Whole-Book Recognition
using Mutual-Entropy-Driven Model Adaptation

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 an approach to unsupervised high-accuracy recognition of the textual contents of an entire book using fully automatic mutual-entropy-based model adaptation. Given images of all the pages of a book together with approximate models of image formation (*e.g.* a character-image classifier) and linguistics (*e.g.* a word-occurrence probability model), we detect evidence for disagreements between the two models by analyzing the mutual entropy between two kinds of probability distributions: (1) the* a posteriori *probabilities of character classes (the recognition results from image classification alone), and (2) the* a posteriori *probabilities of word classes (the recognition results from image classification combined with linguistic constraints). The most serious of these disagreements are identified as candidates for automatic corrections to one or the other of the models. We describe a formal information-theoretic framework for detecting model disagreement and for proposing corrections. We illustrate this approach on a small test case selected from real book-image data. This reveals that a sequence of automatic model corrections can drive improvements in both models, and can achieve a lower recognition error rate. The importance of considering the contents of the whole book is motivated by a series of studies, over the last decade, showing that isogeny can be exploited to achieve unsupervised improvements in recognition accuracy.*

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

1. INTRODUCTION

Millions of books are being scanned cover–to–cover and the resulting page images, together with the results of automatic optical character recognition, are being made available on the Web[1].[13] The accuracy of the present generation of OCR systems varies widely from book to book.[9] When, for a particular book, state-of-the-art OCR accuracy is low, a user's only recourse is tedious and expensive manual correction. Thus there is a growing need for improved methods for *whole-book* recognition, which accept as input the images of the book's pages and an initial transcription, along with a dictionary that might be incomplete, and then proceed to improve the quality of the transcription on both iconic models and the linguistic models (dictionary).

We are investigating fully automatic methods for whole-book recognition. Research over the last decade has shown that adaptive classifiers can sometimes improve accuracy substantially without human intervention.[8] Tao Hong[5] showed that within a book, strong "visual" (image-based, iconic) constraints support automatic post-processing that reduces error. This appears to be due largely to the fact that many documents (and, especially, books) are strikingly *isogenous*, that is, each particular document contains only a small subset of all the typefaces, languages, topics, layout styles, image qualities, and other variabilities that can and do occur in large collections of documents and books.[12] Now it is well known that if models of the specific faces, languages, etc that occur in the book were known, even if only approximately, a strategy of optimizing recognition jointly across all the models can dramatically improve accuracy.[2,7,11] Motivated by these recent technical developments, we are now investigating techniques for locating significant disagreements between models—here, *iconic* and *linguistic* models—and interpreting these disagreements as evidence for potential corrections of one or the other of the two models so that, when the updated models are reapplied to perform recognition, a lower overall error rate results.

In a long, highly isogenous book, we expect that identical (or similar) character images will occur multiple times, and the same word will also occur many times; these events are independent of one another to a considerable degree. By examining all the occurrences of character images in the book and measuring the (to speak informally) 'consistency' of each with its linguistic context, we can estimate the 'fitness' of the iconic model for that character image. Similarly, by summing up 'consistency' scores for a word across the entire book, we can estimate the 'fitness' of the linguistic model for that word. Section 2 will make these informal intuitions precise.

We propose a mutual-entropy-based function to evaluate disagreements between iconic and linguistic models. This makes possible an automatic model-adaptation technique that consists of the following four steps: (a) identifying the characters or the words where the two models strongly disagree; (b) interpreting this disagreement as evidence for corrections to one or the other of the models; (c) applying corrections to the models; and (d) reapplying the updated models for (one hopes) improved recognition results.

We illustrate one such algorithm using a small test case using real book-image data.

2. FORMAL FRAMEWORK

2.1 Probabilistic Models

In our framework, two different kinds of models are required: an iconic model and a linguistic model. We impose four conditions on iconic models:

1. The iconic model, when applied to recognition, must compute *a posteriori* probabilities for all the character classes. (Of course, many such models are known;[4] we'll give details of our choice later.)
2. We expect that, in general, any given iconic model may be imperfect; however, we want it to be good enough to allow our mutual-entropy-based methodology to identify model contradictions and eliminate them. (We do not yet know exactly how accurate the model needs to be for this to happen reliably.)
3. Also, the iconic model should be static: that is, identical character images should be assigned identical classes with the same probabilities.
4. An iconic model should be continuous in some image metric space: that is, it should give similar results on samples whose images are nearby one another under the metric. One example of such a metric is Hamming distance, but of course many others are known. Associated with this continuity assumption is the requirement that if one sample changes its *a posteriori* probability distribution among the classes, then image samples in its neighborhood should also be affected similarly. (We do not yet know how best to enforce these requirements.)

