

# Modeling the Sample Distribution for Clustering OCR

Thomas M. Breuel

Xerox Palo Alto Research Center, Palo Alto, CA, USA

## ABSTRACT

The paper re-examines a well-known technique in OCR, recognition by clustering followed by cryptanalysis, from a Bayesian perspective. The advantage of such techniques is that they are font-independent, but they appear not to have offered competitive performance with other pattern recognition techniques in the past. The analysis presented in this paper suggests an approach to OCR that is based on modeling the sample distribution as a mixture of Gaussians. Results suggest that such an approach may combine the advantages of cluster-based OCR with the performance of traditional classification algorithms.

**Keywords:** Optical character recognition, Bayesian, font-independent, sample distribution, mixture of gaussians

## 1. INTRODUCTION

This paper examines the effects of clustering character images prior to recognition in optical character recognition (OCR) of printed documents. This approach has a long history in OCR, and prior work has addressed the questions of how to build a clustered representation quickly,<sup>1</sup> as well as how to label the resulting clusters.<sup>2,3</sup> Clustering, mixture models, and mixture-based Bayesian recognition itself, of course, has a long history in statistics and pattern recognition.<sup>4</sup> In this work, we make a connection between the two approaches. The key point is that the clustering of the character templates is, in effect, a mixture density estimation of the sample distribution. This connection allows us to reexamine issues of cluster validity, style adaptation,<sup>5</sup> and cluster label assignment within a Bayesian framework.

## 2. MOTIVATION

While unsupervised clustering prior to recognition has been proposed a number of times, it should not be at all obvious that carrying out recognition by first clustering tokens of a document image and then assigning labels to the individual clusters should actually result in either high recognition speed or good recognition performance. The operations of clustering character images into groups is computationally expensive, since it involves either extensive image comparisons (for example, using the minimum Hausdorff distance<sup>6</sup>), or feature extraction followed by simpler comparisons; in terms of speed, for most isolated character recognition methods, it would likely be faster to apply the recognition method directly to each character image. Furthermore, the clustering process itself involves assigning distinguishable characters to the same cluster. But this means that characters on different sides of a decision boundary of the classifier used for recognizing the cluster may end up in the same cluster, resulting in confusions. At the very least, the clustering and recognition methods need to be carefully chosen and trained together.

Nevertheless, unsupervised clustering prior to recognition has a number of intuitively attractive features that justify renewed interest in this approach:

- Token-based compression<sup>6,7</sup> of document images make recognition after unsupervised clustering attractive. For an end-user applying OCR to the token compressed image, it means that OCR can be carried out without fully uncompressing the image. Furthermore, many applications require storage of both the compressed document image and its transcription; applying OCR to the tokens and assigning character identities to tokens allows the image and textual information to be compressed together, likely resulting in better compression.
- Applying recognition algorithms to clustered representations of character images makes adaptation to completely novel styles and novel image degradations possible. Reliable automatic adaptation to novel fonts and conditions is probably the most important open problem in OCR systems today.
- Clustering has also been proposed for improving OCR via preprocessing<sup>8</sup> and post-processing.<sup>9</sup>

### 3. THE OCR PROBLEM

For the purposes of this paper, we will define the OCR problem in the following simplified manner. We assume that there is a fixed, finite set of characters (digits, lower case letters, upper case letters, special characters)  $\mathcal{C}$ . Furthermore, we assume that there is an open-ended set of possible styles  $\mathcal{S}$ . The notion of style encompasses character properties like font, size, and idiosyncracies of the particular rendering engine used. Picking a character  $c \in \mathcal{C}$  and a style  $s \in \mathcal{S}$  uniquely determines an idealized image (bitmap)  $B_{c,s}$  for the character. During document creation, this idealized bitmap is printed on a piece of paper. When the document is scanned back in again, a degraded bitmap of the character,  $\tilde{B}_{c,s}$  is obtained, usually by the addition of noise, blurring, thresholding, sampling error, and various forms of geometric distortions.<sup>10</sup> If the image  $\tilde{B}_{c,s}$  consists of  $n \times n$  pixels, we view it as a feature vector in an  $n^2$ -dimensional feature space. The image  $\tilde{B}$ , or a linear (e.g., principal component analysis) or nonlinear (e.g., gradient magnitude) transformation of  $\tilde{B}$ , then forms the input to the OCR system. For simplicity, the following discussion will be expressed using  $\tilde{B}$  as the feature vector, but exactly the same arguments apply in the case of a transformed feature vector.

