Minimum Error Discriminative Training for Radical-based Online Chinese Handwriting Recognition

Yaodong Zhang
Shanghai Jiao Tong University
Shanghai, China
zhydong@sjtu.edu.cn

Peng Liu
Microsoft Research Asia
Beijing, China
pengliu@microsoft.com

Frank K. Soong
Microsoft Research Asia
Beijing, China
frankkps@microsoft.com

Abstract

Free style Chinese handwriting recognition continues to pose a challenge to researchers due to the variety of Chinese writing styles. To recognize handwritten characters in an online mode, Hidden Markov Model (HMM) has been naturally adopted to model the pen trajectory of a character and a decent recognition performance is achieved. In this study, we start from a maximum likelihood trained HMM model and focus on minimizing the errors on the radical (sub-character) level to optimize the recognition performance. A novel Minimum Radical Error discriminative training criterion is proposed, and compared with the discrimination on the character level, our new approach further reduces the character error rate by 15.55% relatively (totally 29.00% reduction from the maximum likelihood baseline model) on a Chinese database,

1. Introduction

Online Chinese handwriting recognition continues to be a challenging pattern recognition problem due to the highly variable writing styles of a Chinese character: printed, fluent and cursive. In past decades, researchers have developed different strategies to tackle this problem [8] mainly in two directions. From the data perspective, characters in Chinese and other Eastern-Asian languages can be considered as a stand-alone holistic figure [9], or can be decomposed into basic, sub-character units of radicals [12]. Further decomposition is also possible, i.e., each radical can be decomposed into predefined strokes or sub-strokes [2]. Holistic approach usually yields higher recognition accuracy [4] since it captures the entire character shape information, but it is cumbersome to collect character specific training data for more than 20,000 Chinese characters. Also, the model size becomes very large. On the other hand, stroke-based approach defines basic units based upon the analysis of printed fonts, which lead to more compact models. However, for more free-styled cursive writing, it is hard to align the conceptual strokes with the pen trajectory, especially in a writer independent mode. In Chinese and other Eastern-Asian languages, radicals are relatively stable and they form a compact set. With adequate number of basic radicals, we can effectively characterize the entire character set, which provides a promising alternative approach to Chinese online handwriting modeling.

From the modeling perspective, Hidden Markov Model (HMM) is intuitively suitable for modeling the temporal pen trajectory [4] [6], and is expected to achieve a decent recognition performance. Thus, we focus on radical-based trajectory modeling by HMMs in this study. Although using the online trajectory modeling approach, it is still a challenging problem of how to normalize different writing styles and different writing orders among different writers. A data-driven approach is to build models aiming at achieving better classification than better description of the training data, or to train a discriminative HMM. In past years, discriminative criteria, such as Minimum Classification Error (MCE) [5] and Maximum Mutual Information (MMI) [10], have been shown to outperform the maximum likelihood criterion in speech recognition. It is noted that in the framework of minimum error training [11], by defining errors in a higher resolution, we can further improve model discrimination, demonstrated in Minimum Phone Error (MPE) [11] and Minimum Divergence (MD) [3]. In this research, we apply the concept of minimum error training to handwriting recognition. We investigate how to define recognition error at the character and radical level, which leads to MMI training and a novel Minimum Radical Error (MRE) training.

The rest of the paper is organized as follows: in Section 2, we briefly introduce the radical-based Chinese character. In Section 3, we propose the Minimum Radical Error criterion to discriminatively train the radicals models. Next we conduct experiments and analyze errors to demonstrate the effectiveness of discriminative model training. Finally, in Section 5 we conclude the paper and discuss future works.

2. Radical-based Chinese Character

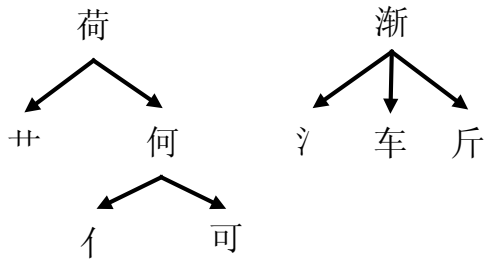


Figure 1. Chinese Character Decomposition

Chinese character is formed in a complex but hierarchical structure as illustrated in Figure 1. Radicals or sub-character roots are commonly used as the basic semantic or phonetic units to construct Chinese characters. Each Chinese character consists of one or several radicals which are similar to speech that each word can be decomposed into one or several sub-word units, saying phones. With a reasonable set of small number of radicals, we are able to represent all Chinese characters. Thus, the problem of modeling all Chinese characters becomes easier since we only need to model a relatively small number of radicals. Another advantage of this approach is that by using a character-to-radical dictionary, we can synthesis, hence recognize characters which are not included in the training data.