For a linguistic model, we expect to be given a lexicon (a dictionary containing valid words). The lexicon should cover most words appearing in the testing images, but may be incomplete. We also expect probabilities of occurrence to be assigned to each word in the lexicon: we can of course infer such statistics from a given corpus.

2.2 Independence Assumptions and Word Recognition

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

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

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:

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

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.[6]

The *linguistic model* is $P(S)$, the prior probability of occurrence of any word S.

Our independence assumption implies that

$$P(X|S) = P(x_1, x_2, \cdots, x_T|s_1, s_2, \cdots, s_T) = \prod_{i=1}^{T} P(x_i|x_{i-1}, \cdots, x_1, s_1, s_2, \cdots, s_T) = \prod_{i=1}^{T} P(x_i|s_i) \tag{3}$$

By elementary definitions,

$$\begin{aligned} P(x_1, x_2, \cdots, x_T) &= \sum_{(s_1 s_2 \cdots s_T)} \left[P(x_1 x_2 \cdots x_T | s_1 s_2 \cdots s_T) \cdot P(s_1 s_2 \cdots s_T) \right] \\ &= \sum_{(s_1 s_2 \cdots s_T)} \left[\prod_{i=1}^{T} P(x_i|s_i) \cdot P(s_1 s_2 \cdots s_T) \right] = \alpha \cdot \prod_{i=1}^{T} P(x_i) \end{aligned} \tag{4}$$

where

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

Our *iconic model* (which provides posterior probabilities for all the classes) 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:

$$P(S|X) = \frac{P(S, X)}{P(X)} = \frac{\left[\prod_{i=1}^{T} P(x_i|s_i) \right] \cdot P(S)}{\alpha \cdot \prod_{i=1}^{T} P(x_i)} = \frac{1}{\alpha} \cdot \prod_{i=1}^{T} P(s_i|x_i) \cdot \frac{P(S)}{\prod_{i=1}^{T} P(s_i)} = \frac{1}{\alpha} \cdot \prod_{i=1}^{T} P(s_i|x_i) \cdot \frac{P(S)}{\prod_{i=1}^{T} P(s_i)} \tag{6}$$

2.3 Mutual Entropy Model On Word Recognition

The *mutual entropy* $M(P, P')$ between the distributions $P(S|X)$ and $P'(S|X)$ is defined as:

$$\mathcal{M}(P, P') = -\sum_{S} P \cdot \log P' \tag{7}$$

which measures the difference or “disagreement” between the two 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 with each other.

The $\mathcal{M}$ has an attribute: the more the distribution P and P' differs from each other, the greater the $\mathcal{M}(P, P')$ will be. Also, $\mathcal{M}$ can be further decomposed into the character-level disagreement measurements as follows:

$$\begin{aligned} \mathcal{M} &= -\sum_{i=1}^{T} \sum_{s_i} P(s_i|X) \log P(s_i|x_i) & (8) \\ &= \sum_{i=1}^{T} M\left(s_i|X, s_i|x_i\right) & (9) \end{aligned}$$

Where

$$M\left(s_i|X, s_i|x_i\right) = -\sum_{s_i} P(s_i|X) \log P(s_i|x_i) \tag{10}$$

Which measures the disagreement on individual character x_i. And $P(s_i|X)$ is the marginal probability.

$$P(s_i|X) = \sum_{s_j, j \neq i} P(S|X) \tag{11}$$

If the iconic output "agrees" with the linguistic model, the two distributions should be close to each other, resulting in a smaller $\mathcal{M}$; otherwise, the linguistic information $P(S)$ will make $P(S|X)$ quite different from the iconic output $P(s_1|x_1) \cdot P(s_2|x_2) \cdot \cdots \cdot P(s_T|x_T)$. As a result, mutual entropy measures the disagreement between the iconic and linguistic models. If the iconic models give out the correct answer but there is no corresponding entry in the dictionary, then the disagreement between the two model should be high, which results in a high value on $\mathcal{M}$ for that word.

$M\left(s_i|X, s_i|x_i\right)$ indicates disagreements between the *a posteriori* probability and the iconic probability for an individual character in the word. 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. In order to change the linguistic model, we can modify the $P(S)$ for some word 'S'.

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 overall mutual entropy of the whole passage $\sum \mathcal{M}$: this measures the overall disagreements of the iconic model and linguistic model overall. 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 argue that the overall mutual-entropy measurements (on the entire passage) are correlated with recognition rates.

1. If recognition performance is high, we expect small overall disagreement $\sum \mathcal{M}$. This is easy to understand: if character recognition is poor, either the iconic model has many errors, or the language model is incomplete: highly probably, they have a strong disagreement.
2. Within a word, if $\mathcal{M}$ is high, there are two possibilities: one, the word to be recognized may not in the dictionary; or, two, the word may contain incorrectly recognized characters due to an inaccurate iconic model.
3. For a single character, if $M\left(s_i|X, s_i|x_i\right)$ is high, there are two possibilities: one, the iconic model is wrong on this character; or, two, the language model may be incomplete.