We define  $N$ , the noise vector, to be the difference  $\tilde{B}_{c,s} - B_{c,s}$ . In general, the degradation may be composed of both a deterministic and a stochastic component and may depend on the input image  $B_{c,s}$  in complicated non-linear ways. Therefore, for each  $c, s$ , the degradation induces a probability distribution  $P(\tilde{B}|c, s)$ . In a Bayesian decision theoretic framework under a zero-one loss function, the task of an OCR system is to find the  $c$  and  $s$  that maximize  $P(c, s|\tilde{B})$ . For simplicity, we also assume a uniform prior  $P(c, s)$ .

### 4. A MIXTURE MODEL

The detailed procedure for OCR-by-clustering described by Casey<sup>2</sup> is fairly typical. It can be interpreted as randomly covering each class conditional distribution  $P(\tilde{B}|c, s)$  with balls of size  $\theta$  in feature space that are centered on samples  $\tilde{B}$ . In fact, the heuristic methods described in both Casey<sup>2</sup> and Nagy *et al.*<sup>3</sup> appear to require that the number of  $\theta$ -balls covering each class conditional distribution is small on the average and that the probability that multiple characters are assigned to the same cluster is low; otherwise, the problem of assigning labels to the clusters based on dictionary or language models becomes computationally hard and unreliable. We argue that this implies a noise model in which, with high probability, the noise vector  $N$  is bounded by a  $\theta$  ball and in which prototype features  $B_{c,s}$  are separated by a distance of at least  $3\theta$ .

While determining the exact distribution of the noise vector  $N$  is still the subject of research,<sup>10</sup> bounding by a  $\theta$ -ball appears to be not a very good approximation for the kind of representation usually used with clustering methods, the raw image  $\tilde{B}$ . An intuitive way of seeing this is the following. Each entry in the diagonal of the covariance matrix  $\Sigma$  of the noise vector  $N$  corresponds to the variance of a pixel in the input image  $\tilde{B}$ . But it is clear that entries corresponding to pixels near the boundary of the character prototype  $B_{c,s}$  are going to have much higher variances than pixels elsewhere.

There are a number of alternative models for the noise vector  $N$  we could use. For example, if we would like to stay within a bounded error framework, we could use ellipsoidal error bounds. Alternatively, we might switch to a Gaussian model with a less constrained covariance matrix  $\Sigma$ . We argue that, given the separation of prototypes implied by traditional clustering OCR methods, the Gaussian model can closely approximate such bounded error methods, and that there is no *a priori* reason to prefer one method over the other anyway. Ultimately, the question of what a good approximation to the distribution of the noise vector  $N$  in feature space is something that will be addressed in a future paper (it can be related both analytically and empirically to the distributions implied by the document image defect model described in<sup>10</sup>).

Therefore, for concreteness, let us assume that the degradation of the input consists of the class and style-specific addition of Gaussian noise:  $\tilde{B} = B_{c,s} + G(\Sigma_{c,s}, \mu_{c,s})$ . Such a model can model the differences in variances and covariances among the pixels of a character image. It has the additional advantage that mixture model estimation methods are best studied and have the most mature implementations available for the Gaussian case. Modern mixture estimation methods for Gaussian mixtures can find statistically valid representations of empirical distributions in terms of a small number of Gaussian mixture components.

Note that the class and style dependence of the noise vectors makes this a significantly more general model than merely adding noise to each character image. In particular, it allows each character to undergo its own preferred distortions and allows variability to depend on the location of a pixel with respect to the character model. A more

detailed and concrete analysis of these phenomena will be presented. Another argument in favor of this kind of model for OCR is that it already performs well on the related problem of handwritten digit recognition; see, for example, the results in Hastie and Tibshirani.<sup>4</sup>

In the above considerations, we have left open the scope of the available styles and degradation models that an OCR system needs to cope with on any given document. A simple approach is to consider the set of styles that the system might encounter to be the set of all possible styles. Furthermore, we might model the distribution of degradations that the system might encounter as the distribution over all possible inputs. In fact, that is the model that is implicit in many isolated character handwriting recognition systems, where the recognition system is trained on very large training sets derived from (hopefully) statistically representative samples of a large population of writers and response forms.