3. Minimum Radical Error Training

Given S training samples, the objective function of minimum error training can be represented as:

$$\mathcal{F} = \frac{1}{S}\sum_{i=1}^{S} f\left(\log \frac{\sum_C P_\lambda^\kappa(\mathcal{O}_i|C)P(C)\mathcal{A}(C, C_i^{\text{ref}})}{\sum_C P_\lambda^\kappa(\mathcal{O}_i|C)P(C)}\right)$$

where i is the index of a training sample, C denotes a hypothesis class (Chinese character), λ denotes the set of parameters of the HMM model, $\mathcal{O}_i$ and C_i^{ref} represents the observation and reference class of the i'th training sample, respectively. f is a smooth function and κ is a scaling factor to normalize the probability. A key difference between the various criteria in this framework is the accuracy term $\mathcal{A}$, which measures the similarity between hypotheses and the reference class. In speech recognition, by increasing the error resolution from sentence level to phone level, a new criterion of MPE is introduced and outperforms sentences level criteria such as MMI [11].

In the radical based online handwriting modeling, we can intuitively adopt the concept of minimum error training. When defining error at the character level, the criterion

Table 1. Comparison of MMI and MRE training

Criteria	$f(x)$	$\mathcal{A}$
MMI	x	$\delta(C, C_i^{\text{ref}})$
MRE	$\exp(x)$	$\lvert C_i^{\text{ref}}\rvert - \text{LEV}(C, C_i^{\text{ref}})$

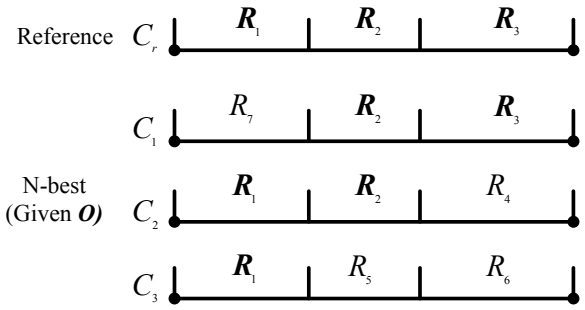


Figure 2. Example

of MMI can be directly used. Inspired by MPE, we can also use the radical level accuracy between a hypothesis character and the reference to increase the error resolution. Correspondingly, the criterion is termed as Minimum Radical Error (MRE). A brief comparison between MMI and MRE is listed in Table 1, where $\delta(C, C_i^{\text{ref}})$ is a binary function that is equal to 1 if C and C_i^{ref} are identical, $\text{LEV}(C, C_i^{\text{ref}})$ represents the Levenshtein distance [7] between the radical sequence of C and that of C_i^{ref}, and $|C_i^{\text{ref}}|$ represents the number of radicals of which $|C_i^{\text{ref}}|$ is composed.

Both MMI and MRE are to maximize posterior probability of the reference, which can be directly related to the recognition error in Bayesian sense. By counting radical errors in the accuracy function, we maximize the sum of posterior probability of all the radicals in the reference, which can be illustrated by the example showed in Figure 2. Suppose that we are recognizing a character labeled with C^{ref} and come up with three hypotheses results C_1, C_2 and C_3 with the corresponding probabilities p_1, p_2 and p_3, respectively. Here the probability is defined by $P_\lambda^\kappa(\mathcal{O}_i|C)P(C)$. Assuming all the hypotheses have the same number of radicals as in the reference, and the radical boundaries are aligned, the objective function of MRE can be written as:

$$\mathcal{F}_{\text{MRE}} = \frac{2p_1 + p_2 + 2p_3}{p_1 + p_2 + p_3}$$

On the other hand, the posterior probabilities of a radical R starting from time s and ending at e is defined as:

$$P([R, s, e]|\mathcal{O}) = \frac{\sum_{C:[R,s,e]\in C} P_\lambda^\kappa(\mathcal{O}|C)P(C)}{\sum_C P_\lambda^\kappa(\mathcal{O}|C)P(C)}$$

Accordingly, the sum of the posterior probabilities of all reference radicals is $S_{\text{PP}} = \sum_{[R,s,e]\in C^{\text{ref}}} P([R,s,e]|\mathcal{O})$. In the special case depicted in Figure 2, we obtain:

$$S_{\text{PP}} = \frac{2p_1 + p_2 + 2p_3}{p_1 + p_2 + p_3} = \mathcal{F}_{\text{MRE}}$$

which means the objective function of MRE is to maximize the radical posterior. In other words, the error measure is refined to the radical level.

