Scaling Up Whole-Book Recognition

Pingping Xiu & Henry S. Baird

Computer Science & Engineering Dept
Lehigh University
19 Memorial Drive West, Bethlehem, PA 18017 USA

E-mail: pix206@lehigh.edu, baird@cse.lehigh.edu
URL: www.cse.lehigh.edu/˜pix206, www.cse.lehigh.edu/˜baird

Abstract

We describe the results of large-scale experiments with algorithms for unsupervised improvement of recognition of book-images using fully automatic mutual-entropy-based model adaptation. Each experiment is initialized with an imperfect iconic model derived from errorful OCR results, and a more or less perfect linguistic model, after which our fully automatic adaptation algorithm corrects the iconic model to achieve improved accuracy, guided only by evidence within the test set. Mutual-entropy scores measure disagreements between the two models and identify candidates for iconic model correction. Previously published experiments have shown that word error rates fall monotonically with passage length. Here we show similar results for character error rates extending over far longer passages up to fifty pages in length: we observed error rates were driven from 25% down to 1.9%. We present new experimental results to support the motivating principle of our strategy: that error rates and mutual-entropy scores are strongly correlated. Also, we discuss theoretical, algorithmic, and methodological challenges that we have encountered as we scale up experiments towards complete books. [1]

Keywords: *document image recognition, book recognition, isogeny, adaptive classification, anytime algorithms, model adaptation, mutual entropy*

1. Introduction

We are investigating fully automatic methods for whole-book recognition. In [11] we introduced an information-theoretic framework for identifying significant disagreements between models—the *iconic* model and the *linguistic* model—and interpreting these as candidates for corrections of one or the other of the two models so that, when the updated models are reapplied to perform recognition, a lower error rate on the entire passage results.

Our research builds on over a decades' work showing that adaptive classifiers can improve accuracy without human intervention[7]. Tao Hong[4] showed that within a book, strong "visual" (image-based, iconic) consistency-constraints support automatic post-processing that reduces error. These successes appear, to us, to be due largely to *isogeny*— the tendency of particular documents to contain only a small subset of all the typefaces, languages, image qualities, and other variabilities that occur in large collections[9]. It is well known that if models of the typefaces, languages, etc were known, even if only approximately, optimizing recognition jointly across all the models improves the accuracy[1, 6, 8].

In a long, highly isogenous book, identical (or similar) character images will occur multiple times, and the same word will also occur multiple times, independently. If the models are inaccurate, the resulting errors cause repeated disagreements between the models, which can be measured at character, word, and passage scales. Correct model adaptation, which leads to a better accuracy, will presumably also lower passage-scale model disagreement. Therefore passage-scale mutual entropy can drive model correction and reduce error rates.

In [11], a small-scale experiment, on a single textline, using an adaptation algorithm we now call ME1.0, illustrated policies that allowed automatic corrections to be made to both models, and showed empirically that both character error-rates and word-error rates could fall as a result. In [10], using an improved algorithm (ME2.0) which copes with segmentation errors and runs faster, we experimented on passages up to ten pages in length, and observed that the word recognition rate for longer words increased significantly as passage length increased.

In this paper, we report experiments scaled up to fifty pages in length, and we focus our attention on iconic-model corrections. We have observed character error rates falling as a function of passage length, from an initial 25%, down to 1.9% on a passage of fifty pages. We report compelling evidence of strong correlation between word error rate and passage-scale mutual entropy, validating a key premise of our strategy. We have measured the effects of varying the number of prototype images per character class in the

[1] Published in *Proceedings, IAPR 10th Int'l Conf. on Document Analysis and Recognition (ICDAR2009)*, Barcelona, Spain, July 26-29, 2009.

iconic model, and two or three perform better than one. We also describe a successful randomization strategy for coping with the computational complexity challenge of long passages.