In practice, OCR systems are not presented with isolated characters, but with whole documents. It has been recognized in work on both handwriting recognition and OCR that an important property of handwriting and OCR problems is that the set of styles and the degradations that occur within a single document are correlated.<sup>5</sup> In different words, individual documents usually use only a small subset of styles  $S' \subset S$ , and the distribution of degradations  $D'$  encountered within a single document has a much lower entropy than that for the whole population of possible inputs. To simplify the subsequent analysis, we will assume that each document uses, in fact, only a single style; however, this does not affect the substance of the argument significantly.

Another issue that is relevant to both token based compression and OCR is that of segmentation of the input page into individual characters. For the purposes of this paper, let us assume that there is some bottom up procedure that reliably returns all the individual characters on an input page. For good quality input, heuristics based on connected component and line extraction can, in fact, do a good job at returning individual characters from the document reliably. For degraded input, however, significantly more complex procedures are needed. In fact, the segmentation problem itself can be formulated in a framework related to the clustering analysis we develop below, possibly allowing recognition-free bottom-up segmentation by preferring segmentations of the input that generate a sample distribution consisting of a small number of compact mixture components.

Finally, let us mention that another important aspect of the OCR problem is that of language modeling. That is, each document is not just a bag of degraded character images  $\tilde{B}$ , to be individually recognized, but it is a collection of sequential arrangements of degraded character images  $\tilde{B}_i$ . The sequences of characters  $c_i$  corresponding to these character images usually represent words and sentences in some natural language and therefore are subject to strong statistical regularities imposed by orthography and grammar. We will return both to the question of segmentation of degraded documents and language modeling briefly in the conclusions to the paper.

## 5. THE GAUSSIAN MIXTURE MODEL AND OCR

With these definitions and preliminaries, let us now look in more detail at the statistical structure of the OCR problem in a traditional framework in which training data (possibly incorporating style information) is used to train a Gaussian mixture classifier. By assumption,  $\tilde{B} = B_{c,s} + G(\Sigma_{c,s}, \mu_{c,s})$ , or  $P(\tilde{B}|c, s) = G(\Sigma_{c,s}, B_{c,s} + \mu_{c,s})$ , a Gaussian distribution with covariance  $\Sigma_{c,s}$  and mean  $B_{c,s} + \mu_{c,s}$ . This is the class conditional density for a particular character instance, determined by the character  $c$  and the style  $s$ . The sample distribution is given by the sum  $P(\tilde{B}) = \sum_{c,s \in C, S} P(\tilde{B}|c, s)P(c, s) = \sum_{c,s} G(\Sigma_{c,s}, B_{c,s} + \mu_{c,s})P(c, s)$ , a mixture of Gaussians. Finally, the posterior density is given by  $P(c, s|\tilde{B}) = P(\tilde{B}|c, s)P(c, s)/P(\tilde{B})$ . Since, for recognizing any particular character image  $\tilde{B}$  the probability  $P(\tilde{B})$  is fixed, maximizing  $P(c, s|\tilde{B})$  is the same as maximizing (over all  $c$  and  $s$ )  $P(\tilde{B}|c, s)P(c, s)$ , which is itself a pure Gaussian distribution in our noise model.

In the absence of any style information, for optimal recognition under a zero-one loss function, we wish to maximize  $P(c|\tilde{B})$ , which is proportional to  $\sum_s P(\tilde{B}|c, s)P(c, s)$ . That is, the class conditional density, as well as the posterior distribution, are both mixtures of Gaussians. Let us call this the global, style-unaware classifier  $P_0$ . This approach would therefore be equivalent to training a mixture-of-Gaussians classifier on a large collection of training examples representing all possible styles, an approach called mixture discriminant analysis (MDA).<sup>4</sup>