Our strategy is to minimize these disagreements through a process of model adaptation: that is, applying a sequence of corrections to both models. However, changing the models is not always safe. In fact, changes can sometimes lead to unrecoverable errors. Consider a extreme situation in which the language model $P(S)$ is the uniform distribution on S, and the iconic models assign the top candidate with probability 1. Then, all the disagreement measurements are zero, but the result is not necessarily correct, and we are not able to correct errors because we detect no disagreements. As a result, we should change models conservatively.

Based on this principle, we may utilize the disagreement measurements to discover model disagreements and fix them to improve the recognition result. In the following section, we use an illustrative experiment to show an example process.

3. ALGORITHM DESIGN

In this section, we define both the iconic model and the linguistic model and describe the model-adaptation algorithm.

The criteria for designing the iconic model are: first, it should produce the probability $P(s_i|x_i)$; second, the character classifier for the iconic model should not be intrinsically complex, or from the perspective of statistical learning theory, the classifier's VC dimension should be low. This implies that, if a change to the iconic model affects one character image, it should impact all similar images; that is, changes to the iconic model should propagate to similar images.

In our experiment, the iconic model is initialized by first assigning to every isolated character image the class given by the OCR result (right or wrong)*. Then we choose, for each character class s, a single character image to act as its template T_s in a minimum-Hamming-distance classifier. under Hamming distance, we assign each Later, given a testing character image I, we use this Hamming distance to calculate the confidence value of each code $\mathbf{conf}_{s,I} = hamming(I, T_s)$, and then we calculate the $P(s|x)$ by the formula:

$$P(s|x) = \beta \cdot e^{-\alpha \cdot \mathbf{conf}_{s,I}} \tag{12}$$

Where β is the normalization factor:

$$\beta = \frac{1}{\sum_{s} e^{-\alpha \cdot \mathbf{conf}_{s,I}}} \tag{13}$$

This defines the behavior of the iconic model.

Now we say how we make changes to the iconic model. If the iconic model and linguistic model have high disagreement on one character, i.e. $M\,(s_i|X, s_i|x_i)$ is high, we can change the distribution $P(s_i|x_i)$ to more closely fit $P(s_i|X)$ and so to lower $M\,(s_i|X, s_i|x_i)$. We change the template of the top code c_{max} in $P(s_i|X)$ ($P(s_i = c_{max}|X) \geq P(s_i = c|X)$) to the image of x_i, in order to increase the ranking of the code c_{max} in the distribution of $P(s_i|x_i)$. For example, if we want a character's top candidate code to change from "b" to "h", we may change "h" 's template to this character's image. This increases the probability of the candidate "h" for this character, meanwhile lower down the ranking of "b" in the candidate list of the modified iconic model.

The linguistic model is a set of probability functions related to different lengths of words: $P^1(S^1), P^2(S^2), P^3(S^3), \cdots$, where $P^i(S^i)$ represents the language model with word length i. For each $S = \{s_1 s_2 \cdots s_T\}$ in the dictionary, $P(S)$ has a non-zero value. For each $S = \{s_1 s_2 \cdots s_T\}$ that is not in the dictionary, $P(S)$ equals zero. For example, the word "entry" should have a non-zero value in $P^5(S^5)$: $P^5(S^5 = entry) \neq 0$. The lingistic model is initiatized, in this experiment, by adding an entry for each word in the training passage shown in Figure 1(b), and assigning them equal probabilities.

Changing the language model means changing the function $P^i(S^i)$ by adding or deleting one word entry S^i in the dictionary. We immediately recompute probabilities to ensure all $P^i(S^i)$ are equal.

Thus model-adaptation algorithm is as follows: First, using the initial models, we recognize the entire test image (the passage of one text-line in Figure 1(a)), and calculate $\sum \mathcal{M}$, $\mathcal{M}$ and $M\,(s_i|X, s_i|x_i)$. We select a list of characters whose similar characters (under Hamming distance) have a larger summation $M\,(s_i|X, s_i|x_i)$, that is,

$$M_c(x_i) = \sum_{distance(x,x_i)<d} M\,(s|X, s|x) \tag{14}$$

Intuitively, the greater $M_c x_i$, the more reduction in $\sum \mathcal{M}$ if the model disagreements are resolved successfully. Thus characters with greater $M_c x_i$ should be dealt with before the smaller ones. Among those characters whose $M_c x_i$ are ranked higher, we choose the one that can drive the $\sum \mathcal{M}$ down most by replacing the corresponding template in the model with its own image. When no characters are able to lower $\sum \mathcal{M}$ any further, we then switch to making changes to the linguistic model.