In Section 2, we introduce the mathematical framework of our approach. In Section 3, we motivate the design of the present experiments and give details of the algorithmic enhancements tested. In Section 4, we present and analyze the results of the experiments. In Section 5, we discuss the results and draw conclusions. In Section 6, we list future algorithmic enhancements and experiments.

2. Mathematical Framework

2.1. Probabilistic Models

In our framework, two different kinds of models are required: an iconic model and a linguistic model. The iconic model, when applied to recognition, must allow the computation of *a posteriori* probabilities for all the character classes. (Of course, many such models are known [2]; we use Hamming-distance matching to multiple character image templates.) For a linguistic model, we expect to be given a lexicon (a dictionary containing valid words). The lexicon should cover most valid words, but may be incomplete; we also expect probabilities to be assigned to each word in the lexicon.

2.2. Independence Assumptions and Word Recognition

Now let X denote a sequence of T observations of character images (*i.e.* a word that is T characters long), and let S denote the true classes of these characters (in communication-theory terms, it is the inner state sequence that generates X):

$$X = (x_1, x_2, \cdots, x_T)\,, S = (s_1, s_2, \cdots, s_T) \qquad (1)$$

where x_i are character images, and s_j are symbols of an alphabet. We adopt the following independence assumption, that each x_i is solely determined by its associated s_i:

$$P(x_i|s_i, \mathcal{F}) = P(x_i|s_i) \qquad (2)$$

Where $\mathcal{F} = (Y, K)\,, Y \subseteq X - \{x_i\}$ and $K \subseteq S - \{s_i\}$. This assumption is similar to the one chosen by Kopec and Chou in their Document Image Decoding theory[5].

Our *linguistic model* is $P(S)$, the prior probability that word S is valid. Our independence assumption implies that

$$P(X|S) = \prod_{i=1}^{T} P(x_i|s_i) \qquad (3)$$

And

$$P(x_1, x_2, \cdots, x_T) = \alpha \cdot \prod_{i=1}^{T} P(x_i) \qquad (4)$$

where

$$\alpha = \sum_{(s_1 s_2 \cdots s_T)} \left[\prod_{i=1}^{T} P(s_i|x_i) \cdot \frac{P(s_1 s_2 \cdots s_T)}{\prod_{i=1}^{T} P(s_i)} \right] \qquad (5)$$

Our *iconic model* is denoted by the function $P(s|x)$ for all symbols s and all character images x. So we can derive $P(S|X)$, the result of word recognition informed by both the iconic and linguistic models:

$$P(S|X) = \frac{1}{\alpha} \cdot \prod_{i=1}^{T} P(s_i|x_i) \cdot \frac{P(S)}{\prod_{i=1}^{T} P(s_i)} \qquad (6)$$

2.3. Mutual Entropy Measurements On Word Recognition

The *mutual entropy* $M(P, P')$ between two distribution P and P' is defined as:

$$\mathcal{M}(P, P') = -\sum P \cdot \log P' \qquad (7)$$

and we apply it to measure the difference or "disagreement" between the distributions $P(S|X)$ and $P'(S|X)$, where $P(S|X)$ is the *a posterior* probability distribution of the character string S given the image of the whole word X, and $P^{'}(S|X) = P(s_1|x_1) \cdot P(s_2|x_2) \cdot \cdots \cdot P(s_T|x_T)$ is the distribution of the character string assuming that there is no linguistic constraints or the distributions of individual characters are independent of one another.

A property of mutual entropy which is critically important to us is that the more the distributions P and P' differ from one another, the greater $\mathcal{M}(P, P')$ will be. Also, $\mathcal{M}$ can be further decomposed into per-character disagreement measurements as follows:

$$\mathcal{M} = \sum_{i=1}^{T} M\left(s_i|X, s_i|x_i\right) \qquad (8)$$

where

$$M\left(s_i|X, s_i|x_i\right) = -\sum_{s_i} P(s_i|X) \log P(s_i|x_i) \qquad (9)$$

This measures disagreement on an *individual character image* x_i: that is, the disagreement between two probability distributions on character classes: (1) the distribution resulting from application of the iconic model alone, and (2) the distribution resulting from the application of both the iconic and liguistic models. Thus we call this *character-scale* mutual entropy.

