Document Style Census for OCR

George Nagy
DocLab, Rensselaer Polytechnic Institute
Troy, NY, USA 12180
nagy@ecse.rpi.edu

Prateek Sarkar
Palo Alto Research Center
Palo Alto, CA, USA 94087
psarkar@parc.com

Abstract

Four methods of converting paper documents to computer-readable form are compared with regard to hypothetical labor cost: keyboarding, omnifont OCR, style-specific OCR, and style-constrained or style-adaptive OCR. The best choice is determined primarily by (1) the reject rates of the various OCR systems at a given error rate, (2) the fraction of the material that must be labeled for training the system, and (3) the cost of partitioning the material according to style. For large corpora, sampling strategies are proposed both for estimating conversion costs and for taking advantage of style homogeneity.

1. Introduction

The cost of scanning a collection of paper documents is fairly predictable, but the cost of converting the resulting bitmap images to a searchable form (i.e., by OCR and keyboarding) is not. We discuss issues related to estimating the cost of conversion, including the type of data that must be collected to plan and execute such a conversion and the choice of conversion methodology. We call the selective collection of relevant data a *document census*. As in the case of a demographic census, both exhaustive enumeration and sampling are required.

The conversion of documents to digital form is of interest for increasing accessibility to both content and form. Broad access to the latter is the primary purpose of *document preservation*, which usually targets relatively few but precious documents. The problem of preserving important historical documents is similar to that of producing high-quality facsimile editions. Even when this is accomplished with digital scanners or digital cameras, as opposed to analog reproduction (i.e., film photography), the production of a transcription is secondary to the retention of precise image detail. (Of course, however elaborate, facsimiles cannot be expected to preserve some aspects, like paper thickness or ink composition.) We consider here only documents where the *content* is of interest, rather than the document artifact itself. Document *format* is another matter: it is generally agreed that at least some format information must usually be retained for effective access to content.

In many collections, the separation of text from illustrations is difficult. Since we are now investigating primarily the OCR aspects of document conversion, we concentrate on mainly-text documents, and assume that automated algorithms combined with human interaction accomplish text-graphics separation. We also neglect contextual methods based on language or application-specific models, which can greatly benefit all the methods discussed herein.

2. Specialized document collections

The vast majority of all documents on the web are freely accessible. A library can also offer its clients some exclusive collections,

Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)
0-7695-2088-X/04 $20.00 © 2004 **IEEE**

such as professional society publications, specialized bibliographies or reports, and collected works, biographies or critiques of a particular author. Such material is more often relevant to the world of scholarship than to that of commerce or entertainment.

These restricted collections, most often owned by non-profit organizations, are among the most valuable nuggets on the web. Many are derived from a single pre-existing collection, have a cohesive theme, and ample supportive metadata. The original hardcopy collections were typically assembled by an organization or a person with a specific interest, and such collections are often more selective than comprehensive. In size they may range from a single volume, such as *A Flora of California* [Jepson 43], or the complete works, including multiple editions, of a particular author, to subject-specific collections or lesser of greater specificity of topic, time, and place.

Current endeavors in digital libraries strive to put more such collections "on line" [Baird 03, GT02]. It is, of course, far more expensive to convert existing paper collections to digital form than to post current publications – novels, newspapers, and technical journals – that already have some digital incarnation. Furthermore, *all* (nonprofit) libraries are poor, at least in the sense that whatever funds they use for conversion encroach on funds for acquisition, conservation, and client services.

The relative homogeneity of the content of collections of interest does not imply homogeneity with regard to the characteristics that determine the cost of conversion. As will be seen, this cost may depend on the extent that the documents can be grouped according to *OCR style*. We attempt to provide an analytical framework for investigating the relationship between conversion cost and OCR style.

We define *style* as a grouping of character images (*glyphs*) that increases the average ratio of between-class to within-class separation in feature space within a style-homogenous group. There is a useful distinction between weak style and strong style [Sarkar 00, SB03]. *Weak style* partitioning decreases the average within-class distance. For example, all the glyphs for each symbol can be clustered into similarity groups, with clustering carried out independently for each symbol. Significant new ideas for exploiting weak style are reported in [MB 02, Breuel 03]. *Strong style* increases the between-class distances in addition to reducing the within-class distances, thereby reducing confusions between glyphs belonging to different symbols *and* different style categories. It exploits which character styles co-occur with which other character styles. Examples of strong grouping include pages in the same font, handwriting by the same writer, and reproduction with a specific copying machine.

3. Omnifont vs. style-specific recognition

Before discussing style-constrained and adaptive OCR, we consider three conventional approaches to entering documents:

1. Manual key entry;
2. Omnifont OCR;
3. Style-specific recognition.

One objective of a document census is to partition the collection into subsets for the most economical application of each method. Given the distribution of documents by style-equivalent pages, and the cost of processing pages of each style by all three methods, one can choose the least expensive method for each page.

Manual entry consists of simply keyboarding the entire collection, followed by whatever verification is appropriate – including multiple entry – for the specified error rate. In the absence of strict time and

security constraints, manual data entry is often contracted out to take advantage of lower wages abroad.

