A Statistical Mask-Matching Approach for Recognizing Handwritten Characters in Chinese Paleography*

Te-Wei Chiang
Department of Accounting Information Systems
Chihlee Institute of Technology
Taipei, Taiwan, R.O.C.
ctw@mail.chihlee.edu.tw

Tienwei Tsai
Department of Information Management
Chihlee Institute of Technology
Taipei, Taiwan, R.O.C.
twt@mail.chihlee.edu.tw

Abstract - *Because most of the characters in the rare books transcribed by ancient calligraphers were contaminated by noise, to retrieve information from these books, we resort to the statistical approach instead of the structural approach. In our approach we generate one positive mask and one negative mask for each distinct class of characters in the training phase. The positive mask of a class of characters is built by finding the bits of the character images that are reliably black. Likewise, the negative mask is built by finding the bits that are reliably white. Then, we can recognize an unknown character by finding the prototype character whose masks are best fitted for the unknown character. Experimental results show that our approach performs well in this application domain.*

Keywords: Handwriting recognition, classification, mask-matching approach.

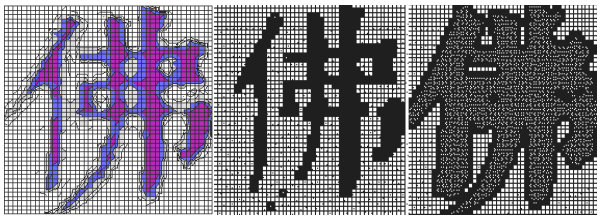
1 Introduction

In ancient Chinese a lot of significant books were transcribed by elite calligraphers. Even though the characters in these rare books were handwritten in regular forms, their integrity is greatly damaged by various imperfections. The strokes are of various thicknesses, even within the confines of a single character. Additionally, characters accompanied by spots and smears outside the theoretical limits of the characters and voids within them produce enormous variations in their shapes, even though they remain legible to human readers.

In order to retrieve information from these books, transforming paper documents into the contents that are accessible is inevitable. One way of digitizing the contents of documents is through human typing. However, this kind of labor-intensive approach is not only time consuming, but also expensive. For over the last decade, optical character recognition (OCR) technique [10] has been introduced as a practical approach for converting paper documents into computer codes. There are two distinct philosophies of OCR in general, known as "statistical" and "structural". In the statistical approach the measurements that describe an object are treated only formally as statistical objects (variables), neglecting their "meaning". On the other hand, structural information has long been suggested as a useful feature for OCR [1], [11] since it contains all the units of characters. The structural approach attempts to recognize characters the same way we generate them – as a sequence or at least a combination of strokes, straights lines, and curves. In practice, few algorithms can consistently extract structural information from characters and most structural matching approaches, however, depend largely on the correctness of the structural information. Moreover, most of the characters in these rare books were contaminated by various noises. It makes the structural approach inapplicable. Therefore, we resort to the statistical approach.

In the literature, some methods had proposed for OCR, such as methods based on neural networks (NNs) [2], [6], graph-matching methods [5], and methods based on hidden Markov model [9], etc. Due to the ease of realization, neural networks seem to be the most popular method; however, its result is not as good as the basic nearest neighborhood method [14], especially in the recognition of characters with low quality. Even though graph-matching method has strong theoretically background, its effect is not satisfactory because of the difficulty in combination with feature extraction methods. Methods based on hidden Markov model have been employed to cope with the recognition of connected characters (such as English characters). However, Chinese handwriting recognition is much more difficult than

Figure 1. Normalized sample data.



(a) (b) (c)

Figure 2. Mask generation. (a) Superimposed characters. (b) The positive mask. (c) The negative mask.

alphabetic characters. The difficulty is due to three factors: the character set is very large, the structure of a Chinese character is much more complex than that of an alphabetic character, and many Chinese characters have similar shapes.

A survey of optical recognition of handwritten Chinese characters can be found in [4]. Handwriting recognition is divided into online [5] and offline [2], [6], [12] categories. The online recognition takes the advantage of capturing the temporal and dynamic information of handwriting, such as the number of strokes, the order of the strokes, the direction of the writing for each stroke, and the speed of the writing within each stroke. As contrasted with online recognition, the offline case is much more difficult because it lacks the above mentioned knowledge. In this paper, we focus on offline recognition.