And $P(s_i|X)$ is the projection probability of $P(S|X)$ onto a particular element of a field: $P(s_i|X) = \sum_{s_j, j \neq i} P(S|X)$. If the iconic output "agrees" with the linguistic model, the two distributions should be close to one another, resulting in a smaller $\mathcal{M}$. If classification based on the iconic model yields the correct word interpretation but there is no corresponding word entry in the dictionary, then the disagreement between the two models should be high, which results in a high value of $\mathcal{M}$ for that word. As a result, *mutual entropy measures the disagreement between the iconic and linguistic models.*

The disagreement for one character can be interpreted as a measure of the urgency of changing one model or the other. In order to change the iconic model, we can modify the $P(s_i|x_i)$ for that character's image: one way of doing this is to swap in a new character template image. In order to change the linguistic model, we can modify the $P(S)$ for some word 'S': one (crude) way to do this is to delete or insert words from the dictionary.

As a result, we have three different kinds of measurements:

1. The *character-scale mutual entropy* $M\left(s_i|X, s_i|x_i\right)$: this measures the model disagreements in regard to a specific character. It can indicate the urgency of changing the iconic model for that character.
2. The *word-scale mutual entropy* $\mathcal{M}$ measures the model disagreements in regard to a particular word. It can indicate the urgency of changing the linguistic model for that word.
3. The *passage-scale mutual entropy* $\sum\mathcal{M}$: this measures the overall disagreements of the iconic model and linguistic model . We choose to use this as the objective function to drive improvements of both models.

So far, we've defined different measurements that operate at three different scales: character-scale, word-scale, and passage-scale. Do they have any relationship to the recognition rate? We conjecture that passage-scale mutual-entropy is strongly negatively correlated with recognition rate. Our strategy is to minimize these disagreements through a process of model adaptation: that is, applying a sequence of corrections to both models.

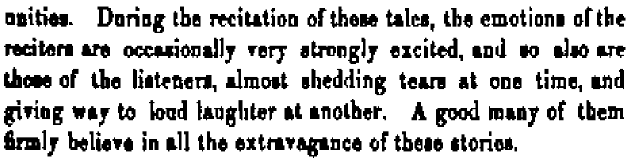
3. Experimental Design

The principal goals of the work reported here are to test the performance of the ME2.0 algorithm on long passages, and to characterize the efficacy of two algorithmic enhancements: (1) speeding-up by randomization of a potentially expensive inner-loop computation that decides when to accept an adaptation; and (2) allowing each character code to possess more than two prototype templates in the iconic model.

In the experiment reported here (using ME2.0), model adaptation proceeds by a sequence of *epochs*. In one epoch, every word in the passage is examined: its top-choice word interpretation (resulting from the current models) assigns a character class label s_i to each character x_i in the word. Among these, the algorithm chooses the pair $(x*, s*)$ with the highest character-scale disagreement within the word, then attempts to adapt the iconic model for character class $s*$ by picking one of its templates at random and replacing it with $x*$. This attempted adaptation is evaluated, and may be accepted as a *correction*. or undone and discarded.

Thus the total number of adaptations attempted in an epoch equals the number of words in the passage, and is in general larger than the number of corrections accepted. Evaluating an attempted adaptation is accomplished, within our theoretical framework, by recomputing the passage-scale mutual entropy due to the adaptation: if it decreases, the adaptation is accepted. However, if this recomputation is performed in a brute-force manner, it will take time proportional to the passage length; and the number of words is also proportional to passage length; so each epoch would in time quadratic in the passage length.