For the OCR options, we consider only OCR systems where a reject threshold can be set to guarantee that the residual error is below some specified value. For example, if the maximum tolerable word error is 0.5%, then on a tightly typeset, low-contrast page the reject threshold may be set to reject 10% of the words, while on easier material it would be set at a 5% word reject rate.

Omnifont OCR requires setting a higher reject rate for the same error rate than style-specific OCR tuned to each category. In either case, the rejects must be entered manually.

Style-specific OCR requires keyboarding a sufficient fraction of each targeted category to provide representative samples of at least the frequently occurring glyphs (letters, digits, punctuation, and special characters) for tuning the classifier to each (strong or weak) style. Symbol distributions are highly skewed, and the total number of classes varies from about 100 in straight prose to over 500 in technical material like patent applications. Therefore the volume of material required for tuning a classifier may vary from a few paragraphs to many carefully selected pages.

To decide which method is most appropriate, we need to know the distribution of documents by style, i.e., the amount of text in each category. We must also know the relative costs of manual entry for each style (keyboarding material with limited context is more expensive than plain English text). We assume that *preprocessing* consists only of entering the transcript of a sufficient amount of each category, and that the cost of automatically tuning (retraining) the classifier given such a transcript is negligible. To estimate data-entry costs incurred in preprocessing, we need to know how much material must be entered for each style, but we assume that the keyboarding cost per word is that of manual entry of the same style. Finally, we must take into account the cost of *post-processing*, i.e., that of manually entering all rejected words. As a first approximation, here too we assume that the cost per word is the same as that of manual entry. (For a better approximation, we can assign a higher cost to entering rejects than to entering straight text.)

Given the reject/error curve of the classifier and a maximum acceptable error rate, the above costs are sufficient to determine which method to use for any distribution of documents into style-based categories. However, we can make the further simplifying assumptions that all the categories contain the same number of pages, and that the relevant costs are the same for each style. This yields the simple result that style-specific recognition with tuning is cheaper than multifont recognition if

$$K > \Delta R$$

where K is the ratio of the amount of necessary training text to the total volume of each style, and $\Delta R = R_s - R_m$ is the difference between the multifont and style-specific reject rate at the specified error rate. Manual entry is required only when

$$K+R_s \approx 1 \text{ or } R_m \approx 1.$$

In the above analysis, the cost of all automated steps is set to zero, and the cost of correcting residual errors is considered the same for multifont and style-specific classification. It is not, of course, necessary to physically group all the pages of each category. Once the scanned page bitmap files have been style-labeled, the necessary preprocessing, classifier training, OCR, and reject entry can be ordered by scripted file manipulation. We shall present a more complex model of cost comparison in Section 6.

4. Style-constrained and adaptive recognition

The scenarios we considered above are the conventional ones: both omnifont and style-

specific ("single font") training and recognition have been studied and applied for decades. The essential requirement for the style-specific approach is to maintain the correspondence between the equivalent document classes in the training set and in the field application. If the first several pages of a particular edition of Moby Dick are manually entered, then when the remaining pages are processed, the OCR system must know that it *is* Moby Dick.

It is possible, however, to improve on omnifont accuracy even without preserving such correspondences. In *style-constrained classification*, the training set is divided into style-homogenous batches, and the statistical characteristics of the individual batches, as well as their relative volumes, are estimated. When a batch of new material is to be recognized, the system automatically applies the classification most appropriate to the style characteristics of the new text. The correspondence between the style of the new material and a specific part of the training data need not be explicitly specified.

Style-constrained recognition is generally not as accurate as style-specific recognition with a classifier trained to each "font." It is, however, more accurate than multifont recognition. Under sensible assumptions, it can be shown that it converges to the style-specific error rate as the number of training and test samples of each style grows to infinity.

Style-constrained recognition has not yet been applied to any digital library conversion task, or indeed to any practical OCR task. It is still at the research stage. In the following, we give an intuitive account of four different variants, each of which is appropriate under different conditions. We reference technical articles and two doctoral dissertations in support of our advocacy. A broader view is presented in [SB03]. After this essential digression, we will return to document census for data collection.

Optimal style classifier for continuous styles and very short fields. The conventional quadratic classifier for Gaussian distributions takes into account the statistical dependence between the features of each pattern. Style-constrained classification makes use, in addition, of the between-class dependence between the mean values of the features of *several* patterns of the same style. Intuitively, this means that we expect the shape of a pattern of class i reveals something about the shape of a pattern of class j from the same style. Formally, instead of a $d \times d$ covariance matrix, where d is the number of features, the classifier makes use of a $d^s \times d^s$ matrix, where s is the number of patterns in a single-style field [VN01, VN02a, VNLF02, Veeramachaneni 02].

On NIST hand-printed digits, the style-constrained optimal quadratic classifier parameters were estimated from pairs of digits from several hundred writers. 100 blurred directional features were used [FL03, WF78]. The resulting classifier reduced the field error rate on new writers by 10 to 25% as the field size was increased from 2 to 5 digits, compared to a conventional singlet classifier. (Even for longer fields, the parameters are estimated only from pairs of patterns.). Such a classifier might be useful under the weak assumption that the font never changes within a word.