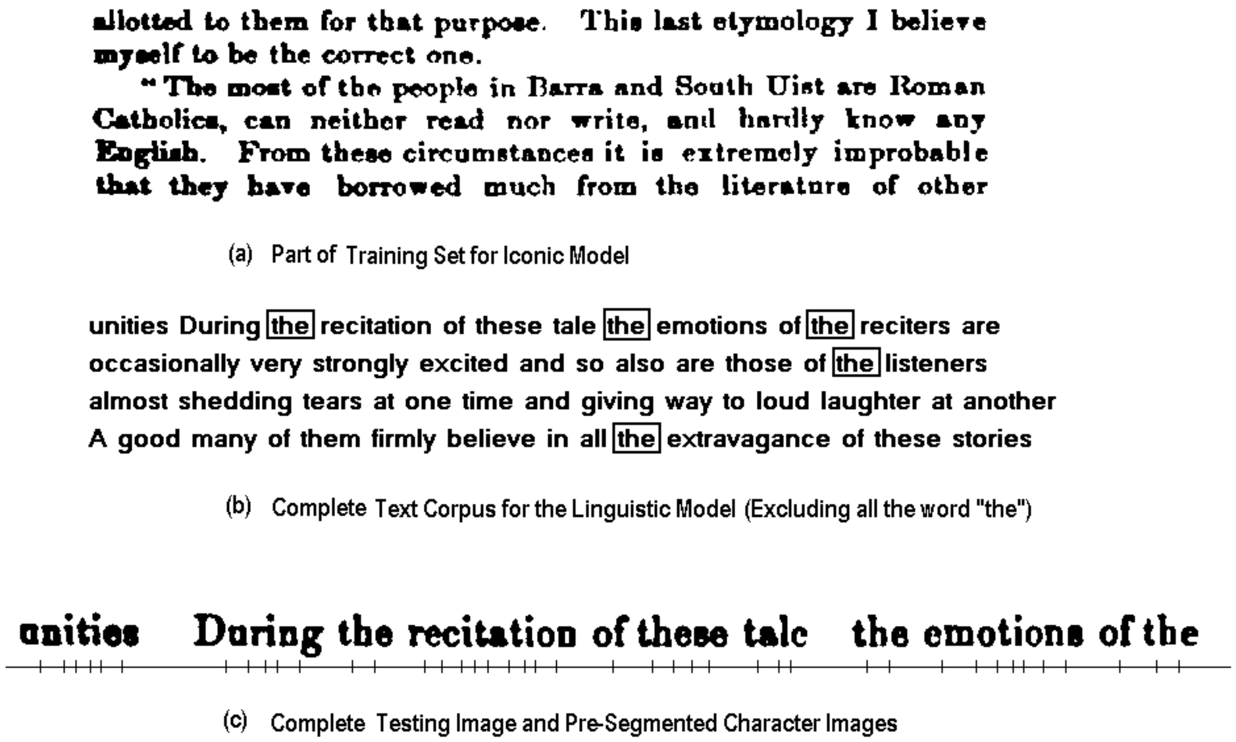
We define the normalized disagreement, $\mathcal{M}/T$ to be the word-scale disagreement $\mathcal{M}$ divided by the word length T; this allows comparisons between words of different lengths. We find the word that with the maximal $\mathcal{M}/T$ and add the first candidate iconic output of that word as a new entry to the dictionary—unless, of course, the entry already exists in the dictionary. In order to make the algorithm stable, we add a null word for each character class as a smoothing factor. Null words are those whose characters are all the same, *e.g.* "aaa", "bbb", "eeeeee", "11111111", etc.

4. EXPERIMENTAL DESIGN

We have conducted a test on real book-image data, in order to illustrate our model–adaptation algorithm in detail. We chose book-images from Volume 0000, page 28, of a Google Book Search Dataset[13] provided by Google, Inc. From this page, we chose: (a) thirty textlines to serve as the training set for the iconic model (six of these are shown in Figure 1(a)); (b) five textlines to serve as the training set for the linguistic model (the resulting text shown in Figure 1(b)); and (c) a single text-line to serve as the test set (Figure 1(c)). The iconic model is trained as follows: the OCR output, which is sometimes mistaken, is nevertheless assumed to be the ground-truth when training the iconic model: thus the iconic model, in its initial stage, is imperfect. The linguistic model was trained as follows: the chosen OCRed text was corrected manually, and then a single error was introduced: the word "the" was removed.

This test, though small and in some ways artificial, nevertheless has characteristics of a real whole-book recognition task: the font is consistent; many characters appear multiple times; some of the words in the test set may not be found in the dictionary; and the consequences of inaccuracies in the models occur multiple times in the test set (*e.g.* the word "the" occurs three times).