As we mentioned above already, we can improve on this result by taking advantage of the fact that each document contains only a single style (or, more generally, only a small number of styles). This means that we determine a single, overall document style  $s_d$  and then classify characters by maximizing  $P(\tilde{B}|c, s_d)P(c, s_d)$  (which is a single Gaussian distribution in our model) over all  $c$ . Because  $P(\tilde{B}|c, s_d)P(c, s_d)$  is only one term in the mixture  $\sum_s P(\tilde{B}|c, s)P(c, s)$ ,

the entropy of the style-specific distribution is going to be less and discrimination is (usually) going to be better using the style-specific approach. If the different styles that may occur are known in advance, the training data to such a system consists of a large collection of triples  $(\tilde{B}, c, s)$ , and we can easily build style-specific models  $P(\tilde{B}|c, s)$  or  $P(c, s|\tilde{B})$  using standard techniques. If the set of possible styles is unknown, training examples consist of triples  $(\tilde{B}, c, d)$ , where  $d$  is a document identifier. In that framework, the correspondence between document identifiers and styles is unknown (multiple document identifiers map to the same style), but it can be recovered using an expectation-maximization approach during training. For illustrations and a more in-depth discussion of this approach, see.<sup>5</sup>

## 6. THE SAMPLE DISTRIBUTION APPROACH

Let us now look at the OCR problem in the framework of a mixture model. Here, we assume that the input to the OCR system is data that is unlabeled and possibly from a different distribution of any training data available to the OCR system on prior occasions. This approach is based on the following observation. Under the assumption that each document represents only a single style  $s_d$ , the characters occurring in a document represent a sample from the sample distribution  $P_{s_d}(\tilde{B}) = \sum_c P(\tilde{B}|c, s_d)$ . This is itself a Gaussian mixture distribution,  $P_{s_d}(\tilde{B}) = \sum_c G(\Sigma_{c,s_d}, B_{c,s_d} + \mu_{c,s_d})$ .

The core idea presented in this paper is the following. Under the assumption that, within a style, a character is represented by a prototype  $B_{c,s}$  corrupted by Gaussian noise, each Gaussian that contributes to  $P_{s_d}(\tilde{B})$  corresponds to exactly one character. (More generally, we could assume that each character is represented by a small number of prototypes corresponding to a small number of mixture components.) Therefore, if we can recover the components of the sample distribution, which is the mixture  $\sum_c G(\Sigma_{c,s_d}, B_{c,s_d} + \mu_{c,s_d})$ , we obtain the class conditional densities and posterior distributions corresponding to each class, without ever having received any labeled training examples. We can use standard methods based on the EM algorithm for recovering the mixture components of the sample distribution. Of course, recovering the mixture components does not give us the class labels corresponding to each mixture component. That is, the mixture estimation algorithms will return a representation of the sample distribution in terms of Gaussians  $G(\Sigma'_i, \mu'_i)$ .

This mixture distribution is the analog of the  $\theta$ -cover of the sample distribution in previous clustering approaches. Each mixture component corresponds to a single  $\theta$ -ball in the cluster based approach. Because mixture estimation algorithms have principled, statistically valid ways of adjusting the mixture components (corresponding to the cluster centers and  $\theta$  parameter in the clustering model), we expect that the quality of the mixture model is going to be better than that of the clustering approach. While this seems plausible based on the analysis above, ultimately, that question needs to be resolved by experiment.

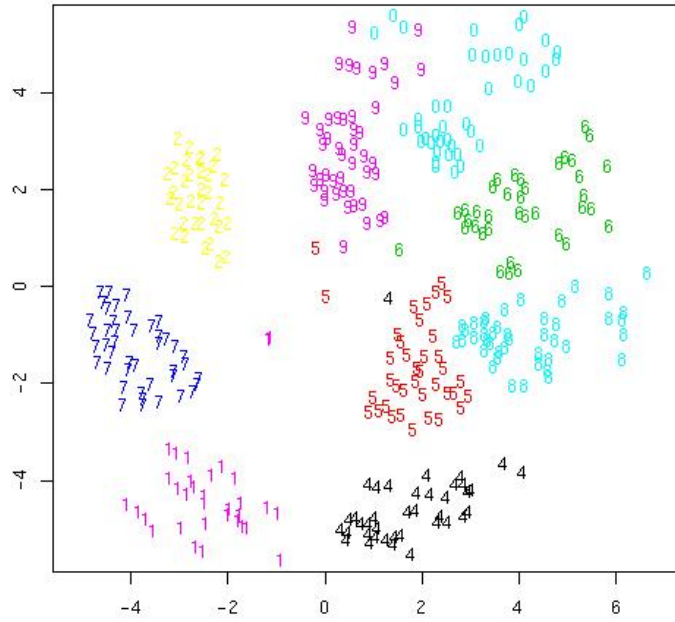
Just as in the clustering approach, still have to determine the correspondence between each mixture component  $i$  and a character classes  $c$ . In the clustering OCR systems<sup>2,3</sup> that motivated this paper, labels are assigned based on simple statistical or cipher-based language models. Such an approach could, of course, also be used within the mixture modeling framework describe in this paper.