Optimal classifier for discrete styles and short fields. Instead of assuming a Gaussian distribution of style means, as above, it may be appropriate to consider a discrete number of styles with Gaussian feature distributions and arbitrary mean vectors [Sarkar 00, SN00, SN01]. A linear discrete-style classifier (modeling independent Gaussian features) was tested on a set of typeset digits of five different typefaces, using only naïve features, with fields of 13 same-style digits used for training and fields of 4 same-style digits used for testing. On this data, the discrete style-constrained classifier based on diagonal covariance matrices had less than

half the error rate of a Gaussian singlet classifier with a diagonal covariance matrix. It was about 25% better than a singlet classifier that modeled the same number (5) of Gaussians (one mode for each typeface). The improvement over the multi-modal singlet classifier is due to the fact that the four patterns in a field yield a better estimate of the probability of the parent style of each field than a single pattern can. The style-constrained classifier was about 10% worse than separate classification of each typeface with a Gaussian classifier trained only on that typeface (which is optimal when the training-set – test-set correspondence is known).

Theoretical arguments indicate that under the usual assumptions the discrete style-constrained classifier is more effective than font classification followed by style-specific classification.

Adaptive optimal classifier for continuous styles, long fields. When many samples of each style must be classified, as in longer documents, then it is possible to re-estimate the means and covariances of the classes from the *unlabeled test samples* via either maximum likelihood (ML) or maximum a posteriori (MAP) estimation by Expectation Maximization [Veeramachaneni 02, VN02b, VN03]. On test fields of hand-printed NIST digits, ML estimation yielded an average decrease in error rate of a factor of two on fields of 60 digits or more from the same writer [VN03]. MAP adaptation is significantly better on shorter fields (10-50 digits), but is essentially the same as ML on longer fields (because of the decreased weighting of the original class-conditional distributions). The number of digits per field in this case allowed only unsupervised re-estimation of the means and variances, but not of the covariances.

On isolated printed characters, long-field adaptation was more than ten times as effective at reducing the error than style-constraints applied to fields of three characters [VN02b]. Theory and simulations show that with long-enough fields, long-field adaptation converges to the optimal quadratic classifier trained on patterns of the same source as the test set. On the NIST data, the maximum number of about 100 digits per writer was insufficient for estimating the writer-specific covariances of 100 features, yet the twofold reduction in error rate is significant. For the much longer documents in digital libraries, this method provides a trade-off between manual data entry and increased computation that is bound to become economical as the cost of machine cycles continues to decrease.

Heuristic algorithms. The simplest method of taking advantage of unlabeled characters is to OCR an entire style category of documents with an omnifont classifier, then use the labels assigned by this classifier (some of which will be incorrect) to retrain the classifier. The retrained classifier is then used to OCR the same set of characters again. Retraining followed by classification may be iterated several times, but most of the benefits usually accrue on the first iteration. On accurately segmented single-font alphabets in twelve fonts, reductions in error rate of a factor of 5 were obtained [NS66, Nagy 68]. The results were replicated on a larger data set generated by a pseudo-random defect model [BN 94].

This *self-correcting* idea has recently surfaced in the machine learning community as bootstrapping for active learning. (*Active learning* is the notion of using the classifier for selecting specific samples to be labeled by an operator and used for classifier retraining.) It has been successfully applied to text categorization and information retrieval [NMTM00, WCZ03]. It must be noted, however, that this method is suboptimal for Gaussian distributions with unequal variances.