[*] These may be extracted from hOCR[3] files provided by Google.



allotted to them for that purpose. This last etymology I believe myself to be the correct one.

" The most of the people in Barra and South Uist are Roman Catholics, can neither read nor write, and hardly know any English. From these circumstances it is extremely improbable that they have borrowed much from the literature of other

(a) Part of Training Set for Iconic Model

unities During the recitation of these tale 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

(b) Complete Text Corpus for the Linguistic Model (Excluding all the word "the")

unities During the recitation of these talc the emotions of the

(c) Complete Testing Image and Pre-Segmented Character Images

Figure 1. The configuration for our test. (a) Part of the training image for the iconic model. (b) The text corpus for building the dictionary. (c) The image that is to be tested on.

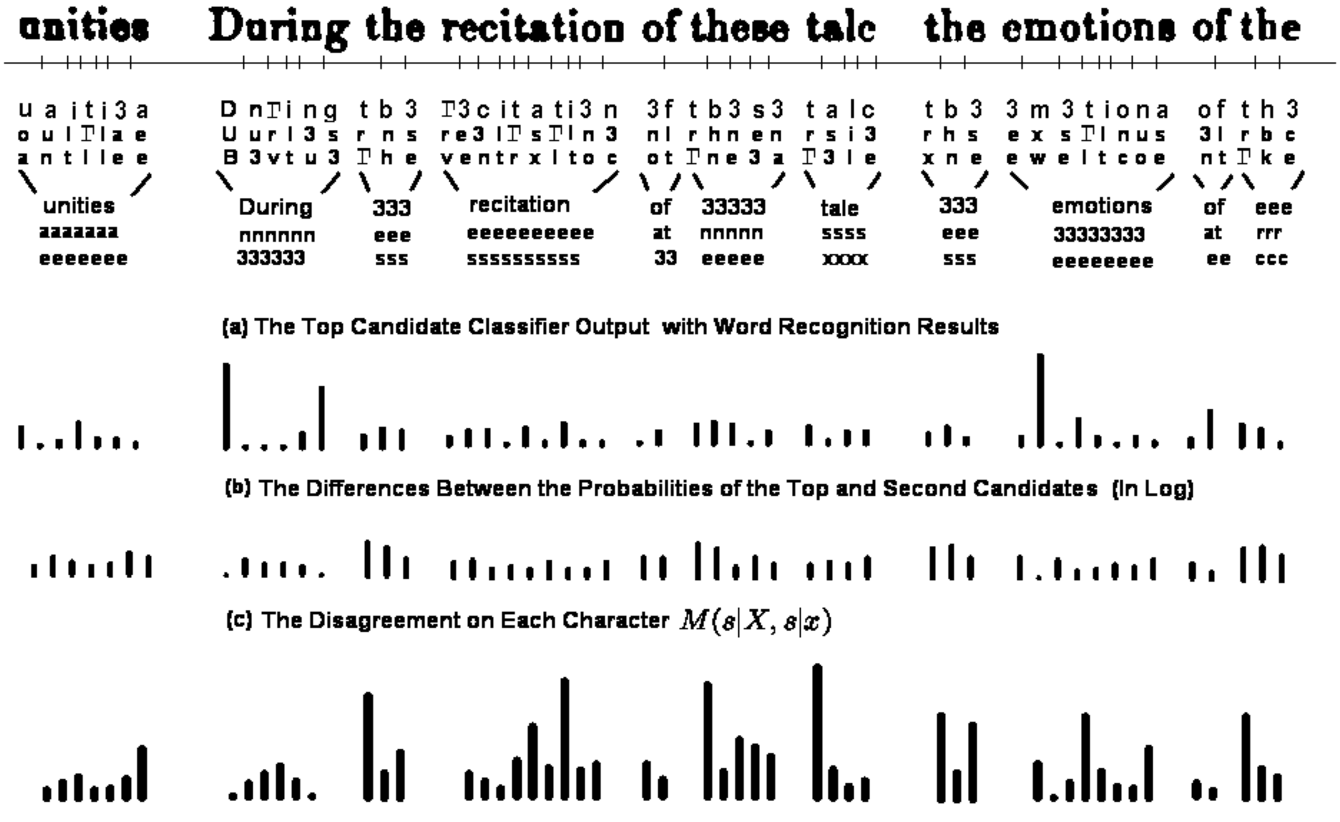
In the first stage, the iconic outputs are erroneous. By the dictionary defined in Figure 1, we can obtain a fairly good word recognition result. However, for all the "the" words, the recognition results are wrong because there is no "the" entry in the dictionary (see Figure 2). At this stage, $\sum \mathcal{M}$ is 101.22.