"Mask matching" is the most obvious technique for visual pattern recognition. Imagine the unknown image being projected through a cutout (mask) onto a photosensor array. The response will be proportional to the degree of matching. Based on this concept, we can store a mask in the computer for each distinct class of characters to be recognized, and to compare the unknown characters with the stored set to find the best matching. Since the scanned characters are supposed to be two-tone – all black on white backgrounds. The matching degree can be calculated by counting their matching bits, which is known as the Hamming distance criterion. However, this kind of global Boolean mask is unreliable. It is because the bits along the edge of a character image are often subject to unpredictable variations and noises.

Based on above observations, we were motivated to devise a statistical mask-matching approach for recognizing handwritten characters in Chinese paleography. In our approach we generate one positive and one negative mask for each distinct class of characters in the training phase. We superimpose a number of images of the "same" character and calculate the fraction of time that a given bit is black. The bits that are reliably black are served as the positive mask. In other words, the positive mask is built by finding the core bits of the character images. The core bits are the most central and reliable bits, shielded from the edge noise by lying deep within the characters. In contrast, the negative mask is built by finding the bits that are reliably white. Then, we can recognize an unknown character by finding the prototype character whose masks are best fitted for the unknown character. Experimental results show that our approach performs well in this application domain.

This paper is organized as follows. The next section introduces how we generate positive and negative masks. Section 3 describes our statistical mask matching approach. Section 4 presents experimental results. Finally, conclusions are drawn in Section 5.

2 Mask generation

Before getting into the main topic, let us consider briefly the normalization issue. The mask matching method can be regarded as a variant of the template matching method. In the template matching method, usually, normalization of position and size is done prior to the matching process [7], [13]. Template matching is based on the assumption that the position of each point is known, the correspondence of the sampling point is already taken, and the similarity is measured only by the difference of the sampling "value". Therefore, the normalization process becomes very important in order to fix the sampling position in advance. Before our classification process, we use a simple geometrical transformation to "normalize" the circumscribed rectangle about a handwritten character to make the shapes of the given characters more uniform. Figure 1 shows four normalized sample data extracted from the rare books. We can find that the left two characters are clear and the others are contaminated.

We know that the border bits are the most unreliable; the bits at the edge of a character image are often subject to writing and scanning noise. We can see this by superimposing a number of images of the "same" character and calculating the fraction of time that a given bit is black. What we need to find are those bits that are reliably black in certain characters and reliably white in others. For example, if we superimpose the same four characters (means "Buddha"), we obtain the result of Figure 2. From Figure 2(a) we see that the unreliable bits are along the edges. The darkest bits in the figure

represent the bits that are the most reliably black, i.e., the darker the bits are the more reliably black the bits are. Then, the darkest bits form a positive mask of, as shown in Figure 2(b), and the pure white bits generate a negative mask (see Figure 2(c)).

In the general case we use thresholds α and β to define the meaning of "reliably black" and "reliably white", respectively. For example, as α = 0.9, a bit is regarded as reliably black when the percentage of the corresponding bits of the superimposed characters is above 90%.

3 Statistical mask-matching approach

Most commonly used optimization methods in statistical approach are based on Bayes' theorem. Our mask-matching approach is also derived from Bayes' theorem.

3.1 Bayes classificaton

In statistical pattern recognition, we recognize that features may be measured with error and that some of the features are useful for identification of the class while others are not. Our goals are then to obtain useful sets of features and to use these features such that the identification is as accurate as possible. If there is an object that is to be classified on the basis of a feature x, into N possible classes (c_1, c_2, …, c_N), then the probability of x in class i when x is observed can be described by $P(c_i|x)$. From the "theorem on compound probabilities" [3], we obtain

$$P(x\&y) = P(x|y)\cdot P(y) = P(y|x)\cdot P(x) \tag{1}$$

In our situation, x is the feature and y represents the class variable c_i. Substituting for x and y in Eq. (1), we obtain the probability that the class is i when the feature x is observed.

$$P(c_i \mid x) = \frac{P(x \mid c_i)P(c_i)}{P(x)} \tag{2}$$

This is Bayes' theorem, which gives the probability of a class i being present when a feature x is observed, provided we know the probability of the feature being observed when the class is present, the probability of that class being present, and the probability of that feature.

3.2 Measures for mask matching

Before the mask-matching process, we have to define a measure to indicate the degree of matching between a sample character and a mask.