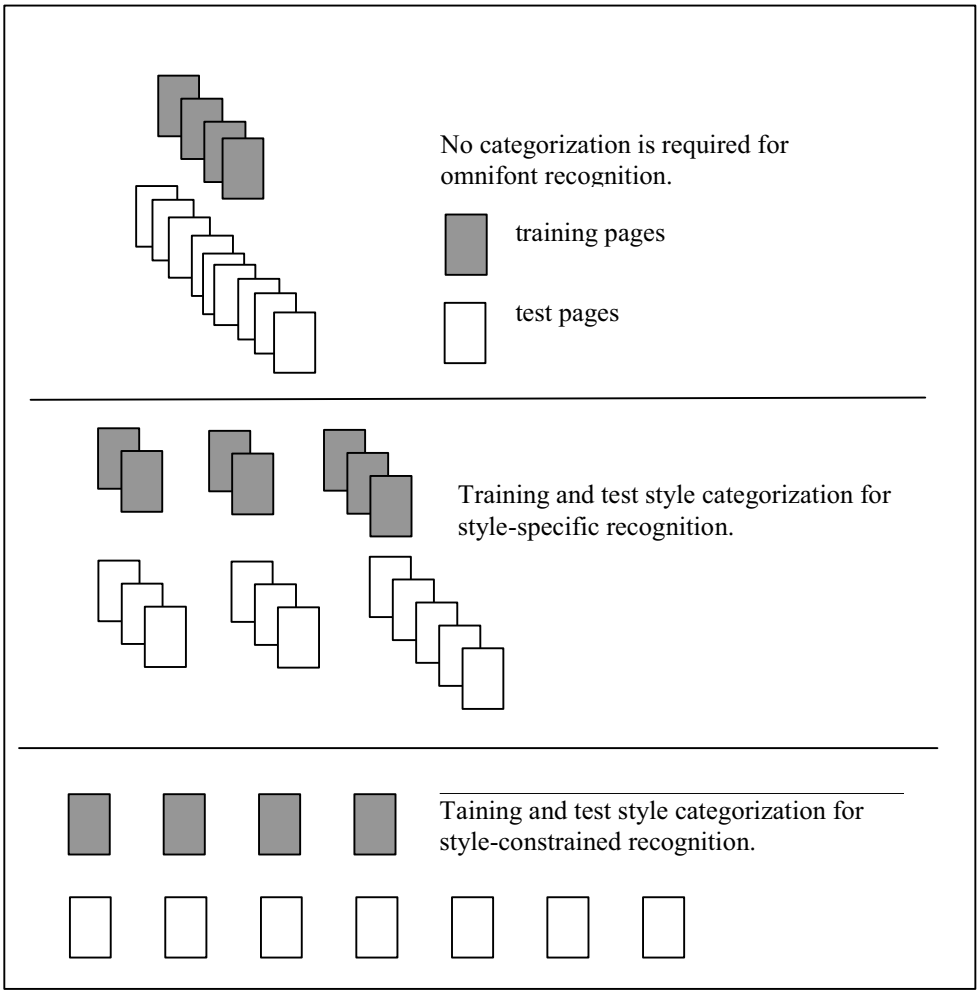

Figure 1. Categorization for different recognition methods.

Another heuristic algorithm classifies longer fields as overlapping pairs of patterns, using any available pair-classifier that yields some confidence value [VLFN02]. The scores of all the pairs are then combined for the highest field score. We have not yet found an optimal method of combining the scores even under the restrictive assumption that they represent the probability of correct pair recognition, but several different combination rules do reduce the error rate over that of singlet classification. In some ways, this may be considered as a type of *classifier combination*, which was first used with commercial OCR systems in the late eighties [Bradford 91, BN91, HH91], and became a popular topic of research in the nineties [HHS94, Ho 02].

5. Style categorization

Both style-specific OCR and style-constrained OCR require partitioning the training set into style-homogeneous subsets. The style-specific method also requires labeling the page bitmaps to be recognized consistently with the training set. In contrast, style-constrained methods need only partitioning the bitmaps into style-homogenous categories, without any style labels. Manual entry and omnifont OCR require no style categorization at all. The

differences between the partitioning methods required for the three OCR methods are illustrated in Figure 1. For style-specific recognition, all of the data must be consistently style-labeled. Labeling all of the data (part of which will serve as the training set and must eventually be labeled by character class as well as style) is difficult and expensive. Font recognition [ZI97] trained on a broad variety of typefaces and sizes may help, but the style categories for the most efficient style-specific classifiers don't necessarily correspond to conventional typefaces.

For style-constrained recognition, it is sufficient to partition all of the data into style-consistent subsets. It is not necessary to determine which partitions share the same style, only that each partition is style-homogenous. The size of the partitions is application dependent. For instance, the text of an entire novel may sometimes be considered style-homogenous, but in a 1960's volume of typed conference papers (provided by the authors in camera-ready form), only one paper may be assigned to each partition. In a typeset technical journal, most lines of text will be style-homogenous, but a whole page will have header, title, author, affiliation, and footnote style lines. Some reference works (dictionaries, telephone books, flower guides) will need elaborate document analysis to separate out the single-style components. In many cases, this will be more expensive than the OCR itself. *Style-constrained recognition is most applicable when there is a clear, algorithmically definable partition (by document, page, isolated text block) into homogenous style categories.*

A style-constrained classifier can group elementary partitions by style automatically during the training process. Adaptive classification does not require partitioning the training set at all, but the data to be classified must still be grouped into homogenous sets. With either method, the characteristics of each style-homogenous partition in the test data, rather than of an isolated character, determine how each character is classified. As mentioned earlier, the separation of training or test data into elementary same-style categories may be accomplished most economically by interactive or automatic processing of the scanned bitmaps, rather than of the physical documents.

The notion of style consistency in printed matter applies not only to typefaces and sizes, but also to printers, copiers and scanners. Some of the attributes that have been suggested to affect OCR are contrast, speckle, skew, typeface and size, line density, format complexity, and layout uniformity. With current technology, scanner and printer distortions are minor compared to iterated copying, except for the inevitable edge noise due to the random phase of spatial sampling [SNZL08].

The cost of labeled categorization increases the cost of style-specific recognition. Unlabeled categorization, while cheaper, adds to the cost of style-constrained and adaptive recognition. Automated style categorization is therefore a research priority.

6. Analysis of quality control costs

In Section 4, we stated a simple rule of thumb for choosing between two different methods of OCR. Here we shall introduce a few concepts relevant to digitization workflows for large volumes of data that are typical for digital library projects, with emphasis on quality control.

