Improving Handwritten Chinese Text Recognition by Confidence Transformation

Qiu-Feng Wang, Fei Yin, Cheng-Lin Liu
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation of Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, P.R. China
{*wangqf, fyin, liucl*}*@nlpr.ia.ac.cn*

***Abstract*—This paper investigates the effects of confidence transformation (CT) of the character classifier outputs in handwritten Chinese text recognition. The classifier outputs are transformed to confidence values in three confidence types, namely, sigmoid, softmax and Dempster-Shafer theory of evidence (D-S evidence). The confidence parameters are optimized by minimizing the cross-entropy (CE) loss function (both binary and multi-class) on a validation dataset, where we add non-character samples to enhance the outlier rejection capability in text recognition. Experimental results on the CASIA-HWDB database show that confidence transformation improves the handwritten text recognition performance significantly and adding non-characters for confidence parameter estimation is beneficial. Among the confidence types, the D-S evidence performs best.**

***Keywords*-Handwritten text recognition; confidence transformation; cross-entropy; non-characters**

I. INTRODUCTION

The recognition of unconstrained handwritten Chinese texts is a great challenge due to the diversity of writing styles, large character set, the character segmentation problem caused by variable character size, confusing within-character and inter-character gaps, character touching and overlapping, etc. A general approach to overcome the ambiguity of character segmentation is to generate candidate characters by over-segmentation and search for the optimal path in the candidate segmentation-recognition lattice. Candidate paths are usually evaluated by combining character classification scores, linguistic context and geometric context [1], which can be seen as a combination of multiple classifiers. Confidence transformation (CT) converts the classifier outputs to approximate the class posterior probability, and has been shown to benefit the combination of multiple classifiers [2], [3].

Some works of confidence transformation for character recognition have been reported. Li et al. [4] used the logistic regression model to directly convert the classifier outputs into confidence values, which inherently considered the multi-class problem as multiple one-versus-all binary problems. Jiang et al. [5] got confidence values with the multi-class softmax framework, with the parameter of softmax estimated by minimizing a squared error criterion on a validation dataset [6]. These works have not considered the influence of non-characters because they either tested on isolated characters or assumed that the text strings have been segmented into characters. Handwritten text (character string) recognition involves classification of non-characters and the non-character resistance (outlier rejection) capability of the classifier is very important [7]. When the classifier outputs are transformed to confidence values, the outlier rejection capability should be taken into account. To our best of knowledge, the outlier resistance of confidence transformation has not been considered in the context of character recognition.

In this paper, we evaluate the effects of confidence transformation of character classifier in handwritten Chinese text recognition. We consider some common confidence types, namely, sigmoid, softmax and Dempster-Shafer theory of evidence (D-S evidence) [3]. The confidence parameters are estimated by minimizing the cross entropy (CE) loss function on a validation dataset. We formulate the confidence measures in multi-class as well as binary classification frameworks incorporating an outlier class, and add non-character (outlier) samples in the validation dataset to enhance the outlier resistance of transformed confidence measure. In our experiments of unconstrained handwritten Chinese text recognition on the CASIA-HWDB database using two classifiers (modified quadratic discriminant function (MQDF) [8] and nearest prototype classifier) and character-level n-gram language model, the confidence transformation of classifier outputs is shown to significantly influence the text recognition performance. Adding non-character samples in confidence parameter estimation is shown to benefit the text recognition performance. While comparing the confidence types, the D-S evidence performs best in our experiments.

II. SYSTEM OVERVIEW

Our handwritten Chinese text recognition system framework is based on our previous work [9], and the block diagram is shown in Fig. 1. First, the input text line image is over-segmented into a sequence of primitive segments (Fig. 2a) using the connected component-based method [10]. Then, consecutive segments are combined to generate candidate character patterns, forming a segmentation candidate lattice (Fig. 2b). After that, each candidate pattern is classified to assign several candidate character classes, forming a character candidate lattice (Fig. 2c). Last, each character sequence **C** paired with candidate pattern sequence **X** (the pair is called a candidate segmentation-recognition path) is

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.110

evaluated by combining the classification scores, linguistic context and geometric context.