This motivates our randomization enhancement: instead of recomputing passage-scale mutual entropy on *all* words in the passage, we choose a certain fraction of the words at random, and estimate the change on them.



unities. During the recitation of these tales, the emotions of the reciters are occasionally very strongly excited, and so also are those of the listeners, almost shedding tears at one time, and giving way to loud laughter at another. A good many of them firmly believe in all the extravagance of these stories.

Figure 1. A sample textual image used in the experiments

In these experiments, we use page images plus an imperfect OCR transcript for one of the books ("Popular tales of the west highlands") provided in the publically released Google Book Search Dataset [3](an example image is shown in Figure 1). In this book, each page contains roughly 350 words, and we use up to 50 pages on the experiments in this paper. We used this OCR transcript to perform word segmentation alignment, and we proofread the transcript and the alignment manually. We are grateful to Prof. Cheng and his students in the Beijing Information Science and Technology University for assistance in ground-truthing.

We initialized the iconic model from a short passage, yielding a low inital accuracy of sixty percent words correct and fifty-five percent characters correct. The linguistic model was initialized with the 4562 words occurring in 50 pages' groundtruth: thus it is a "perfect lexicon" for the 50 pages, and a superset for smaller passage lengths. (This contrasts with our previous papers, where the linguistic model was initialized from a public-domain dictionary containing around 50,000 words which did not fully cover the test set.) The joint recognition results from these initialized models yielded an approximately 25% character error rate: this is the "initial error rate" to which we compare our adaptation results.

4. Experimental Results

The principal experimental result is that character error rate falls as a function of the length of the passage operated upon by our adaptation algorithm. In Figure 2 the initial character error rate of 25% is shown as circles (o), and the average final character error rate, after adaptation over three epochs, is shown as stars (*). The horizontal axis (passage length in pages) and the vertical axis (character error rate) are displayed in log scale. Character error rates are measured on the word recognition result given by Equation 6. Passage lengths include 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 16, 25 and 50 separately. The 50-page result was computed in a single experiment; the others, for $m < 50$ pages, were computed as averages over $\lfloor 50/m \rfloor$ experiments on nonoverlapping subsets of pages. In every case, the final error rate was smaller than the initial error rate, and error rates fall almost monoonically as a function of passage length.

We plot the results of linear regression in log-log scale: the cross-correlation coefficient was -0.95, which indicates a strong negative linear relationship. The residuals suggest that the fall in error rate is still strongly marked up to about ten pages, and it continues to fall significantly up to twenty-five pages. It is hard to project these data beyond fifty pages, but further statistically significant improvement does not seem to be ruled out.

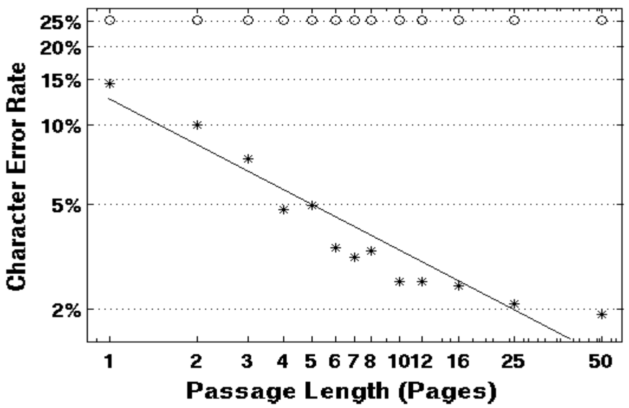


Figure 2. Character error rate (%) decreases as a function of passage length (in pages). Stars(*) indicate average character error rates after three epochs of the model adaptation algorithm, and circles(o) indicate the initial error rate before adaptation. Both axes are displayed in log-scale. The straight line is a plot of a linear regression for the star data points.