Suppose the character images and the masks are of the same size ($n \times n$ bitmap). The black bits are those bits with value 1 in the bitmaps, and white bits are those with value 0. Let $N_b(p)$ be the number of black bits in bitmap p, and $M_b(p, q)$ be the number of black bits with the same positions in both bitmap p and bitmap q. Then, the degree of matching between an unknown character x and the positive mask of class i, m_+^i, can be defined by:

$$d_+(x, m_+^i) = \frac{M_b(x, m_+^i)}{N_b(m_+^i)} \tag{3}$$

Similarly, let $N_w(p)$ be the number of white bits in bitmap p, and $M_w(p, q)$ be the number of white bits with the same positions in both bitmap p and bitmap q. Then, the degree of matching between character x and the negative mask of class i, m_-^i, can be defined by:

$$d_-(x, m_-^i) = \frac{M_w(x, m_-^i)}{N_w(m_-^i)} \tag{4}$$

Definition 1. If x matches to the positive mask of class i at the degree of α, i.e., $d_+(x, m_+^i) = \alpha$. It is called x α-match the positive mask of class i, and denoted by $x_{\alpha+}^i$.

Definition 2. Similarly, if x matches to the negative mask of class i at the degree of β, i.e., $d_-(x, m_-^i) = \beta$. It is called x β-match the negative mask of class i, and denoted by $x_{\beta-}^i$.

3.3 Statistical mask-matching

If there is an unknown character x that is to be classified on the basis of its matching result into one of N possible classes (c_1, c_2, …, c_N), then the probability of x in class i when $x_{\alpha+}^i$ is observed can be described by $P(c_i \mid x_{\alpha+}^i)$. From Bayes' theorem, we get

$$P(c_i \mid x_{\alpha+}^i) = \frac{P(x_{\alpha+}^i \mid c_i)P(c_i)}{P(x_{\alpha+}^i)} \tag{5}$$

This equation gives the probability of a class i being present when the positive mask of the class is matched, provided we know the probability of the mask being matched when the class is present, the probability of that class being present (called the "prior probability"), and the probability of that mask being matched. Similarly, we get

$$P(c_i \mid x_{\beta-}^i) = \frac{P(x_{\beta-}^i \mid c_i)P(c_i)}{P(x_{\beta-}^i)} \tag{6}$$

Finally, to decide the expected class of the input pattern x, the following decision rules are used:

1) Rule *PMD* (Positive Matching Degree):

$$E(x) = \arg\max_{1 \le i \le N} \{ d_+(x, m_+^i) \} \quad (7)$$

2) Rule *NMD* (Negative Matching Degree):

$$E(x) = \arg\max_{1 \le i \le N} \{ d_-(x, m_-^i) \} \quad (8)$$

3) Rule *PMP* (Positive Matching Probability):

$$E(x) = \arg\max_{1 \le i \le N} \{ P(c_i \mid x_{\alpha+}^i) \} \quad (9)$$

4) Rule *NMP* (Negative Matching Probability):

$$E(x) = \arg\max_{1 \le i \le N} \{ P(c_i \mid x_{\beta-}^i) \} \quad (10)$$

5) Rule *AMP* (Average Matching Probability):

$$E(x) = \arg\max_{1 \le i \le N} \{ \left(P(c_i \mid x_{\alpha+}^i) + P(c_i \mid x_{\beta-}^i) \right) / 2 \} \quad (11)$$

where $\alpha = d_+(x, m_+^i)$ and $\beta = d_-(x, m_-^i)$.

The concept underlying rules PMD and NMD is very intuitively. The class whose mask matches the input pattern most will be regarded as the expected class of the input pattern. On the other hand, rules PMP and NMP select the most likely class as the expected class of the input pattern according to the probability of the class being present when its mask is matched. As to rule AMP, it takes average of the probabilities evolved from the positive and negative masks of a class. Through the combination of different viewpoints, we expect AMP will outperform the other rules.

4 Experimental results

A preliminary experiment has been made to test our approach. There are a total number of 6000 samples (about 500 categories) extracted from one of the famous handwritten rare books, Kin-Guan bible. Each character pattern was transformed into a 20×20 bitmap pattern by means of simple size normalization. The positive and the negative masks for each character category were generated by superimposing the character patterns of the same category. The thresholds α and β used to generate masks are equal to 0.8. 1000 of the 6000 samples are used for testing. Table I shows the recognition rate for each decision rule. As we expected, AMP outperforms the other rules. It is worthy of notice that NMD performs better than most of the other rules.