The primary factors that affect the degree of complexity and cost in an OCR project are

1. *Complexity of layout*. Simple single or multicolumn text, with simple reading order and no non-text objects are easiest, while pages with pictures, drawings, tables, forms are difficult to process and prone to high error. We shall focus our attention on the former - typewritten documents, books on history, literature,

are examples of documents with primarily textual content.

2. *Typestyles, typesizes, linespacing.* Frequently changing typesetting formats makes both layout segmentation, and OCR difficult and error-prone.

3. *Image quality*. Non-standard scan resolutions, scanner brightness and contrast settings, page skew and bending, paper quality and age, non-print marks, and ink-spread adversely affect OCR error rates.

If the document content is in a familiar linguistic domain, both OCR and operator keyboarding can be more accurate and efficient. However, the cost of skilled operators or sophisticated language models would have to be taken into account. We ignore the content domain factor in our analysis because this can affect all OCR methods equally.

It is well recognized that the cost of document conversion from paper to digital medium is dominated by quality control costs, especially when the goal is to digitize content rather than simply scanning to image. In a quality controlled scan-OCR workflow the major costs are:

1. Document type detection (language, script, skew, noise level – any information that downstream processes, including pre-processing, may use).
2. Text-zoning quality evaluation
3. Zoning correction and reject processing (manual zoning).
4. OCR quality evaluation.
5. OCR correction and reject processing.
6. Workflow and system maintenance.

Factors 1 through 5 can account for almost all of OCR cost once the workflow system has been setup. With any state of the art zoning software, factors 3 & 4 dominate this cost for most simple text-only document layouts.

In a typical high volume OCR task, it is not sufficient to specify quality requirements as an aggregate character or word error rate over the entire collection to be digitized. Since OCR errors are bursty, often smaller units (such as documents, articles, or pages) in the collection get poorly OCRed even as the aggregate error rate stays low. When OCR output is used for indexing and retrieval of these smaller units, poorly OCRed units are virtually lost.[1]

Therefore, error rates are measured for each unit, and it is required that less than x% of units should have error rates higher than y%. For example, in a digital library project x and y could be 1.0 and 2.0 respectively. From now on we shall consider pages as units, without loss of generality.

In the face of such two-tiered quality requirements, systems for thetriage of pages from OCR output, such as in[SBH02], have to be implemented to translate confidence levels output by the OCR engine (at the word or character level) to a page level confidence. As a result the error vs. reject relationship can be extremely complex.

We shall now present a simple model for purposes of an intuitive evaluation of alternative OCR methods. The following is a summary of notations.

- P: Number of pages to be OCRed.
- N: Number of words per page (we base our simple analysis on average quantities, rather than ranges or probability distributions).
- k_c: Cost, per word, of continuous manual keyboarding, as in typing in a page of a novel. Costs may be measured in monetary units.
- k_d: Cost, per word, of discontinuous manual entry, as in OCR error correction.
- k_e: Cost, per word, of OCR evaluation (quick visual comparison of image and corresponding OCR output).
- c_m: Cost of style markup (chunking of style-homogeneous sections, style-

[1] There have been attempts to model OCR error patterns, so that indexing could be made relatively more robust to OCR errors. See, for example, [MHFS97].

labeling of such sections if necessary) expressed as a per word cost. Unlike the previous three costs (denoted by $k_?$), this cost is much more dependent on the complexity of pages, and the OCR system.

- T: fraction of pages accepted by triage[2] *without* manual evaluation or correction). For any given required quality standard, this is a function of the OCR system, and the difficulty of the task at hand.
- e: residual error rate, i.e., average error rate on pages assigned by triage to be checked and corrected. Therefore e is higher than the threshold error rate set for quality control.

Operator costs k_c, k_d, k_d, and c_m incorporate multiple corrections/entry by independent operators that may be necessary to meet extremely high quality requirements. Typically $k_e < k_c < k_d$. Style markup cost can be extremely low if pages are style-homogeneous, if style homogeneous chunks can be identified automatically by geometric analysis, or if page-zoning segments are themselves style-homogeneous. On the other hand if style-partitioning and style-labeling of such partitions require heavy operator assistance, c_m can be much higher than the highest typing cost (k_d).

Cost of full manual entry of P pages of text with N words/page is then

$$k_c \,.\, N \,.\, P.$$

Cost of OCR assisted conversion without further training of system is:

$$(k_e + k_d.e) \,.\, N \,.\, (1\text{-}T).P$$

where e is the average error rate on pages that require further processing. When a fraction, f_{Tr} of pages are required to for training or adaptation, this cost is:

$$[f_{Tr}.k_c + (1\text{-}f_{Tr}).(k_e+k_d.e).(1\text{-}T) + c_m] \,.\, N \,.\, P$$

Here we assume that labeling the training data is as costly as continuous manual entry. This is true if it is sufficient to simply provide the transcription of the content of training pages: the training algorithm automatically performs the necessary alignment of text-images and transcription (extremely costly to provide manually). DID [Kopec94, SBZ03], HMM based methods [LBKMNS99], and template alignment methods [NX99] are examples of such methods. It may be appropriate to include a sampling cost to represent time and skill required to carefully create representative training data, but we ignore that. Some adaptive systems may bootstrap themselves without further manual groudtruthing. Our costs formulations ignore costs of transportation and storage of physical documents.