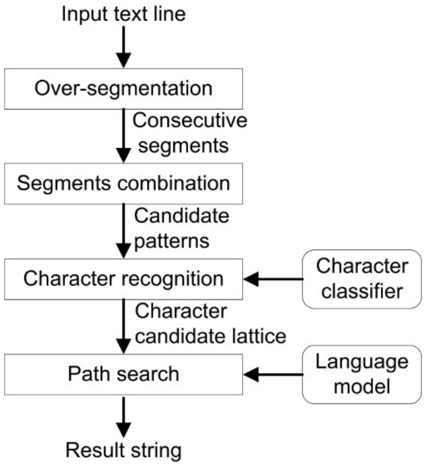
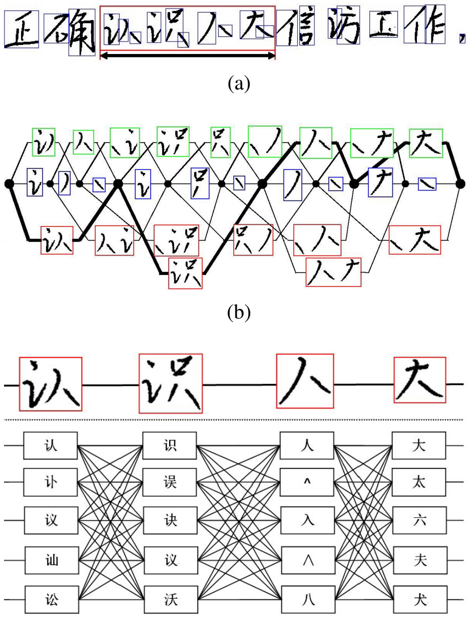


Figure 1: System diagram for handwritten Chinese text recognition.



(a)

(b)

(c)

Figure 2: (a) Over-segmentation; (b) Segmentation candidate lattice; (c) Character candidate lattice of a segmentation.

In this work, we ignore the geometric context because our focus is the effect of confidence transformation. Specifically, we evaluate the candidate segmentation-recognition path by

$$f(\mathbf{X},\mathbf{C}) = \log P(\mathbf{C}) + \lambda \cdot \sum_{i=1}^{L} k_i \cdot \log P(c_i|\mathbf{x_i}). \qquad (1)$$

This formula is modified from $\log P(\mathbf{C}|\mathbf{X})$, and $p(c_i|\mathbf{x_i}) = p(\mathbf{x_i}|c_i)p(c_i)/p(\mathbf{x_i})$ is calculated by confidence transformation of the character classifier. The probability $P(\mathbf{C})$ is given by the character-level tri-gram language model. The combining weight λ is optimized by a method similar to Minimum Phone Error (MPE) training [11]. To overcome the bias of $P(\mathbf{C}|\mathbf{X})$ to small number of segmented characters, we weight the classification score of each character pattern with its number k_i of constituent segments (similar to the variable length HMM of [12]). The path of maximum score over all combinations ($\mathbf{X}$,$\mathbf{C}$) gives the segmentation-recognition result.

The summation nature of (1) guarantees that the optimal path can be found by dynamic programming (DP) search. The search proceeds in frame-synchronous fashion: at each primitive segment s_t, examine all the candidate patterns $\mathbf{x_i}$ ending at s_t and the candidate class c_i assigned to $\mathbf{x_i}$. Denote the preceding candidate pattern of $\mathbf{x_i}$ as $\mathbf{x_{i-1}}$ ending at segment s_{t-k} and assigned classes c_{i-1}. For each pair (s_t, c_i), an optimal partial path with maximum partial score over (k, c_{i-1}) is retained (the c_{i-2} for tri-gram calculation has been known at preceding stage ending at (s_{t-k}, c_{i-1})).

III. Confidence Transformation

For probabilistic fusion of classifier outputs, the transformed confidence measures are desired to approximate the class posterior probability $P(\omega_j|\mathbf{x})$ (ω_j refers to the j-th class) [3], which can be obtained in several ways.

A. Confidence Types

The a posteriori probability can be directly obtained by the Bayes formula given the a priori probability and the conditional probability density of each class. Since the probability density functions are not trivial to estimate, there are many ways to approximate the a posteriori probability from classifier outputs. Under the assumption of Gaussian distribution with equal identity covariance matrix, the a posteriori probability is proportional to the exponential [6]:

$$P(\omega_j|\mathbf{x}) \propto \exp\left[\frac{-d_j(\mathbf{x})}{\theta}\right], \quad j = 1, 2, \ldots, M, \qquad (2)$$

