Geometrical-Statistical Modeling of Character Structures for Natural Stroke Extraction and Matching

Xiabi Liu, Yunde Jia, Ming Tan

Department of Computer Science and Engineering, Beijing Institute of Technology,
Beijing 100081, P. R. China
{liuxiabi, jiayunde, tanming_82}@bit.edu.cn

Abstract

This paper proposes a model-based approach to extract natural strokes in handwritten Chinese character images. We model the distortion between an input stroke and its counterpart in the character model as the result of an affine transformation followed an uncertainty distribution. Based on this modeling, a statistical method is presented to measure similarities between strokes or characters, and a probabilistic relaxation process is put forward to assign line segments in the skeleton of the input character to model strokes transformed by detected affine transformation. All line segments assigned to a model stroke constitute the corresponding input stroke. For achieving high robustness, affine transformation estimation and stroke extraction are performed alternately until the similarity between the input character and the character model is maximized. The proposed stroke extraction and matching method was applied to off-line handwritten Chinese character recognition, whose effectiveness is confirmed by the experimental results.

Keywords: Stroke extraction, stroke matching, handwritten Chinese character recognition.

1. Introduction

Stroke matching methods are promising to solve the hard problem of off-line handwriting recognition. In this category of solutions, stroke extraction from character images is a major difficulty. Therefore, many works and fruitful results on stroke extraction have been reported. Stroke ambiguity, such as amphibolous stroke breaking [1], is an obstacle to the reliability of stroke extraction. Since they can be overcome by referring to character models, model-based approaches to stroke extraction were introduced and developed [2-5].

Intuitively, a stroke means a natural stroke which is a trajectory of pen-tip from pen-down to pen-up in regular writing. It's highly difficult to extract natural strokes from character images, so a stroke is defined as a line stroke, a line segment or a short segment more often. In this paper, we propose a new model-based approach to natural stroke extraction and matching for off-line handwritten Chinese character recognition. In the proposed approach, characters are modeled as sets of natural strokes, each of which is represented as a polygonal line with ordered vertices. The distortion between the input stroke and the corresponding model stroke is assumed as the result of two factors: an affine transformation from the character model to the input character, and an uncertainty between the transformed model stroke and the corresponding input stroke. By modeling the uncertainty as the Gaussian Mixture, a method is presented to measure similarities between strokes or characters. We further put forward a probabilistic relaxation process to extract input strokes and match them with model strokes by assigning line segments in the skeleton of the input character to model strokes transformed by detected affine transformation. For getting robust result, affine transformation estimation and stroke extraction are performed alternately to maximize the character similarity. An eigenvector based point matching method [6] is used to compute the initial affine transformation for starting this alternating optimization process. In order to test the effectiveness of the proposed approach to stroke extraction and matching, we conducted experiments of off-line handwritten Chinese character recognition. The recognition rate on HCL2000 database of handwritten Chinese characters [7] is above 92%, which shows that the proposed approach is effective and promising.

The rest of this paper is organized as follows. Section 2 introduces the measures of similarities between strokes or characters. Section 3 presents the probabilistic relaxation process for stroke extraction and matching. Section 4 describes the alternation of estimating affine transformation and extracting strokes. Experimental results on handwritten Chinese character recognition are discussed in Section 5 and conclusions are given in Section 6.

2. Stroke and character similarity measure based on uncertainty modeling

The similarity between two strokes, or equivalently the dissimilarity between them, is key information for stroke extraction and matching. We measure stroke similarity according to the statistical modeling of the uncertainty between the input stroke and the corresponding model stroke, where the affine transformation between two strokes is supposed to be removed.

Since strokes are represented as polygonal lines, we uniformly sample points in two polygonal lines and

define a feature vector based on sample points to represent the distortion between strokes. Given a polygonal line s and a point p_i on s. At p_i, the local feature of s can be described by a 3-dimensional vector (x_{i1}, x_{i2}, x_{i3}), where

x_{i1}, x_{i2}: horizontal and vertical coordinates of p_i

x_{i3}: tangent slope angle at p_i.

$x_{i3} = \arg((x_{i+1} - x_{i-1}) + J(y_{i+1} - y_{i-1}))$, with $J^2 = -1$ and "arg" the phase of the complex number above, is an approximation of the tangent slope angle at the point [8].