When the choice is between full manual conversion and zero-training omni-font OCR, the cost comparison comes down to

$$k_c <> (k_e + k_d.e)*(1\text{-}T).$$

With appropriate operator interfacing, k_e is negligible compared to k_c. So we check if

$$k_c <> k_d*e*(1\text{-}T).$$

If we find, for a given performance threshold, that 50% of pages in a reasonable sample can be exempted by triage from correction, leaving behind pages with an average of 20% errors on them, then complete manual entry is justified if correction cost per word exceeds 10 times its continuous entry cost. Of course this is a very rough calculation based solely on average values of each quantity.

When style markup cost is negligible a brief manipulation of the formulae reveals that the same criterion holds in deciding between full manual entry and trained/adapted OCR. Indeed, after a fraction of pages (training fraction) is manually entered in both cases, the nature of the decision is the same as in the previous paragraph. Only the volume of the task shrinks to $(1\text{-}f_{Tr})*P$ pages. However, when the system is tuned to the style of documents to be OCRed, we would expect a higher rate of triage (e.g., 80%), and a lower average residual error rate (e.g., 5%). Under these situations, style-specific OCR would be better unless correction costs exceeded direct entry costs by a factor of 100!

[2] OCR triage means the (automated) classification of pages after OCR according to subsequent processing requirements.

Table I. Summary comparison of various factors affecting OCR cost

	e	T	c_m
Full manual Entry	-	0	-
Omnifont OCR (no training)	HIGH	LOW	-
Style-specific OCR	LOW	HIGH	HIGH
Style constrained OCR	LOW	HIGH	LOW

When style markup represents significant costs, the comparison of omnifont OCR with style-specific OCR also involves careful consideration of the systems' respective triage rates and residual error rates.

The choice between style-specific OCR and style-constrained OCR is more subtle to analyze. Style specific OCR can be expected to have higher triage rate, and a lower residual error rate. But it also requires skilled operators for manual identification of an appropriate number of styles in data, collecting sufficient training data for each style, and categorizing of each page (or unit) of data by style-label. This results is high style markup costs, c_m. Especially if the units for quality evaluation (and therefore triage) (typically pages or documents) are not style homogeneous, as in documents formatted with multiple fonts and type-sizes, then style-specific OCR quickly becomes expensive. Style-constrained OCR benefits in the comparison because it does not require style-label assignments for each automatically or manually marked-up unit.

Choice of manual vs. OCR assisted conversion can also be affected by logistical considerations, such as different physical locations of computer farms and manual operators, so that shipping data back and forth at various stages of workflow may become impractical.

Given the nature of human intervention tasks, it is possible to form approximate intuitions about the relative costs of different operations from empirical studies in human computer interaction tasks [CMN83, EN81]. For example it is known that experienced typists can type continuous text thrice as fast (keystrokes/sec) as random disconnected text [CMN83]. After locating an OCR error, correction may involve moving a mouse or pointer to the location of an error, before typing in the correct text. This could make the cost of disconnected typing 10 times slower than continuous typing. Empirical cost estimates directly related to scan conversion can also be obtained from cost analyses of digital library projects such as the Million Book Project. Project Gutenberg has launched a successful effort to harness volunteer power online for distributed full manual keyboarding/correction of texts over the internet.

7. Sampling

The difference between academic and "real" character recognition is that in a research setting, the training set is usually much larger than the test set, while the training sets of even the largest OCR companies are much smaller than the amount of data processed by their customers.

OCR tends to work best when the training data is statistically representative of the test data. Digital library conversion tasks facilitate style-based recognition because, in principle at least, all of the hardcopy data is available before the start of the conversion. Therefore a representative subset for training can be drawn by sampling. The chosen sampling scheme can be either a "flat" design, or a "stratified sampling" design. We are interested in three key parameters about a set of documents: the *OCR triage rate* at bounded error, the *cost of manual entry,* and the cost of *style partitioning*. One option is to take a sample of the documents, scan the sample, then group the bitmap files by style, and both OCR and manually enter the samples. There are two problems: how to obtain a truly random sample, and how to estimate the triage rate and the reject entry cost.

Numbering every document explicitly and each page in every document implicitly solves the first problem. Then each page has a unique page number. Sampling is performed according to page numbers generated by a pseudo-random number generator that produces the required number of integers uniformly distributed within the page range. The solution to the second problem depends on the sampling scheme. If the sample is truly random (i.e., *flat*), then the reject rate and the reject entry cost on the sample will be representative of the entire collection. If a stratified sampling scheme is used, then the triage rate of each document category is weighted according to the number of documents of that category (other weighting schemes for finding population statistics with stratified sampling exist). Some possible sampling units for different types of documents are shown in Table II. We have more confidence in predicting triage and error rates by sampling than by attempting to analyze measurements on digitized pages. Past and current research on image measurement and filtering to predict and decrease OCR error is discussed in [BKNG95, CHK99, Summers 03].