TABLE I
Recognition Rate for Each Decision Rule

Decision Rule	Number of Candidates		
	Top 1	Top 3	Top 10
PMD	76.16%	86.57%	90.79%
NMD	85.05%	91.44%	94.26%
PMP	82.23%	88.95%	92.96%
NMP	83.86%	90.03%	94.47%
AMP	87.87%	91.77%	94.80%

5 Conclusions

This paper presents a statistical mask-matching approach for recognizing handwritten characters in Chinese paleography. After generating the masks for each prototype character and calculating some prior probabilities in advance, we can obtain the probability of a class being present when the mask of the class is matched. In our preliminary experimental results, the recognition rate is about 88 percent for a unique candidate, and 95 percent for multichoice with 10 candidates.

Since only preliminary experiment has been made to test our approach, a lot of works should be done to improve this system:

- For accurate and reliable character recognition we require at least 8 or 10 pixels per millimeter. Thus in the space of a single Chinese character, say 4×4mm, it seems that for a single Chinese character 40×40 bitmap is more appropriate than current 20×20 bitmap.
- Since features of different types complement one another in classification performance, by using features of different types simultaneously, classification accuracy could be improved.
- In order to distribute the load of the character recognition, clustering techniques need to be involved in our system.

References

[1] F. Chang, Y. C. Lu, and T. Pavlidis, "Feature analysis using line sweep thinning algorithm," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 2, pp. 145-158.

[2] J. P. Drouhard, R. Sabourin and M. Godbout, "A neural network approach to off-line signature verification using directional PDF," *Pattern Recognition*, Vol. 29, No. 3, pp. 415-424, 1996.

[3] W. Feller, "An Introduction to Probability Theory and its Applications," vol. 1, 2nd ed., Wiley, New York, 1957.

[4] T. H. Hildebrandt and W. Liu, "Optical Recognition of Handwritten Chinese Characters: Advances Since 1980," *Pattern Recognition* , vol. 26, no.2, pp. 205-225, 1993.

[5] A. J. Hsieh, Kuo-Chin Fan, and Tzu-I Fan, “Bipartite Weighted Matching for On-line Handwritten Chinese Character Recognition," *Pattern Recognition,* Vol. 28, no. 2, pp. 143-151, 1995.

[6] S. W. Lee, “Off-line recognition of totally unconstrained handwritten numerals using multilayer cluster neural network,” *IEEE Trans. Patt. Anal. Machine Intell.*, Vol. PAMI-18, no.6, pp. 648-652, 1996.

[7] S. W. Lee and J. S. Park, “Nonlinear Shape Normalization Methods for the Recognition of Large-Set Handwritten Characters,” *Pattern Recognition* , vol. 27, no. 7, pp. 859-902, 1994.

[8] C. Y. Liou and H. C. Yang, “Handprinted character recognition based on spatial topology distance measurement,” *IEEE Trans. Patt. Anal. Machine Intell.*, Vol PAMI-18, no.9, pp. 941-945, 1996.

[9] M. Mohamed and P. Gader, “Handwritten word recognition using segmentation-free hidden Markov modeling and segmentation-based dynamic programming techniques,” *IEEE Trans. Patt. Anal. Machine Intell.*, Vol PAMI-18, no.5, pp. 548-554, 1996.

[10] M. Nadler and E. P. Smith, “Pattern Recognition Engineering,” Wesley Interscience, 1993.

[11] C. Y. Suen, R. Legault, C. Nadal, M. Cheriet and L. Lam, “Building a new generation of handwriting recognition systems,” *Pattern Recognition Letters*, vol. 14, pp. 303-315, 1993.

[12] Y. Y. Tang, L. T. Tu, J. Liu and S. W. Lee, “Offline recognition of Chinese handwriting by multifeature and multilevel classification,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 5, pp. 556-561.

[13] T. Wakahara, K. Odaka, “Adaptive Normalization of Handwritten Characters Using Global/Local Affine Transformation,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no.12, pp. 1332-1341, 1998.

[14] Hong Yan, “Comparision of Multilayer Neural Network and Nearest Neighbor Classifiers for Handwritten Digit Recognition," *Int. J. of Neural Systems,* Vol. 6, No.4, pp. 417-423, 1995.