However, another approach is to use a classifier that is not document specific (but may be style-aware in the sense of Sarkar<sup>5</sup>) to assign labels to most or all mixture components; remaining ambiguities could be resolved using statistical language models.

## 7. EXTENSIONS

Of course, in this discussion, we have made a number of simplifications. In a real system, the number of mixture components representing a particular class is going to be small, but probably greater than one. Furthermore, in some styles, some pairs of characters may be difficult to distinguish or even by identical (e.g., some typewriter styles use the letter “l” to represent the digits “1”). This means that the mixture estimation algorithm needs to be able to determine the Bayesian optimal number of mixture components automatically. Standard methods for addressing this problem have been described in the literature.

An objection to this approach (or any other kind of clustering approach) might be that algorithms for estimating mixture densities cannot reliably find mixture components that overlap strongly and may conflate such multiple mixture components into a single component. If these mixture components of the sample distribution represent contributions from different character classes, this could lead to a significant number of classification errors. However,



**Figure 1.** This example shows the result of a Sammon mapping of the feature vectors of a set of degraded characters into a 2D space. This mapping gives us a rough idea of the relationship of feature vectors in feature space. However, separability and cluster shape are better than suggested by this mapping (see text).

in such cases, there is a high intrinsic rate of classification errors anyway so that the overall impact on system performance is likely to be small.

More generally, we might ask why mixture components should be identifiable based on the sample distribution alone. We can make an argument for this based on the purpose documents serve. Documents are usually created by their authors to communicate text clearly. This implies that the class conditional distributions should not have significant overlap because otherwise the corresponding characters would be confusable. But if the class conditional distributions do not have much overlap, the sample distribution as a whole will consist of a mixture well-separated Gaussians.

A serious problem with clustering approaches is that specific characters may only occur in a few instances in a given document, or the document itself may only consist of a small number of characters. In that case, we would have insufficient information to recover good mixture density estimates from the given data, and characters might be assigned to mixture components representing different classes. The mixture-based statistical approach described in this paper provides two answers to this problem. First, we can take the approach of constraining the mixture density estimate for the document specific model by the decision boundaries for the non-document specific model  $P_0$ . That is, mixture density estimates that create mixture components that would assign significant probabilities to multiple decision regions in  $P_0$  are rejected. Second, we can take a smoothing approach, in which the mixture density estimate for a particular document is considered an interpolation between the non-document specific model  $P_0$  and the mixture density estimated from the empirical distribution of samples present in the document. Such a smoothing approach has the additional advantage that it will not penalize mixture components for character pairs that, in many styles, are confusable (e.g., “1” and “l”).

## 8. EXPERIMENTS

There are a number of experiments that should be carried out to demonstrate the utility of this approach. First, we would like to see that the approach yields good recognition performance on fairly hard problems, compared to state-of-the-art classification systems. Second, we would like to see that the approach, in fact, outperforms traditional clustering OCR approaches, at least under some conditions. Third, it would be good to test the statistical assumptions (in particular, assumptions about the distribution of the noise vector  $N$ ); while the Gaussian model is almost certainly more accurate than the bounded error model used in prior approaches to OCR by clustering, it may itself allow additional room for improvement.

To demonstrate the first point, a system based on the ideas described in this paper was implemented and applied to 1500 images of digits in the `cmr6` font from the Bell Labs database of severely degraded character images found on the University of Washington Database I.<sup>11</sup> Experiments were carried out interactively in the “R” statistical system.<sup>12</sup>

**Input Data and Feature Extraction** The input data was divided into 1000 training samples and 500 test samples. The images were centered, convolved with a Gaussian of  $\sigma = 1$  and subsampled to a size of  $10 \times 10$ . The resulting image was treated as a raw feature vector and 7 principal components were extracted. These PCA feature vectors were then used as input to a sample-distribution based classifier, as well as a mixture discriminant analysis-based classifier (MDA;<sup>4</sup>).