The next experimental result illuminates the efficacy of allowing more than two prototype templates for each character class in the iconic model. The motive of this new policy is that we expect that a larger number of prototype templates may be needed to represent variations within long passages. We designed experriments to study the relationship between the number of templates per class and the accuracy of adaptation. We choose a set of eight pages for these experiments, one for each number of templates from one to ten. In each experiment, we ran as many epochs are were needed for the character error rate to stabilize. The results are plotted in Figure 3 (note: the data point for templates=9 is missing). These data suggest that, for a passage length of eight pages, the number of prototype templates per character class should be expanded to three at least. Not shown in this figure are our observations that the number of epochs required for convergence increased as a function of the number of templates. So although it appears to be safe to allow more than three templates, this may increase runtime required for best results.

The present implementation of ME2.0 has a runtime quadratic in passage length: it goes through the passage word-by-word to examine the potential iconic correction within each word; and it computes the passage level mutual entropy after each suggested correction to judge whether to adopt it. We have tested a randomization technique to reduce computation complexity: instead of computing passage-scale mutual entropy on the entire passage, this randomly selects words from within the entire passage with a given sampling factor, and estimates the change in passage-scale mutual entropy to judge whether to adopt a correction. Figure 4 shows that with a sampling rate of 12.43%, the error rate of the result rises by only one percent over the optimal (achieved by the

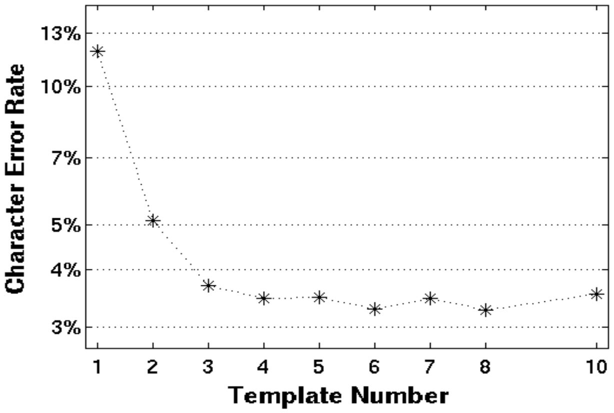


Figure 3. Character error rate (%) as a function of the number of prototype templates used to represent a single character class in the iconic model. Computed on a passage of eight pages with a randomization factor of 0.125. Adaptation was allowed to iterate until error rates stabilized.

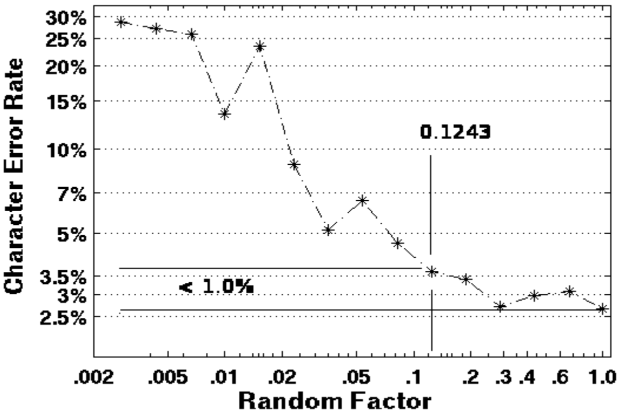


Figure 4. Character error rate (%) as a function of the random sampling factor used to accept adaptations. A randomization factor of 0.124 gives a speed-up of a factor of 8, while achieving an error rate within one per cent of not randomizing at all.

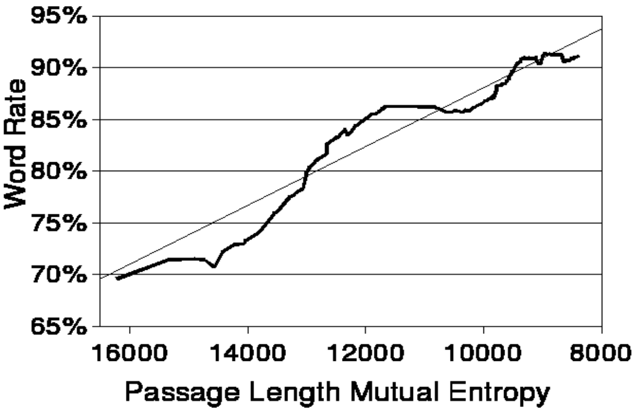


Figure 5. The relationship between the falling passage-scale mutual entropy and the increasing trend of the word accuracy. The straight line is a plot of a linear regression for the data points.

brute force algorithm) but the algorithm speeds up by a factor of 8.

