, *IRE Trans. Information Theory*, 17(24): 17-24, March 1959.

[7] R. W. Lucky, Automatic equalization for digital communication, *Bell Systems Technical Journal*, 45(4): 547-588, February 1966.

[8] R. W. Lucky, Techniques for adaptive equalization of digital communication systems, *Bell Systems Technical Journal*, 45(2): 255-286, 1966.

[9] J. L. Gauvain and C. H. Lee, Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains, *IEEE Trans. Speech and Audio Processin*g, 2(2): 291-298, April 1994.

[10] B.S. Shahshani and D.A. Landgrebe, The effect of unlabeled samples in reducing the small sample size problems and mitigating the Hughes phenomenon, *IEEE Trans. Geoscience & Remote Sensing*, 32(5): 1087-1095, 1994.

[11] G. Nagy and G. L. Shelton Jr., Self-corrective character recognition system, *IEEE Trans. Information Theor*y, IT-12(2): 215-222, April 1966.

[12] G. Nagy, The application of nonsupervised learning to character recognition, in L. Kanal (Editor), *Pattern Recognition*, Thompson Book Company, Washington, pp. 391-398, 1968.

[13] H. S. Baird and G. Nagy, A self-correcting 100-font classifier, In L. Vincent and T. Pavlidis (editors), *Document Recognition*, SPIE Vol.2181, pp. 106–115, 1994.

[14] S. Veeramachaneni and G. Nagy, Adaptive classifiers for multisource OCR, *Int'l J. Document Analysis and Recognition*, in press, 2003.

[15] S. Veeramachaneni and G. Nagy, Classifier adaptation with non-representative training data, in D. Lopresti, J. Hu, and R. Kashi (editors), *Document Analysis Systems V: Proc. 5th Int'l Workshop on Document Analysis Systems*, Princeton, NJ, Springer LNCS 2423, pp. 123-133, August 2002.

[16] M. Yasuda and H. Fujisawa, An improved correlation method for character recognition, *Systems, Computers, and Control*, 10(2): 29-38, 1979 (translated from Trans. IEICE Japan, 62-D(3): 217-224, 1979).

[17] A. El-Nasan and G. Nagy, On-line handwriting recognition based on bigram co-occurrences, *Proc. 16th Int'l Conference on Pattern Recognition,* Quebec City, Vol. III, pp. 740-743, Aug. 2002.

[18] A. El-Nasan, *InkLink: A Writer-Dependent On-Line Unconstrained Handwriting Recognition System*, PhD thesis, Rensselaer Polytechnic Institute, Troy, New York, 2003.

[19] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 1998.

[20] C.-L. Liu, H. Sako, and H. Fujisawa, Discriminative learning quadratic discriminant function for handwriting recognition, *IEEE Trans. Neural Networks*, in press, 2003.

[21] C.-L. Liu, H. Sako, and H. Fujisawa, Learning quadratic discriminant function for handwritten character classification, *Proc. 16th Int'l Conference on Pattern Recognition,* Quebec City, Vol. IV, pp. 44-48, August 2002.

[22] C.-L. Liu, H. Sako, and H. Fujisawa, Integrated segmentation and recognition of handwritten numerals: comparison of classification algorithms, *Proc. 8th Int'l Workshop on Frontiers of Handwriting Recognition* (IWFHR02), Niagara-on-the-Lake, Canada, pp. 303-308, August 2002.

[23] C.-L. Liu, H. Sako, and H. Fujisawa, Performance evaluation of pattern classifiers for handwritten character recognition, *Int'l J. Document Analysis and Recognition*, 4(3): 191-204, 2002.

[24] P. Sarkar and G. Nagy, Classification of style-constrained pattern fields, *Proc. 15th Int'l Conference on Pattern Recognition*, Barcelona, Vol. 2, pp. 859-862, September 2000.

[25] P. Sarkar and G. Nagy, Style-consistency in isogeneous patterns, *Procs. 6th Int'l Conference on Document Analysis and Recognition*, Seattle, pp.1169-1174, Sept. 2001.

[26] P. Sarkar, An iterative algorithm for optimal style-conscious field classification, *Proc. 16th Int'l Conference on Pattern Recognition*, Quebec City, Canada, August 2002.

[27] P. Sarkar and T. Breuel, Amplifying accuracy through style consistency, *Symposium on Document Image Understanding Technology*, Greenbelt, MD, pp. 245-252, April 2003.

[28] S. Veeramachaneni, *Style Constrained Quadratic Field Classifiers*, PhD thesis, Rensselaer Polytechnic Institute, Troy, New York, 2002.

[29] S. Veeramachaneni, G. Nagy, C.-L. Liu, and H. Fujisawa, Classifying isogenous fields, *Proc. 8th Int'l Workshop on Frontiers of Handwriting Recognition* (IWFHR02), Niagara-on-the-Lake, Canada, pp.41-46, August 2002.

[30] V. Castelli and T. M. Cover, The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter, *IEEE Trans. Information Theor*y, 42(6): 2102–2117, November 1996.

[31] M. Perrone, *Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimization*, PhD thesis, Brown University, 1993.

[32]G. Nagy and S. Veeramachaneni, A Ptolemaic model for OCR, *Proc. 7th Int'l Conference on Document Analysis and Recognition*, Edinburgh, pp. 1060-1064, August 2003.

[33] D. M. Jung, M. S. Krishnamoorthy, G. Nagy, and A. Shapira, N-tuple features for OCR revisited, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(7): 734-735, July 1996.

[34] R. G. Casey and G. Nagy, An autonomous reading machine. *IEEE Trans. Computers*, C-17(5): 492-503, May 1968.

[35] R. G. Casey and G. Nagy, Advances in pattern recognition, *Scientific American,* 224(4): 56-71, 1971.

[36] G. Nagy, S. Seth, and K. Einspahr, Decoding substitution ciphers by means of word matching with application to OCR, *IEEE Trans. Pattern Analysis and Machine Intelligenc*e, 9(5): 710-715, September 1987.

[37] G. Nagy, S. Seth, K. Einspahr, and T. Meyer, Efficient algorithms to decode substitution ciphers with applications to OCR, *Proc. 8th Int'l Conference on Pattern Recognition*, Paris, pp. 352-355, October 1986.

[38] G. Hart, To decode short cryptograms, *Communications of the ACM*, 37(9): 102-108, Sept. 1994.

[39] T. K. Ho and G. Nagy, OCR with no shape training, *Proc. 15th Int'l Conference on Pattern Recognitio*n, Barcelona, pp.27-30, September 2000.

[40] G. Nagy, On the frontiers of OCR, *Proceedings of the IEEE*, 40(8): 1093-1100, July 1992.

[41] G. Nagy, What does a machine need to know to read a document? (invited), *Proc. ISRI Symposium on Document Analysis and Information Retrieval*, Las Vegas, pp. 1-10, March 1992.

[42] G. Nagy, Teaching a computer to read (invited), *Proc. 11th Int'l Conference on Pattern Recognition*, The Hague, Vol. 2, pp. 225-229, August 1992.

[43] S. Rice, G. Nagy, and T. Nartker, *Optical Character Recognition: An Illustrated Guide to the Frontier*, Kluwer Academic Publishers, Boston/Dordrecht/London (200 pages), 1999.