7. Conclusions

Until we can automatically convert all documents accurately into searchable formats, the cost of human interaction will dominate the cost of conversion. It is therefore essential to direct human interaction to where it is most effective. We

Table II. Document sampling for style-based recognition

Document type	Sources of variations	Sampling unit	Examples
Book	typeface and size, bound/unbound	volume	novels
		page	conference proceedings
		paragraph block	text books, statutes
		sub-paragraph	dictionaries, telephone books
Journal, magazine	typeface and size	single issue	old philosophy journals
		paragraph block	title, abstract, citation
Typescript	contrast, pitch, typeface and size, bleed-through, copier	complete document	technical reports
		page	business/ private letters
Manuscript	writer, ink/pencil, paper, bleed-through, copier	complete document	diary
		pages by same writer	letters by a single writer
		pages of same letter	multi-source business/private letters
		documents by same census taker	19th Century census forms
		same-clerk documents	early real estate records

have advocated research on "non-supervised" classification for OCR for many decades. We are gradually developing evidence that when a collection of documents can be partitioned into style-homogenous fractions, it is cost-effective to apply style-constrained and adaptive classification techniques that take advantage of the style information distributed throughout the documents.

Acknowledgments

We gratefully acknowledge the long-standing technical and financial support of our OCR research by the Hitachi Central Research Laboratory. Most of the substantial theoretical and experimental results underlying the opinions stated here are due to past graduate students of DocLab, especially Dr. H. Veeramachaneni. We are also thankful to Dr. Henry Baird for some useful pointers.

REFERENCES

[Baird 03] H. S. Baird, Digital libraries and document image analysis, *Procs. ICDAR-03*, Edinburgh, pp. 2-14, August 2003.

[BKNG95] L. Blando, J. Kanai, T. Nartker, J. Gonzalez, Prediction of OCR accuracy, *Procs. Third Int'l Conf. on Document Analysis and Recognition,* pp. 319-322, 1995.

[BN94] H.S. Baird, G. Nagy, A Self-correcting 100-font Classifier, *Proc. SPIE Conference on Document Recognition, Volume SPIE-2181,* pp. 106-115, San Jose, CA, February 1994.

[BN91] R. Bradford, T. Nartker, Error correlation in contemporary OCR systems, *Proceedings of the First International Workshop on Document Analysis Systems, ICDAR01, St. Malo, pp. 516-523, 1991.*

[Bradford 91] R. Bradford, Technical factors in the creation of large full-text databases, *DOE Infotech 91 Conference*, May 1991.

[Breuel 03] T. M. Breuel, Character recognition by adaptive statistical similarity, *Procs. ICDAR-03,* Edinburgh, pp. 158-162, August 2003.

[CHK99] M. Cannon, J. Hochberg, P. Kelly, Quality assessment and restoration of typewritten document images, *Int'l J. of Document Analysis and Recognition*, Vol 2, #2-3, 1999.

[CMN83] S. K. Card, T. P. Moran, A. Newell. The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Inc. Publishers, 1983.

[EN81] D. W. Embley, G. Nagy. Behavioral Aspects of Text Editors, ACM *Computing Surveys* 13, #1, pp. 33-70, March 1981.

[GT02] D. Greenstein, S.E. Thorin, The Digital Library: A Biography, Council on Library and Information Resources, Washington DC, Second Edition, December 2002.

[FL03] H. Fujisawa, C-L Liu, Directional pattern matching for character recognition revisited, *Procs. ICDAR-03*, Edinburgh, pp. 584-598, August 2003

[Gutenberg] Project Gutenberg. http://www.gutenberg.org.

[HH91] J. Handley, T. Hickey, Merging optical character recognition outputs for improved accuracy, *Procs. RIAO 91 Conference*, pp. 160-174, 1991.

[HHS94] T.K. Ho, J.J. Hull, S.N. Srihari, Decision combination in multiple classifier systems, IEEE Trans. on Pattern Analysisss and Machine Intelligence, PAMI-10, 1, pp. 1067-1079, January 1997.

[Ho 02] T.K. Ho, Multiple classifier combinations: lessons and next steps, in A. Kandel, H. Bunke, (eds.), *Hybrid Methods in Pattern Recognition* , World Scientific, pp. 171-198, 2002.

[Hughes03] L. Hughes. The Price of Digitization: New Cost Models for Cultural and Educational Institutions. Report on National Initiative for Networked Cultural Heritage Symposium, April 2003, New York City. http://www.ninch.org/forum/price.report.html

[LBKMNS99] Jointly with Z. Lu, I. Bazzi, A. Kornai, J. Makhoul, P. Natarajan, R. Schwartz. A Robust, Language-Independent OCR System. In: Robert J. Mericsko (ed): Proc. 27th AIPR Workshop: Advances in Computer-Assisted Recognition SPIE Proceedings 3584, 1999.

[Jepson 43] Willis Linn Jepson, A Flora of California, Associated Student Store, Berkeley, CA 1943.