In the second stage, we choose the character 1 ('t')of word 7 ("tale") to change the iconic model because it has the largest $M_c(x)$ in stage 1, and it can achieve a better $\sum \mathcal{M}$ as 101.06, which is smaller than the first stage 101.22.

In the third stage, we first try the character 1 ('t') of word 3 ("the") to change the model since it has the largest $M_c(x)$. However, the $\sum \mathcal{M}$ it achieved is 112.74, larger than that of the previous stage, so we should not apply model change on this character. The character with the second largest $M_c(x)$ is character 7 ('t') of word 4("recitation"), the $\sum \mathcal{M}$ after applying iconic model changing on that character is still larger than previous stage (101.31), so we do not choose it either. Similarly, the character with the third largest $M_c(x)$ still leads to a higher $\sum \mathcal{M}$. The character with the fourth largest $M_c(x)$, character 3 ("3") of word 8 ("tb3"), however, gets a lower $\sum \mathcal{M}$, 97.19. Therefore we choose this character for the model changing action at this stage.

We follow the same routines, and do twenty-four stages of iconic model changing in total, until we can not find any character with which the iconic model changing leads to a lower $\sum \mathcal{M}$. In this stage, the recognition status is shown in Figure 3.

Since we cannot move any further by changing the iconic model, we switch to change the linguistic model. By examining the $\mathcal{M}/T$ measurements shown in Figure 3, we identify the word 3 "the" is with the greatest $\mathcal{M}/T$. We further identify that word 8 and word 11 are "close" to word 3 since the top three candidates of characters of word 8 and 11 contains 't','h','e' separately. So we find a cluster of words that has similar iconic outputs. The sum of $\mathcal{M}/T$ of this cluster is 9.84, higher than 3.99, the largest iconic disagreement $M(s|X, s|x)$(from character 2 of word 8). So we start to change the linguistic model based on word 3. In word 3, 8 and 11, the candidates of the character classes with the highest rankings are 't','h' and 'e'. We combine them and make a new entry "the", then insert it into the dictionary. Since in the dictionary, there is no "the", so the insertion succeeds. With the new linguistic model, the $\sum \mathcal{M}$ drops from 72.26 to 57. 23. And the word recognition results (first word candidate of each word image)are : *"unities During the recitation of these tale the emotions of the"*. (See figure 4)



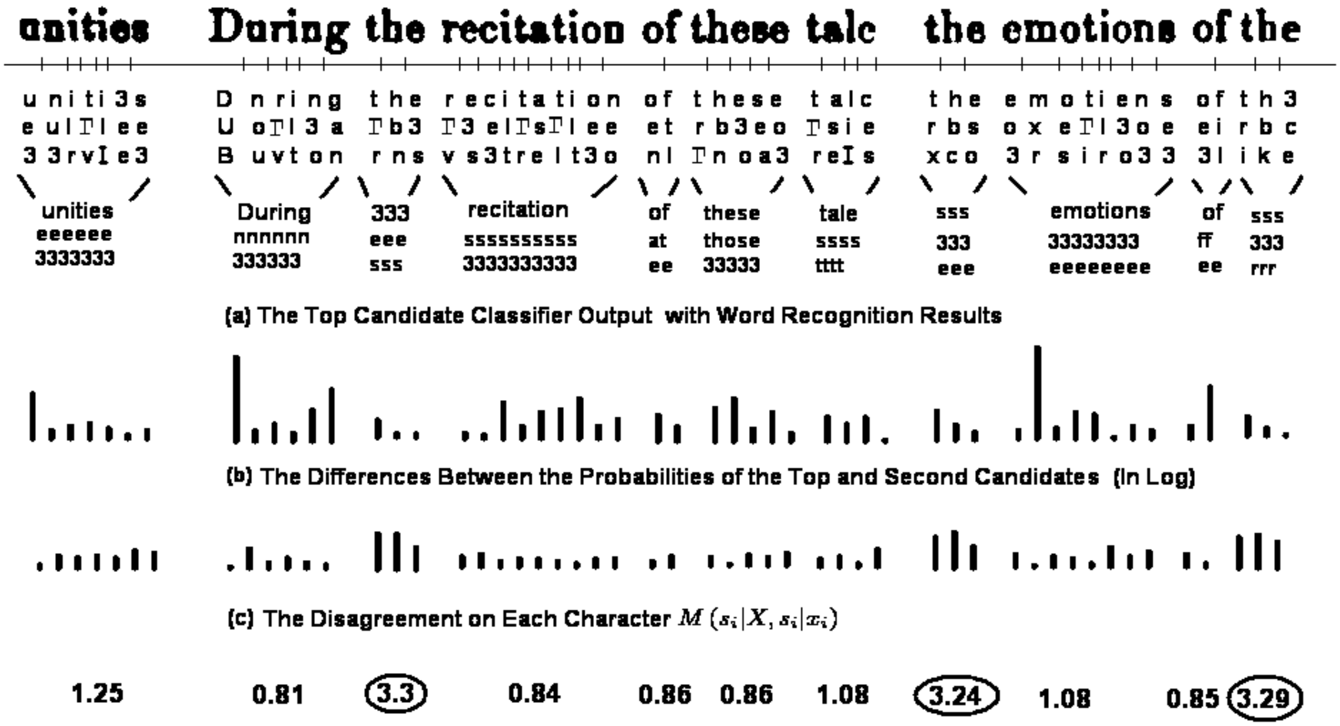
(a) The Top Candidate Classifier Output with Word Recognition Results