**Visualization of Clusters** As a first step, we might want to visualize the distribution of these feature vectors. The Sammon mapping<sup>13,14</sup> is a simple means for visualizing high dimensional distributions in 2D diagrams. It attempts to find a distribution of points in 2D such that each point corresponds to one of the high dimensional vectors and such that the distances between the points approximate distances between the high dimensional vectors as closely as possible. Figure 1 shows the result of generating a Sammon mapping for the feature vectors of a sample of digits. The mapping itself is generated without knowledge of the class labels; the class labels were added in the diagram to visualize the shape of the individual clusters. This mapping suggests that most of the samples are well separated into clusters, even if they are mapped only using their feature vectors, without corresponding class labels.

The Sammon mapping also suggests possible areas where confusions might occur. For example, in Figure 1 several samples corresponding to the digit “0” appear to have intruded into the cluster of feature vectors corresponding to the digit “9”. We can examine whether these clusters actually overlap by repeating the Sammon mapping (or some other projection) with just those two classes, and it turns out that all the apparent class overlaps are artifacts of the mapping into 2D. Note, incidentally, that the cluster shapes in the 2D projection also are a complex non-linear projection of the high dimensional shape; non-convexity of the projected shapes, for example, does not imply non-convexity of the actual distribution of feature vectors.

In summary, the Sammon mapping and other projection techniques can give us some intuition of how clusters in the (unlabeled) sample distribution correspond to class labels. From visualizations like these, it appears that most classes are well separated. However, such a characterization is only qualitative. In future work, it would be desirable to characterize the distributions and their separations more formally.

**Clustering OCR** The clustering OCR system performs its unsupervised clustering using the method described by Fraley and Raftery<sup>15</sup> and implemented by the `mclust` package for the R statistical system. Clusters in this approach are represented as Gaussian distributions. The method first performs hierarchical clustering and follows it by Expectation-Maximization (EM) steps to optimize the cluster shapes. In these experiments, the cluster shapes considered by the algorithm were “spherical” (all clusters have spherical covariance matrices), “uniform” (all clusters have the same covariance matrix), and “unconstrained”.

In this way, the clustering OCR represents the unlabelled sample distribution as a mixture of 15 Gaussians. By assumption of the method, each Gaussian corresponds to a single digit label. Traditionally, clustering OCR methods recover the labels corresponding to a cluster by using techniques analogous to cryptanalysis. While such an approach could easily be generalized to a Bayesian optimal recovery of cluster labels when OCR is performed on actual text, it is not applicable in this evaluation, since the training and test data just consists of random samples of characters.

An alternative approach for unlabeled data is to perform an assignment of class labels based on a non-specific (population) classifier. Concretely, we assume that, as in a traditional OCR system, we have a large population of labeled feature vectors that we can use to build a non-specific classifier (e.g., a mixture density classifier). We can use this non-specific classifier to assign labels to each of the unlabeled clusters that we obtained in the previous clustering step. Such an assignment of labels to individual feature vectors based on assigning labels to the clusters that the feature vectors results in lower error rates than using the non-specific classifier on individual samples, because (intuitively), there is a lot more information available for assigning labels to whole clusters reliably. An empirical and theoretical analysis will be presented elsewhere.

When the assignment of labels to clusters is correct (either based on cryptanalysis or based on a non-specific classifier), the error rate of the sample distribution based recognizer on test data is 0.7% (N=500) in these experiments.

**Mixture Discriminant Analysis** To compare the performance of the clustering OCR with a traditional approach to character recognition, a Mixture Discriminant Analysis (MDA) model<sup>4</sup> was trained. An MDA model represents likelihood functions as mixtures of Gaussians and uses Bayes rule to perform classification. The Gaussian mixtures are estimated using the Expectation Maximization (EM) algorithm. In the experience of the author, as well as based on results reported in the literature,<sup>4</sup> MDA performance is roughly comparable to the performance of other, commonly used classifiers like neural networks and radial basis function methods. The R implementation of MDA (available from the R web site) was used for the experiment. When the MDA classifier was trained on the training set (N=1000), its error rate on the test set was 0.6% (N=500)

## 9. DISCUSSION

Ideas of clustering and style in OCR are not new and have been explored by a number of authors explicitly or implicitly. For example, several authors have suggested using clustering of documents to recover high resolution representations of characters, both for improving document quality and OCR. And the application of clustering and mixture density estimation methods to pattern recognition problems in general, of course, also has a long tradition.