[MBP] Million Books Project. http://zeeb.library.cmu.edu/Libraries/LIT/Projects/1MBooks.html

[MHFS97] K. Marukawa, T. Hu, H. Fujisawa, and Y. Shima, "Document retrieval tolerating character recognition errors -- evaluation and application," Pattern Recognition, vol. 30, no. 8, pp. 1361--1371, 1997.

[MB02] C. Mathis, T.M. Breuel, Classification using a hierarchical Bayesian approach, *Procs. ICPR XVI*, IEEE Computer Society Press, Quebec City, August 2002.

[NMTM00] K. Nigam, A.K. McCallum, S. Thrun, T.M. Mitchell, Text classification from labeled and unlabeled documents using EM, Machine Learning 39, # 2/3, pp. 103-134, 2000.

[NS66] G. Nagy, G.L. Shelton, Self-corrective character recognition system, (with G.L. Shelton), *IEEE Transactions on Information Theory IT-12,* #2, pp. 215-222, April 1966.

[NX99] G. Nagy, Y. Xu. Prototype Extraction and Adaptive OCR, IEEE Trans. PAMI-21, 12, pp. 1280-1296, Dec. 1999.

[Nagy 68] G. Nagy, The application of nonsupervised learning to character recognition, in *Pattern Recognition,* (L. Kanal, Editor), Thompson Book Company, Washington, pp. 391-398, 1968.

[Sarkar 00] P. Sarkar, Style consistency in pattern fields, Rensselaer Polytechnic Institute PhD Dissertation, May 2000.

[SN01] P. Sarkar, G. Nagy, Style-consistency in isogeneous patterns, *Procs. ICDAR-01, 1169-1174, IEEE Computer Society Press, Sept. 2001*.

[Sarkar02] P.Sarkar. An Iterative Algorithm for Optimal Style-conscious Field Classification , Proceedings of the 16th ICPR, August 2002, Quebec City, Canada.

[SB03] P. Sarkar, T. Breuel, Amplifying accuracy through style consistency, *Symposium on Document Image Understanding Technology*, Greenbelt, MD, pp. 245-252, April 2003.

[SBH02] P. Sarkar, H. S. Baird, and J. Henderson. Triage of OCR output using 'confidence' scores, in *Proceedings of SPIE/IS&T 2002 Document Recognition & Retrieval IX Conf. (DR&R IX)*, San Jose, California, USA, pp. 20-25, January 2002.

[SN00] P. Sarkar, G. Nagy, Classification of style-constrained pattern fields, Procs. *ICPR-XV, Vol. 2*, pp. 859-862, Barcelona, September 2000.

[SNLZ08] P. Sarkar, G. Nagy, D. Lopresti, J. Zhou, Spatial Sampling of Printed Patterns, IEEE *Trans. PAMI-20*, 3, pp. 159-170, March 1998.

[Summers 03] K. Summers, Varying effects of image improvement methods on OCR accuracy, , *Symposium on Document Image Understanding Technology*, Greenbelt, MD, pp. 245-252, April 2003.

[VLFN02] S. Veeramachaneni, C-L Liu, H. Fujisawa), G. Nagy,, Classifying isogenous fields, *Proceedings of the 8th International Workshop on Frontiers of Handwriting Recognition* (IWFHR'02), Niagara-on-the-Lake, Canada, published by CEDAR/IAPR, pp. 41-46,August 2002.

[VN03] S. Veeramachaneni, G. Nagy, A Ptolemaic model for OCR, *Procs. ICDAR-03*, Edinburgh, pp. 577-582, August 2003.

[VN02a] S. Veeramachaneni, G. Nagy, Style-conscious quadratic classifier, *Procs. ICPR XVI*, IEEE Computer Society Press, Vol. II, pp. 72-75, Aug. 2002.

[VN02b] S. Veeramachaneni, G. Nagy, Classifier adaptation with non-representative training data, in Document Analysis Systems V*,* Springer LNCS 2423*, Proceedings of the Fifth International Workshop on Document Analysis Systems* (D. Lopresti, J. Hu, R. Kashi, editors), 123-133, Princeton, NJ, August 2002,.

[VN03] S. Veeramachaneni, G. Nagy, Adaptive classifiers for multisource OCR, Int'l Journal of Document Analysis and Recognition, accepted for publication in March 2003.

[Veeramachaneni 02] S. Veeramachaneni, Syle-constrained quadratic field classifiers, Rensselaer Polytechnic Institute PhD Dissertation, August 2002.

[WCZ03] L. Wang, K.L. Chan, Z. Zhang, Bootstrapping SVM active learing by incorporating unlabelled images for image retrieval, *Procs. CVPR03*, IEEE, 2003.

[YF79] M. Yasuda, H. Fujisawa, An improved correlation method for character recognition, *Systems, Computers, and Control,* Vol. 10, #2, pp. 29-38, 1979 (translated from Trans. IEICE Japan, Vol 62-D, #3, pp. 217-224, 1979).

[ZI98] A. W. Zramdini, R. Ingold, Optical Font Recognition Using Typographical Features, *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(8), pp. 877-882 (1998)