where M is the total number of defined classes, $d_j(\mathbf{x})$ is the dissimilarity score for class ω_j output by the classifier, and the parameter θ is optimized on training samples rather than $2\sigma^2$ (the variance from maximum-likelihood) to overcome the deviation from distribution assumption. In this way, we can get the a posteriori probability by the normalization of (2), which results in the so-called softmax:

$$P^{sf}(\omega_j|\mathbf{x}) = \frac{\exp\left[-a \cdot d_j(\mathbf{x})\right]}{\sum_{i=1}^{M} \exp\left[-a \cdot d_i(\mathbf{x})\right]}, \quad j = 1, 2, \ldots, M, \quad (3)$$

where the parameter $a = \frac{1}{\theta}$ is used for more general purposes. In Chinese character recognition, to reduce the computation cost due to the large number of classes, we usually consider a reduced number of top rank classes while viewing the probabilities of the remaining classes as zero [6].

Viewing a multi-class problem as multiple one-versus-all binary problems, the sigmoid function is often taken for binary posterior probability, as commonly used in logistic regression and neural networks [13]:

$$P^{sg}(\omega_j|\mathbf{x}) = \frac{\exp\left[-a\cdot d_j(\mathbf{x})+b\right]}{1+\exp\left[-a\cdot d_j(\mathbf{x})+b\right]},\ j=1,2,\ldots,M. \quad (4)$$

We can also get the multi-class probabilities by combining the sigmoid confidence values according to the D-S theory of evidence [3], [14]. First, we give $2M$ focal elements (singletons and negations) $\omega_1, \overline{\omega_1}, \ldots, \omega_M, \overline{\omega_M}$ with basic probability assignments (BPAs) $m_j(\omega_j) = P^{sg}(\omega_j|\mathbf{x})$, $m_j(\overline{\omega_j}) = 1 - P^{sg}(\omega_j|\mathbf{x})$, then the combined evidence of ω_j is

$$P^{ds1}(\omega_j|\mathbf{x}) = A\cdot m_j(\omega_j)\prod_{i=1,i\neq j}^{M} m_i(\overline{\omega}_i), \quad (5)$$

where

$$A^{-1} = \sum_{j=1}^{M} m_j(\omega_j)\prod_{i=1,i\neq j}^{M} m_i(\overline{\omega}_i) + \prod_{i=1}^{M} m_i(\overline{\omega}_i).$$

Substituting the BPAs from sigmoid confidence (4) in (5) gives

$$P^{ds1}(\omega_j|\mathbf{x}) = \frac{\exp\left[-a\cdot d_j(\mathbf{x})+b\right]}{1+\sum_{i=1}^{M}\exp\left[-a\cdot d_i(\mathbf{x})+b\right]}, \quad j=1,2,\ldots,M. \quad (6)$$

We call this probability as D-S evidence-type confidence.

B. Confidence for Outlier Class

In character string recognition as well as many other pattern recognition problems, there maybe samples out of the M defined classes, which can be viewed as belonging to an "outlier class". The one-versus-all binary probability (sigmoid confidence) directly takes the outlier class as in the negative class of each singlet class. By combining binary probabilities, the D-S evidence confidence also covers the outlier class because the summation of M posterior probabilities in (6) is guaranteed to be smaller than or equal to one. The softmax form of (3) does not consider the outlier class, however. To modify, we assume the outlier class has a constant dissimilarity score $d_o(\mathbf{x}) = \frac{b}{a}$, which can be viewed as a threshold for outlier rejection. Under the $M+1$ classes framework, the softmax probabilities are modified to

$$\begin{aligned} P^{ds2}(\omega_j|\mathbf{x}) &= \frac{\exp\left[-a\cdot d_j(\mathbf{x})\right]}{\exp(-b)+\sum_{i=1}^{M}\exp\left[-a\cdot d_i(\mathbf{x})\right]} \\ &= \frac{\exp\left[-a\cdot d_j(\mathbf{x})+b\right]}{1+\sum_{i=1}^{M}\exp\left[-a\cdot d_i(\mathbf{x})+b\right]}, \\ & \qquad j=1,2,\ldots,M, \end{aligned} \quad (7)$$

which is equivalent to the form of D-S evidence (6), but is derived from a different viewpoint. Also, the parameters (a,b) for (6) (combined from sigmoid probabilities, and we call DS1) and the extended softmax (7) (called DS2) are estimated in different ways. For DS1, the parameters of sigmoid confidence are estimated from one-versus-all perspective; while for DS2, the parameters are estimated from multi-class perspective.