(b) The Differences Between the Probabilities of the Top and Second Candidates (In Log)

(c) The Disagreement on Each Character $M(s|X, s|x)$

(d) The Total Disagreements in the Neighbourhood of Each Character $M_c(x\ )$

Figure 2. The recognition result in stage 1. The total disagreement $\sum \mathcal{M}$ is 101.22 in this stage. (a) The iconic model's output are listed, with the top three candidate character classes and top three candidate words. (b) The more difference between the probabilities of the top and second candidates, the more confident the top choice is. So from this graph, we can learn several characters are recognized with high confidence, like 'D', 'g' and 'm'. (c) Comparing to (b), the characters that are recognized with higher confidence usually have lower $M(s|X, s|x)$ because they have more confidence to select the word recognition result, so that the word *a posteriori* probability distributions projected on those characters are closer to their image classification distributions. (d) The character 't's have the higher $M_c(x)$ than other characters, with the first character of word 7 having the greatest $M_c(x)$.



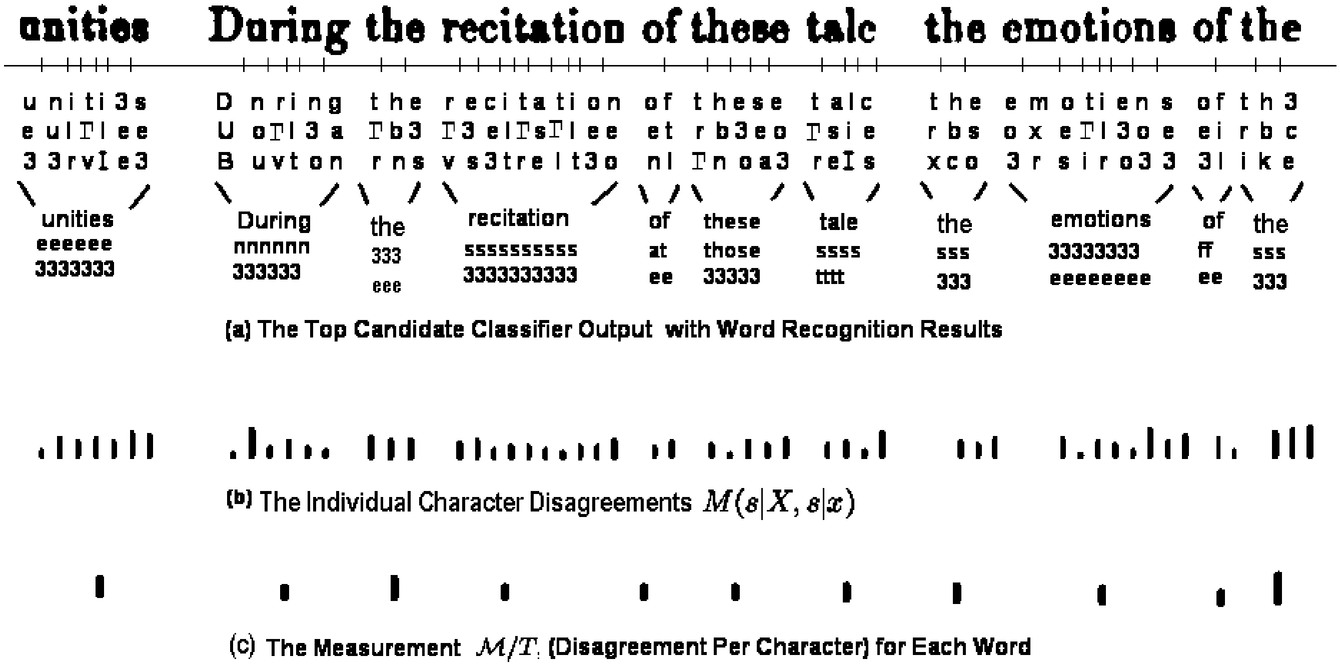
(a) The Top Candidate Classifier Output with Word Recognition Results

(b) The Differences Between the Probabilities of the Top and Second Candidates (In Log)

(c) The Disagreement on Each Character $M\,(s_i|X, s_i|x_i)$

(d) The Measurement $\mathcal{M}/T$ (Disagreement Per Character) for Each Word

Figure 3. The recognition status of the last stage for iconic model changing. In this stage, the total disagreement $\sum \mathcal{M}$ is 72.26. (a) For most characters, the iconic model gives the correct answer. (b) The difference between the probabilities of the top and second candidates indicates the recognition confidence. From the graph, we know that for most characters, the recognition confidence increases, which suggests a better iconic model. (c)The disagreements are now concentrated on the word “the”, which is not in the dictionary. (d) Under the $\mathcal{M}/T$ measurements, the word “the”s, which have higher $\mathcal{M}/T$, are easily distinguished from other words.