We have observed a strong relationship between the final accuracy and the passage-length mutual entropy, as shown in Figure 5. From the curve, we can tell that decreasing passage mutual entropy correlates strongly to increasing word accuracy. Thus we feel more confident that the fundamental technical principle of our approach—that mutual entropy tracks accuracy—continues to hold.

5. Discussion and Conclusions

The often nearly monotonic improvement of word and character accuracy as passage-length increases remains highly encouraging. The enhancement to randomize the passage to evaluating the change of the mutual entropy suggests the algorithm can tackle a scale of experiment much larger than 50 pages this paper has experimented on. (We did not use the randomization technique in all experiments in Figure 2.) The enhancement to the iconic model allowing more than two templates per character is clearly valuable; at the present scale of experiments the sufficient number of templates appears to be three. And the seemingly low asymptote in Figure 3 suggests that it is safe to increase the template upper limit to cope with the arising complexity of the sample set when we scale up the experiment. The accumulation of new evidence continues to support our working hypothesis that passage-scale mutual entropy is strongly negatively correlated with accuracy.

6. Future Work

The most urgent questions concern how well the algorithms, with its various enhancements, will perform as the experiments scale up to approach passages that embrace entire books. The efficacy of randomization of the estimation of passage-scale mutual entropy promises that there are no insurmountable runtime obstacles to scaling-up to hundreds of pages. Also, the evidence suggests that accuracies with continue to rise, and we do not yet see clear evidence of a low asymptote. As passage length increases, we may find that the optimal number of templates per character class may grow beyond three. Policies for applying corrections to the iconic model have several interesting variations including replacing the template which most mismatches the current character. to achieve significant scale-ups.

References

[1] T. Breuel and K. Popat. Recent work in the document image decoding group at xerox PARC. In *Proc., DOD-sponsored Symposium on Document Image Understanding Technology (SDIUT 2001)*, Columbia, Maryland, April 2001.

[2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification, 2nd Edition*. Wiley, New York, 2001.

[3] A. . Google Inc. Book Search Dataset, Version 1.0.

[4] T. Hong. *Degraded Text Recognition Using Visual And Linguistic Context*. PhD thesis, 1995.

[5] G. Kopec and P. Chou. Document image decoding using Markov source models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI–16:602–617, June 1994.

[6] G. Kopec, M. Said, and K. Popat. N-gram language models for document image decoding. In *IS&T/SPIE Electronic Imaging 2002 Proc. of Document Recognition and Retrieval IV*, San Jose, California, January 2002.

[7] G. Nagy and H. S. Baird. A self-correcting 100-font classifier. In *Proc., IS&T/SPIE Symp. on Electronic Imaging: Science & Technology*, San Jose, CA, 1994.

[8] P. Sarkar, H. S. Baird, and X. Zhang. Training on severely degraded text–line images. [submitted to] IAPR Int'l Conf. on Document Analysis & Recognition, Edinburgh, August, 2003.

[9] P. Sarkar and G. Nagy. Style consistent classification of isogenous patterns. *IEEE Trans. on PAMI*, 27(1), January 2005.

[10] P. XIU and H. Baird. Towards whole-book recognition. In *Proceedings., 8th IAPR Document Analysis Workshop (DAS'08)*, Nara, Japan, September 2008.

[11] P. XIU and H. Baird. Whole book recognition using mutual-entropy-based model adaptation. In *Proc., IS&T/SPIE Document Recognition & Retrieval XII Conf.*, San Jose, CA, January 2008.