In the D-S evidence framework, the outlier probability is

$$P^{ds}(\omega_{outlier}|\mathbf{x}) = \frac{1}{1+\sum_{i=1}^{M}\exp\left[-a\cdot d_i(\mathbf{x})+b\right]}, \quad (8)$$

which is the complement probability to the M defined classes.

C. Confidence Parameters Estimation

We optimize the confidence parameters by minimizing the cross entropy (CE) loss function, which is commonly used in logistic regression and neural network training [13]. Depending on the binary or multi-class nature, we use the binary CE loss for the sigmoid confidence and the multi-class CE for the softmax confidence. On a validation dataset (preferably different from the dateset for training classifiers) of N samples, the binary CE is

$$\min CE = -\sum_{n=1}^{N}\sum_{j=1}^{M}\left[t_j^n \log P_j + (1-t_j^n)\log(1-P_j)\right]. \quad (9)$$

Minimizing the multi-class CE is equivalent to maximizing the conditional log-likelihood:

$$\min CE = -\sum_{n=1}^{N}\sum_{j=1}^{M}\left[t_j^n \log P_j\right]. \quad (10)$$

In the above, $P_j = P(\omega_j|\mathbf{x})$, $t_j^n = \delta(c^n, j) \in \{0,1\}$, where $c^n(=1,2,\ldots,M)$ is the class label of the n-th sample $\mathbf{x}^n$. In both cases, a term of weight decay is added to alleviate the overfitting, and we minimize the empirical loss by stochastic gradient descent to estimate the confidence parameters (a,b).

Outlier samples can be used for parameter estimation in either binary (sigmoid) or multi-class (softmax) case. The effect of outlier samples is not crucial for the binary CE because for each singlet class, the samples of the other classes play similar effect as outlier samples. For the multi-class CE, however, outlier samples are crucial to guarantee the outlier probability (8). In our experiments, we will evaluate the effects using variable number of outlier samples.

IV. Experimental Results

We evaluated the handwritten text recognition performance with confidence transformation on a database CASIA-HWDB produced by 1,020 writers, collected by the Institute of Automation of Chinese Academy of Sciences (CASIA). The database includes both isolated characters and handwritten texts, and is divided into a training set of 816 writers and a test set of 204 writers. The training set contains 4,198,494 isolated character images of 7,356 classes (7,185 Chinese characters, 10 digits, 52 English letters and 109 frequently used symbols) from isolated characters and unconstrained texts. We tested on the unconstrained texts of 204 writers, including 1,015 documents, which were segmented into 10,449 text lines and there are 268,629 characters (26,583 symbols, 6,879 digits, 747 letters, 233,329 Chinese characters and 1,091 characters out of the 7,356 classes). The text lines were recognized one by one with the cross-line linguistic dependency incorporated.

The character classifier used in text recognition extracts character features from gray-scale images using the normalization-cooperated gradient feature (NCGF) method [15]. The obtained 512D feature vector is reduced to 160D by Fisher linear discriminant analysis (FLDA) and then classified using an MQDF classifier or a nearest prototype classifier (NPC). The NPC was trained using the algorithm of log-likelihood of margin (LOGM) [16]. We used 4/5 samples of the training character set for training classifiers, and the remaining 1/5 samples for confidence parameter estimation. The non-character samples for confidence parameter estimation were extracted from the unconstrained text lines in the training set, with some examples shown in Fig. 3.

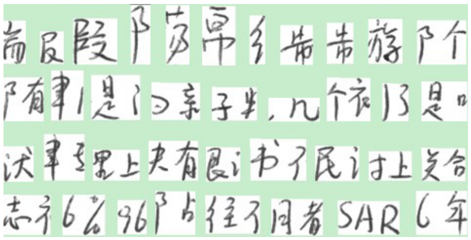


Figure 3: Examples of non-character samples.

In confidence transformation, we took only the 200 top-rank classes in calculating the softmax and D-S evidence for acceleration. After confidence transformation, only the 20 top-rank classes were used in the candidate character lattice as done in our previous work [9]. This is to reduce the lattice and speed up path search.