(a) The Top Candidate Classifier Output with Word Recognition Results

(b) The Individual Character Disagreements $M(s|X, s|x)$

(c) The Measurement $\mathcal{M}/T$ (Disagreement Per Character) for Each Word

Figure 4. The recognition status of the linguistic model changing. (a) Three "the" have been correctly recognized. (b) $M(s|X, s|x)$ for each character. (c) $\mathcal{M}/T$ for each word. The graphs of (b) and (c) are in the same scale.

After this stage, we try to select other words to try changing the linguistic model. However, we find that after do a clustering and sum the $\mathcal{M}/T$ in the cluster of that word, none of them are able to exceed the threshold (in our algorithm, the highest character disagreement $M(s|X, s|x)$ in the text). As a result, the whole process terminates.

So far, we have run an algorithm that terminates successfully, with a very remarkable improvements for both iconic and linguistic models. The iconic model gives higher character classification rate, and the linguistic model finds its missing entry "the". Also, the recognition rate for the whole process on the test image improves (in this case, one hundred percent accurate). From this small example, we illustrate the basic principle of the process in which the mutual entropy measurements help to improve the models. Generally speaking, the mutual entropy measurements identify the characters or words that are with high model disagreements, and then evaluates the soundness of various temptations to change the models, based on which we can select the best one to make the changing.

5. DISCUSSION AND FUTURE WORK

The small test case on real image data described here illustrates one way in which mutual-entropy-based measures can be effective in identifying and ranking disagreements between linguistic and iconic models. Further, our framework motivates policies for deciding, fully automatically, which corrections to the models should to be made in order to drive the recognition system towards lower error rates. The particular method we chose for this text case is greedy and we do not yet possess a proof that it will converge to a global maximum of system performance.

There are reasons to believe that isogeny is more effective in driving recognition as the passage to be recognize lengthens.[10] Thus whole-book recognition is an attractive application for this approach.

In the future, we may go into following ways:

1. Scaling up our experiments: when the scale goes up, we may not be able to do a complete $\sum \mathcal{M}$ in each stage because of the computation cost. We may consider some approximate computational methods to calculate $\sum \mathcal{M}$ more quickly.

2. Develop the iconic models. The single-template iconic model may be too simple to fit the large scale experiment. However, the more templates each character has, the harder to optimize the model. We may design some automatic adaptation approach to find the best number of templates for each character.
3. In our experiment, we assume that the words of the same length have equal probabilities. In the future we may calculate the word probabilities in the linguistic model based on the statistics in the corpus .
4. Try and compare various policies for changing the models. The disagreement measurements only provides a framework, and there are various ways to implement a mutual-entropy-based auto adaptation system.
5. Try to incorporate segmentation model into our framework.
6. Try to apply as few model changes as possible to achieve maximal reduction in overall disagreement $\sum \mathcal{M}$.

Acknowledgements

We are grateful for insights offered by George Nagy, Dan Lopresti, David Stork, Ashok Popat, Prateek Sarkar, and Dar-Shyang Lee.

REFERENCES

1. In challenge to Google, Yahoo will scan books. In *New York Times*, October 2005.
2. 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.
3. Thomas Breuel. The hOCR Microformat for OCR Workflow and Results. In *Proceedings, IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR'07)*, Curitiba, BRAZIL, August 2007.
4. Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification, 2nd Edition*. Wiley, New York, 2001.
5. Tao Hong. *Degraded Text Recognition Using Visual And Linguistic Context*. PhD thesis, 1995.
6. 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.
7. 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.
8. 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.
9. G. Nagy S. V. Rice and T. A. Nartker. *OCR: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers, 1999.
10. P Sarkar. *Style Consistency in Pattern Fields*. PhD thesis, Rensselaer Polytechnic Institute, May 2000.
11. 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.
12. P. Sarkar and G. Nagy. Style consistent classification of isogenous patterns. *IEEE Trans. on PAMI*, 27(1), January 2005.
13. Luc Vincent. Google Book Search: Document understanding on a massive scale. In *Proceedings, IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR'07)*, Curitiba, BRAZIL, August 2007.