What this paper contributes is an analysis of the OCR problem from the point of Bayesian statistics, Gaussian mixtures, and mixture density estimation of the sample distribution. This view gives us clear guidelines for what properties are desirable for the clustering process, whether and how we should assign labels to clusters, and what kind of principled approach we should take in using the resulting clusters for discriminant analysis. It also gives us a roadmap and checklist for examining a particular clustering-based system from the point of view of OCR performance: we can look at the actual likelihood functions and sample distributions and verify that they satisfy our models. This seems far preferable to picking an ad-hoc clustering mechanism and attempting to make it work with a back-end OCR system.

The sample distribution-based approach presents an alternative to other style-constrained recognition methods. Existing approaches to style-constrained recognition assume that styles to be recognized are represented in the training sample, or at least that the set of allowable transformations leading to different styles can be characterized from the training set. The sample distribution based approach to style-constrained recognition, however, can work under a much wider set of conditions and cope with entirely novel transformations of the input. This corresponds to the human ability to cope with entirely new classes of font styles (c.f., for example, the appearance of sans serif styles, or popular styles related to the OCR-A and OCR-B fonts).

While the initial experiments described above are promising, clearly much work remains to be done to arrive at a practical OCR system based on the sample distribution based approach and demonstrate that approaching the problem of clustering OCR from the point of view of approximating the empirical sample distribution as a mixture distribution leads to higher performance in practice.

## REFERENCES

1. R. Casey, S. K. Chai, and K. Y. Wong, "Unsupervised construction of decision networks for pattern classification," in *Proc. ICPR-7*, July 1984.
2. R. Casey, "Text ocr by solving a cryptogram," in *8th International Conference on Pattern Recognition*, vol. 1, pp. 349-351, IAPR, 1986.

3. G. Nagy, S. Sharad, K. Einspahr, and T. Meyer, "Efficient algorithms to decode substitution ciphers with applications to ocr," in *8th International Conference on Pattern Recognition*, vol. 1, pp. 352–355, IAPR, 1986.
4. T. Hastie and R. Tibshirani, "Discriminant analysis by gaussian mixtures," tech. rep., AT&T Bell Laboratories, 1994.
5. P. Sarkar, *Style Consistency in Pattern Fields*. PhD thesis, Rensselaer Polytechnic Institute, May 2000.
6. Huttenlocher D., Felzenszwalb P., and Rucklidge W., "Digipaper: a versatile color document image representation," *Proceedings of 6th International Conference on Image Processing (ICIP'99)*, pp. 219–23 vol.1, 1999.
7. Haffner P., LeCun Y., Bottou L., Howard P., Vincent P., and Riemers B., "Color documents on the web with djvu," *Proceedings of 6th International Conference on Image Processing (ICIP'99)*, pp. 239–43 vol.1, 1999.
8. J. Hobby and T. Ho, "Enhancing degraded document images via bitmap clustering and averaging," in *International Conference on Document Analysis and Recognition*, 1997.
9. T. Hong and J. Hull, "Improving ocr performance with word image equivalence," in *Symposium on Document Analysis and Information Retrieval*, pp. 177–190, 1995.
10. K. T, B. H, and H. R, "Estimation and validation of document degradation models.," in *Proc. 4th Annual Symposium on Document Analysis and Information Retrieval*, (Las Vegas, NV), April 1995.
11. "University of Washington Document Image Database I."
12. R. Gentleman, R. Ihaka, and other contributors, "The R statistical system at <http://www.r-project.org/>."
13. J. W. Sammon, "A non-linear mapping for data structure analysis.," *IEEE Transactions on Computers C-18*, pp. 401–409, 1969.
14. W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-Plus (3rd ed.)*, Springer Verlag, 1999.
15. C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis.," Tech. Rep. No. 329, Dept. of Statistics, U. of Washington, February 1998.