We evaluate the recognition performance using three character-level metrics: recall (rcl), precision (prs) and F-rate (frt), which are defined as

$$
\begin{aligned}
rcl &= \frac{\#correct}{\#truth} \times 100\% \\
prs &= \frac{\#correct}{\#result} \times 100\% \\
frt &= 2 \times \frac{rcl \cdot prs}{rcl + prs} \times 100\% \qquad (11)
\end{aligned}
$$

where $\#correct$ denotes the number of correctly recognized characters calculated by dynamic programming (DP) alignment of the output text string and the ground-truth string, $\#truth$ denotes the number of characters in ground-truth strings, and $\#result$ denotes the number of segmented characters after text recognition. We also give the correct (recall) rates for different types of characters: symbols (sb), letters (lt), digits (dg), and Chinese characters (ch).

For combining the classifier outputs with the language model in handwritten text recognition, we tested five options of confidence transformation: using the output dissimilarity measure directly as log-likelihood (without confidence transformation), softmax without outlier class, D-S evidence (combination of sigmoid confidence), and extended softmax with outlier class, which are abbreviated to “w/o”, “sf”, “sg”, “ds1”, and “ds2”, respectively.

By confidence parameter estimation without outlier samples, the recognition performance using MQDF classifier and nearest prototype classifier (NPC) are shown in Table I and Table II, respectively. The results show compared to recognition without confidence transformation, sigmoid confidence and D-S evidence confidence (“ds1”) improve the recognition performance. For MQDF, the recall rate is improved from 85.64% to 88.01% and 89.40%, respectively. And for NPC, the recall rate is improved from 84.27% to 84.96% and 85.33%. The benefit of confidence transformation is attributed to the fact that the converted class probabilities and the probabilistic language model are more compatible to be combined. “ds1” performs even better than “sg” because it gives multi-class probabilities. However, “sf” and “ds2” give inferior performance, because “sf” does not consider outlier probability while in the parameter estimation of “ds2”, no outlier samples were used.

Table I: Recognition rates (%) of MQDF without outlier samples in confidence parameter estimation.

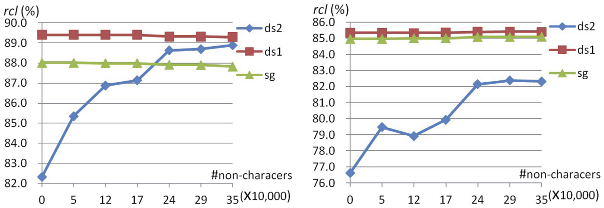
	rcl	pcs	frt	sb	dg	lt	ch
w/o	85.64	84.55	85.09	79.27	83.15	72.82	86.88
sg	88.01	86.02	87.01	77.85	82.99	75.64	89.77
sf	78.01	86.34	81.96	58.30	56.05	47.52	81.36
ds1	89.40	88.77	89.08	82.08	83.85	76.71	90.86
ds2	82.31	88.00	85.06	62.96	65.78	57.83	85.47

Table II: Recognition rates (%) of NPC without outlier samples in confidence parameter estimation.

	rcl	pcs	frt	sb	dg	lt	ch
w/o	84.27	84.68	84.48	79.44	77.95	61.85	85.48
sg	84.96	85.49	85.22	76.40	78.91	65.46	86.57
sf	75.10	82.65	78.69	65.82	55.65	38.69	77.20
ds1	85.33	86.87	86.09	78.99	76.35	64.52	86.79
ds2	76.62	83.39	79.86	67.09	59.27	41.23	78.68

In the confidence parameter estimation of “sg”, “ds1” and “ds2”, non-character samples can be used to improve the outlier resistance. The recognition performance (recall rate) of MQDF and NPC with variable number of non-character samples in confidence parameter estimation shown in Fig. 4, and Table III and Table IV show the performance when using 350,000 non-character samples. The results show that outlier samples are much influential to the performance of “ds2”. When using a large number of outlier samples, “ds2” can even yield higher recall rate than “sg” and close to “ds1” (this is the case for MQDF classifier). On the other hand, the performance of “sg” and “ds1” is almost not influenced by outlier samples in parameter estimation, this is because they are inherently resistant to outliers.

Finally, the performance (recall rate) of text recognition



(a) MQDF (b) NPC

Figure 4: Recall rate with variable number of outlier samples in confidence parameter estimation.