Let a pair of points in two polygonal lines be p_i and p_i', the difference between them can be computed as:

$$\begin{aligned}\mathbf{x_i} &= (\Delta x_{i1}, \Delta x_{i2}, \Delta x_{i3}) \\ &= (x_{i1} - x'_{i1}, x_{i2} - x'_{i2}, x_{i3} - x'_{i3}).\end{aligned} \quad (1)$$

Suppose there are n pairs of sample points in two polygonal lines, the distortion between two polygonal lines can be represented as a $3n$-dimensional vector which is

$$\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}). \quad (2)$$

We assume $\mathbf{X}$ for a model stroke s and its instance s' in the input character is of the distribution of the Gaussian Mixture in this work. Let $\boldsymbol{\mu}_\mathbf{k}$, $\boldsymbol{\Sigma}_\mathbf{k}$, w_k respectively be the mean vector, the covariance matrix and the weight of the k th Gaussian component in the Gaussian Mixture Model, then we have

$$p(\mathbf{X}|s = s') = \sum_{k=1}^{K} w_k N(\mathbf{X}|\mu_k, \Sigma_k), \quad (3)$$

where

$$\begin{aligned}&N(\mathbf{X}|\mu_k, \Sigma_k) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{X} - \mu_k)^T \Sigma_k^{-1} (\mathbf{X} - \mu_k)\right).\end{aligned} \quad (4)$$

Using Bayes' formula, we get

$$P(s = s'|\mathbf{X}) = \frac{p(\mathbf{X}|s = s')P(s = s')}{p(\mathbf{X})}. \quad (5)$$

It's reasonable to use the probability $P(s = s'|\mathbf{X})$ to measure the similarity between s and s'. In Eq. 5, $p(\mathbf{X}|s = s')$ can be estimated from the sample pairs of strokes, but it's very difficult to learn $P(s = s')$ and $p(\mathbf{X})$ since contrary cases are too diverse to be handled. However, they are fixed for all pairs of strokes because $\mathbf{X}$ can be assumed to be distributed uniformly in the cases without apriori knowledge, thus

$$P(s = s'|\mathbf{X}) \propto p(\mathbf{X}|s = s'). \quad (6)$$

According to Eq. 6, we can estimate $P(s = s'|\mathbf{X})$ by embedding $p(\mathbf{X}|s = s')$ in a smooth, monotonically increasing function which takes value in $[0,1]$. The value of this kind of functions is called the pseudo-probability, for computing which the following function is a good choice:

$$\rho(p(\mathbf{X}|s = s')) = 1 - \exp(-\alpha p(\mathbf{X}|s = s')). \quad (7)$$

Consequently, we compute the similarity between two strokes as

$$P(s = s'|\mathbf{X}) \approx \rho(p(\mathbf{X}|s = s')). \quad (8)$$

In the following description of this paper, $P(s = s'|\mathbf{X})$ is simplified to $P(s = s')$ for convenience.

Let $\mathbf{S}$ be the stroke set of the character model C, and $\mathbf{S}'$ be the extracted stroke set of the input character I, then the similarity between $\mathbf{S}$ and $\mathbf{S}'$ is measured as the joint probability that the model strokes in $\mathbf{S}$ are matched to the corresponding input strokes in $\mathbf{S}'$:

$$P(\mathbf{S} = \mathbf{S}') = P(s_1 = s_1', \cdots, s_n = s_n'). \quad (9)$$

By using the 2-order statistical independence, this high-order probability is approximated as

$$P(s_1 = s_1', \cdots, s_n = s_n') \approx \prod_{ij} P(s_i = s_i', s_j = s_j'). \quad (10)$$

$P(s_i = s_i', s_j = s_j')$ reflects the relationship between two strokes. We compute the feature vector $\mathbf{X}_{ij}$ between s_i and s_j, and the feature vector $\mathbf{X}'_{ij}$ between s_i' and s_j'. $\mathbf{Y}_{ij} = \mathbf{X}_{ij} - \mathbf{X}'_{ij}$ is used to represent the difference of the relationship between two strokes, and $P(s_i = s_i', s_j = s_j')$ is computed as $\rho(p(\mathbf{Y}_{ij}|s_i = s_i', s_j = s_j'))$.

After input strokes are extracted under the guidance of the character model, there could be unmatched parts in the input character, so the similarity between the input character I and the character model C is defined as

$$P(I = C) = P(\mathbf{S} = \mathbf{S}') \cdot \frac{L_e + L_l}{L_I}, \quad (11)$$

where L_I is total length of all line segments in the input character, L_e is total length of extracted input strokes, L_l is total length of ligature line segments in the input character, which are not included in extracted input strokes, but link two extracted ones. The definition of line segment is given in next section.