Table III: Recognition rates (%) of MQDF with outlier samples in confidence parameter estimation.

	rcl	pcs	frt	sb	dg	lt	ch
sg	87.82	86.01	86.90	78.38	83.25	74.97	89.48
ds1	89.28	88.66	88.97	82.15	83.98	76.57	90.71
ds2	88.88	90.27	89.57	76.38	79.82	73.49	91.03

Table IV: Recognition rates (%) of NPC with outlier samples in confidence parameter estimation.

	rcl	pcs	frt	sb	dg	lt	ch
sg	85.09	85.58	85.34	76.90	79.01	65.60	86.67
ds1	85.41	86.90	86.15	79.25	76.64	64.52	86.84
ds2	82.31	86.12	84.17	73.37	69.46	54.75	84.18

without using language model in path evaluation is shown in Table V, where "mqdf-o" and "npc-o" denote that outlier samples were used in confidence parameter estimation for MQDF and NPC, respectively. It is again shown that confidence transformation mostly improves the recognition performance.

V. Conclusion

We evaluated the effects of classifier confidence transformation in handwritten Chinese text recognition system using different confidence types and parameter estimation methods. The sigmoid and D-S evidence confidence, due to their coverage of outlier probability, give improved text recognition performance even when estimating without outlier samples. The softmax confidence is inferior because it does not consider the outlier class, while the performance of the extended softmax with outlier class largely relies on outlier samples in confidence parameter estimation. Based on confidence transformation, our future work will integrate more contextual information to further improve the handwritten text recognition performance.

Table V: Recall rate (%) of recognition without language model.

	w/o	sg	sf	ds1	ds2
mqdf	63.02	65.75	66.31	66.90	69.95
mqdf-o	—	65.22	—	66.09	72.04
npc	71.12	75.17	69.39	75.33	70.72
npc-o	—	75.16	—	75.30	74.27

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC) under grants no.60825301 and no.60933010.

References

[1] M. Cheriet, N. Kharma, C.-L. Liu, C.Y. Suen, *Character Recognition Systems: A Guide for Students and Practioners*, John Wiley & Sons, Inc. 2007.

[2] X. Lin, X. Ding, M. Chen, R. Zhang, Y. Wu, Adaptive confidence transform based on classifier combination for Chinese character recognition. *Pattern Recogn Lett.* 19 (10) 975-988, 1998.

[3] C.-L. Liu, Classifier combination based on confidence transformation, *Pattern Recognition* 38 (1) 11-28, 2005.

[4] Y.X. Li, C.L. Tan, X. Ding, A hybrid postprocessing system for offline handwritten Chinese script recognition, *Pattern Analysis and Applications* 8, 272-286, 2005.

[5] Y. Jiang, X. Ding, Q. Fu, Z. Ren, Context driven Chinese string segmentation and recognition, *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR Int. Workshops*, LNCS, Vol.4109, 127-135, 2006.

[6] C.-L. Liu, M. Nakagawa, Precise candidate selection for large Character set recognition by confidence evaluation, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (6) 636-642, 2000.

[7] C.-L. Liu, H. Sako, H. Fujisawa, Effects of classifier structures and training regimes on integrated segmentation and recognition of handwritten numeral strings, *IEEE Trans. Pattern Anal. Mach. Intell.* 26 (11) 1395-1407, 2004.

[8] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to Chinese character recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 9 (1) 149-153, 1987.

[9] Q.-F. Wang, F. Yin, C.-L. Liu, Integrating language model in handwritten Chinese text recognition, *Proc. 10th ICDAR*, Barcelona, Spain, 1036-1040, 2009.

[10] C.-L. Liu, M. Koga, H. Fujisawa, Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading, *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(11): 1425-1437, 2002.

[11] D. Povey, P.C. Woodland, Minimum phone error and I-smoothing for improved discriminative training. *Proc. ICASSP*, Orlando, FL, 2002.

[12] M.-Y. Chen, A. Kundu, S.N. Srihari, Variable duration Hidden Markov model and morphological segmentation for handwritten word recognition, *IEEE Trans. Image Processing*, 4 (12) 1675-1688, 1995.

[13] C.-M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.

[14] J.A. Barnett, Computational methods for a mathematical theory of evidence, *Proc. 7th IJCAI*, Vancouver, Canada, 868-875, 1981.

[15] C.-L. Liu, Normalization-cooperated gradient feature extraction for handwritten character recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 29 (8) 1465-1469, 2007.

[16] X.-B. Jin, C.-L. Liu, X. Hou, Regularized margin-based conditional log-likelihood loss for prototype learning, *Pattern Recognition* 43 (7) 2428-2438, 2010.