3. Stroke extraction and matching by probabilistic relaxation

Stroke extraction is performed according to the skeleton of the input character and the character model

transformed by appropriate affine transformation. The problem of affine transformation estimation will be discussed in next section. In this section, model strokes are supposed to be transformed by appropriate affine transformation. As for the problem of character image skeletonization, the principal curve based algorithm is applied to solve it [9].

Different with traditional thinning methods, the skeleton from principal curve based algorithm is a graph, which means the skeleton topology have been established and the stroke extraction will benefit from it. Fig. 1 shows a handwritten Chinese character and its skeleton computed by our algorithm, where the black dots denote the dominant points in the skeleton. The dominant points include three kinds of points: end vertices of paths, joint vertices of different paths, high-curvature vertices in paths. All paths linking only two dominant points are taken as primitives to constitute input strokes, which are called line segments for obvious reason.

Figure 1. A handwritten Chinese character image with its skeleton and dominant points.

Stroke extraction is completed by assigning line segments in the skeleton of the input character to natural strokes in the character model. All line segments assigned to a model stroke are connected with each other to become the input stroke which is matched with this model stroke. Although a model stroke could possess several line segments, a line segment can be assigned to only one model stroke.

Let $\mathbf{L}=\{l_1,\cdots,l_n\}$ be the set of line segments in the skeleton, $\mathbf{S}=\{s_1,\cdots,s_m\}$ be the set of model strokes, where l_i and s_j both are polygonal lines with ordered vertices. Considering spurious line segments, a null stroke s_{m+1} is inserted into $\mathbf{S}$: $\mathbf{S}=\mathbf{S}+\{s_{m+1}\}$.

We propose a probabilistic relaxation process to solve the assignment of $\mathbf{L}$ to $\mathbf{S}$. The core of the proposed process is a $m\times n$ probability matrix whose elements are probabilities of l_i assigned to s_j. Since a line segment is represented as a polygonal line, the initial probability of l_i assigned to $s_j, j=1,\cdots,m$ is computed as the pseudo-probability between l_i and its projection on s_j using Eq. 8 in previous section. Let l'_{ij} be the projection of l_i on s_j, $P^{[t]}(l_i\to s_j)$ be the probability of l_i assigned to s_j in the t-th iteration of the relaxation process, then we have

$$P^{[0]}(l_i\to s_j)=P(l_i=l'_{ij}). \tag{12}$$

The probability of l_i assigned to the null stroke is calculated as:

$$\begin{aligned}&P^{[t]}(l_i\to s_{m+1})\\&=1-\max\left(P^{[t]}(l_i\to s_j)\right), j=1,\cdots,m\end{aligned}. \tag{13}$$

All probabilities for a line segment are normalized because a line segment can be assigned to only one model stroke:

$$P^{[t]}(l_i\to s_j)=\frac{P^{[t]}(l_i\to s_j)}{\sum_{k=1}^{m+1}P^{[t]}(l_i\to s_k)}, j=1,\cdots,m+1. \tag{14}$$

According to the probability matrix, each line segment is assigned to the model stroke which is corresponding with the maximum probability for this line segment. If two line segments assigned to a stroke are exclusive with each other, one assignment with smaller probability is canceled. The exclusion of two assignments is determined by the overlap degree of projections of two line segments on the stroke. Let L_{ik} and L_{jk} be lengths of projections of l_i and l_j on s_k respectively, L_{op} is the length of the overlapping part of l_i and l_j projected on s_k, then the overlap degree is computed as $L_{op}/\min(L_{ik},L_{jk})$. If it's larger than some threshold, two assignments $l_i\to s_k$ and $l_j\to s_k$ are accepted to be exclusive with each other.

After all line segments are assigned to corresponding strokes, the input stroke for the model stroke s_k are extracted by linking all line segments assigned to s_k. The sequence of line segments is determined according to their projections on s_k. Based on the extraction result, the similarity between the input character and the character model is computed by using Eq. 11 in previous section. If this similarity stops growing, the process of line segments assignment will finish, or else the probability matrix is updated by the following method, and the process above is iterated.

Let $\mathbf{H}_k$ be the set of line segments assigned to the stroke $s_k, k=1,\cdots,m$ in the t-th iteration. For arbitrary line segment l_i, let $\mathbf{H}'_k=\mathbf{H}_k-\{l_i\}=\{l_{k1},\cdots,l_{km}\}$, if $\mathbf{H}'_k=\phi$, the probability $P(l_i\to s_k)$ is unchanged, or else it is

updated as

$$P^{[t+1]}(l_i \to s_k) = P(l_i \to s_k | l_{k1} \to s_k, \cdots, l_{km} \to s_k) = \frac{P(l_i \to s_k, l_{k1} \to s_k, \cdots, l_{km} \to s_k)}{P(l_{k1} \to s_k, \cdots, l_{km} \to s_k)} \quad . (15)$$

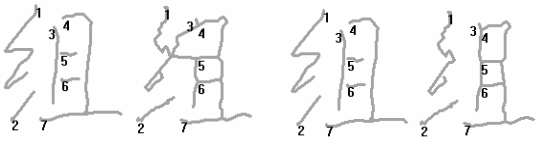
The joint probabilities that several line segments are assigned to s_k in Eq. 15 can be viewed as the similarity between the model stroke s_k and a temporary input stroke formed by linking these line segments, so which can be computed by using Eq. 8 in previous section. After all probabilities of line segments assigned to strokes except the null stroke are updated, the probabilities for the null stroke are computed using Eq. 13, and all probabilities are normalized using Eq. 14 again.

4. Affine transformation estimation

This section solves the problem of affine transformation estimation. In order to achieve robust stroke extraction and matching, we alternate between estimating affine transformation and extracting strokes. An initial affine transformation between the input character and the character model is computed using an eigenvector based point matching method [6]. Then the character model is transformed accordingly, and input strokes are extracted and matched with transformed model strokes using the probabilistic relaxation process presented in Section 3.

According to the extraction result, the affine transformation between the input character and the character model is re-estimated. We uniformly sample pairs of points in model strokes and corresponding input strokes. Based on these point correspondences, the affine transformation is computed according to the minimum square criterion. Then the character model is transformed accordingly and the probabilistic relaxation process is performed again to extract and match strokes. This alternation between affine transformation estimation and stroke extraction continues until the similarity between the input character and the character model is maximized. Fig. 2 shows the effectiveness of alternating optimization, in which the extraction result becomes better and better with alternations. Fig. 2a shows an input Chinese character and its skeleton, Fig. 2b shows the corresponding character model, Fig. 2c-e show the extraction result from the first alternation to the third (last) alternation respectively, where the left image is the character model transformed by the detected affine transformation, and the right one is extracted input strokes. The correspondences between strokes are indicated by numbers at stroke ends.

(a) (b) (c)



(d) (e)

Figure 2. The illustration of alternating optimization: (a) input character and its skeleton; (b) the corresponding character model; (c) the result after the first alternation; (d) the result after the second alternation; (e) the result after the third (last) alternation.

5. Experimental Results

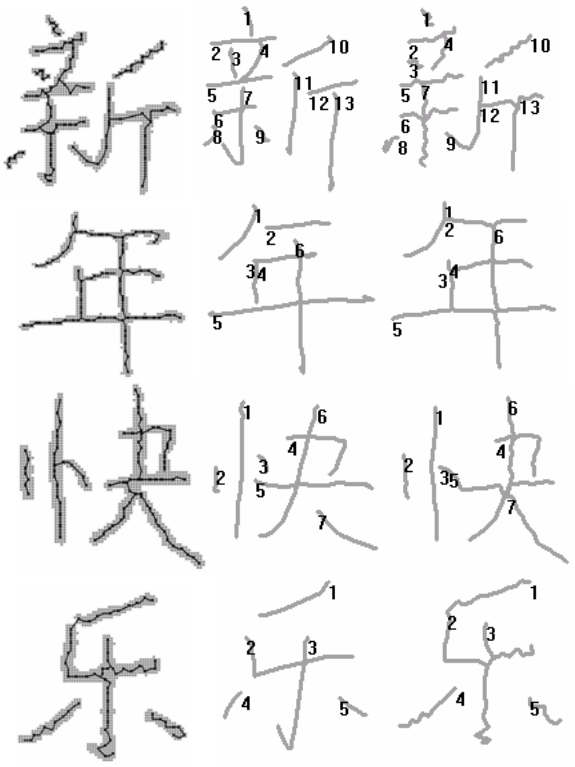
In order to evaluate the effectiveness of the proposed approach to stroke extraction and matching, we not only conducted stroke extraction experiments, but also applied it to off-line handwritten Chinese character recognition. A recognition system with two-level classification scheme was developed, in which a coarse classifier is used to generate top 10 candidates for further classification, and the proposed approach in this paper is used in fine classifier to get final result from 10 candidates. In order to exclude the influence of the coarse classifier for evaluating our approach, the correct result is guaranteed in the 10 candidates. If the correct result is not in the top 10 candidates generated by the coarse classifier, the tenth candidate will be replaced by it.

Consider an input character I and a set of character models $\mathbf{C} = \{C_1, C_2, \cdots, C_M\}$, where M is the number of models. For I and C_i, input strokes in I are extracted under the guidance of C_i by using the proposed approach, and the similarity between I and C_i is computed accordingly. After all similarities between I and character models are computed, I is classified by the following decision rule:

$$I = C_i \quad \text{if} \quad P(I = C_i) = \max_j P(I = C_j),\ j = 1, \cdots, M .$$

We implemented recognition experiments on HCL2000 database of handwritten Chinese characters, which are collected by PRAI laboratory of Beijing University of Posts and Telecommunications [7]. The database has 3763 Chinese characters; each character has 1000 sample images. We built character models by writing each character regularly on the tablet, and then used two set of samples and EM algorithm to train the Gaussian Mixture for computing stroke similarity. Another two sets of samples in the database are used to test the proposed recognition method. The resultant recognition rate is above 92%. This number is lower than results reported in recent works of handwritten Chinese character recognition using model-based stroke extraction and matching [4, 5]. However, the class number in our experiments is 3763, and that in the works mentioned above is 783. Fig. 3 shows four

examples of stroke extraction and matching for handwritten Chinese characters from HCL2000 database, where four Chinese characters are connected to mean ‘happy new year’ in English. Fig. 3a shows input characters and their skeletons; Fig.3b shows the corresponding character models transformed by detected affine transformations, where the numbers denote stroke indices; Fig.3c shows the results of extraction and matching, where each extracted stroke is labeled with the index of the corresponding model stroke.



(a) (b) (c)

Figure 3. Examples of stroke extraction and matching for handwritten Chinese characters from HCL2000 database: (a) input characters and their skeletons; (b) the corresponding character models transformed by detected affine transformation; (c) the results of extraction and matching.

6. Conclusions

In this paper, natural strokes in character images are extracted based on geometrical-statistical modeling of character structures. Characters are represented as sets of natural strokes, and the distortion between the model stroke and its counterpart in the input character is thought as the result of two factors: the affine transformation between the input character and the character model, and the uncertainty distribution between the transformed model stroke and the corresponding input stroke. Based on this modeling, affine transformation estimation and stroke extraction are alternated to get robust result. After the character model is transformed by appropriate affine transformation, the similarities between strokes are measured by a statistical method, and a corresponding probabilistic relaxation process is performed to extract strokes from the skeleton of the input character and match them with model strokes. We apply the proposed method of stroke extraction and matching to off-line handwritten Chinese character recognition, whose effectiveness is confirmed by experimental results on HCL2000 database of handwritten Chinese characters.

The main contribution of this paper is to integrate key stages of stroke matching for handwriting recognition, including transformation estimation, stroke extraction and character recognition. These procedures cooperate with each other towards a common optimization object, i.e. maximizing the character similarity. We think this is the way to achieve high robustness. Furthermore, the definition of pseudo-possibility for measuring object similarity, and a probabilistic relaxation method of stroke extraction and matching are advised in this paper.

References

[1] P.N. Suganthan, H. Yan, “Recognition of handprinted Chinese characters by constrained graph matching”, *Image Vision Computing*, Vol. 16, No. 3, 1998, pp. 191-201.

[2] H.T. Tsui, “Guided stroke structure extraction for the recognition of handprinted Chinese characters”, *Proceedings of the sixth International Conference on Pattern Recognition*, 1982, pp. 786-788.

[3] D.S. Yeung, H.S. Fong, “A fuzzy substroke extractor for handwritten Chinese characters”, *Pattern Recognition*, Vol. 29, No. 12, 1996, pp. 1963-1980.

[4] C. L. Liu, I. J. Kim, and J. H. Kim, “Model based stroke extraction and matching for handwritten Chinese character recognition”, *Pattern Recognition*, Vol. 34, 2001, pp. 2339-2352.

[5] I. J. Kim and J. H. Kim, “Statistical character structure modeling and its application to handwritten Chinese character recognition”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 25, No. 11, 2003, pp. 1422-1436.

[6] X. Liu, Y. Jia, and Y. Wang, “An eigenvector approach based on shape context patterns for point matching”, *Technical Report*, 2005.

[7] J. Guo, Z. Q. Lin, and H. G. Zhang, “A new database model of off-line handwritten Chinese characters and its applications”, *Acta Electraonic Sinica*, Vol. 28, No. 5, 2000, pp. 115-116, In Chinese.

[8] Claus Bahlmann and Hans Burkhardt, “The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 26, No. 3, 2004, pp. 299-310.

[9] X. Liu and Y. Jia, “A bottom-up algorithm for finding principal curves with applications to image skeletonization”, *Pattern Recognition*, Vol. 38, No. 7, 2005, pp. 1079-1085.