Practically, we can conduct minimum error training in two steps: first, we concentrate on character level errors and applying MMI to train the model discriminatively; second, based upon the model trained by MMI, we apply MRE to refine errors at a higher resolution.

4. Experiments

4.1. Data Set

We use a large database of online Chinese handwritten characters collected from the Tablet-PC platform. The data set for training contains a total of 563,250 samples of 3,755 (150 samples per character) most frequently used Chinese characters in printed, fluent and cursive writing styles. The samples from these three writing styles are non-uniformly distributed over the whole data set. 1,142 radicals are selected to form the radical set to represent all these characters. There are 8 special imaginary radicals which represent the direction of connection between two radicals. In the testing stage, we use a database of 56,325 samples over these 3,755 characters (15 samples per character). The samples from the three writing styles are also non-uniformly distributed over the test set.

4.2. Baseline System

To build a baseline HMM model, we use a sequence of four dimensional feature vector to represent the pen trajectories of a radical. The four dimensional feature is composed of slope features $(\cos\theta, \sin\theta)$ and curvature related features $(\cos\Delta\theta, \sin\Delta\theta)$, as illustrated in Figure 3.

We use HTK (v3.2.1) [1] to train the maximum likelihood (ML) HMM model as our baseline system. Left-to-right HMMs are adopted to model each radical, and the output distributions are Gaussian. There are a total of 3,374 untied states over these 1,142 models. Since the training database is mixed with three kinds of writing styles and no special method is adopted to deal with various writing orders and styles, ML based trajectory modeling is not so successful in building effective models. Accordingly, the character error rate (CER) of this baseline model is 18.79%.

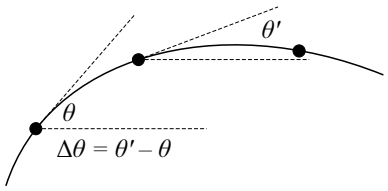


Figure 3. Features

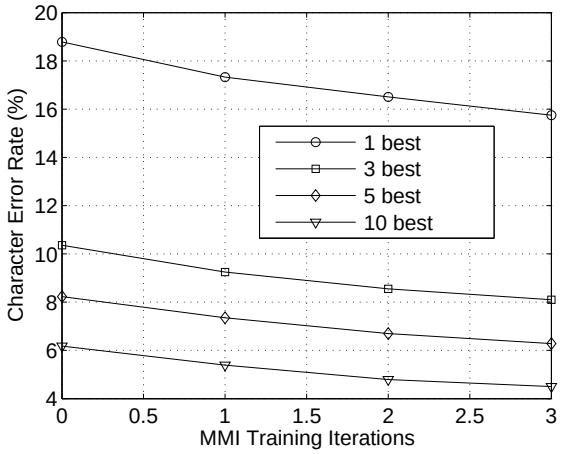


Figure 4. MMI Training Results

4.3. MMI training

We apply MMI training on the baseline model. The scaling parameter is set to $\kappa = 1/15$ and we perform isolated character recognition without any context information. Figure 4 shows the result. The training process converges after three iterations (the following curve is omitted). Finally, we reduce the CER to 15.75%, with a the relative reduction of 16.18%. Compared with ML training, MMI training process takes into account the samples of other competitive categories other than just describes them individually, which will give better classification performance on some samples that are easy to be confused on the ML trained model. That is possibly the main reason why MMI training leads to better recognition performance. We report the detailed results in Table 2.

4.4. MRE training

MMI criterion investigates the error resolution on the character level, while, to further localizing the error resolution to a lower level – radical level, we apply our MRE training by using the resultant model got from the MMI training as the seed model. We still set the scaling factor $\kappa = 1/15$. The training process converges after six itera-

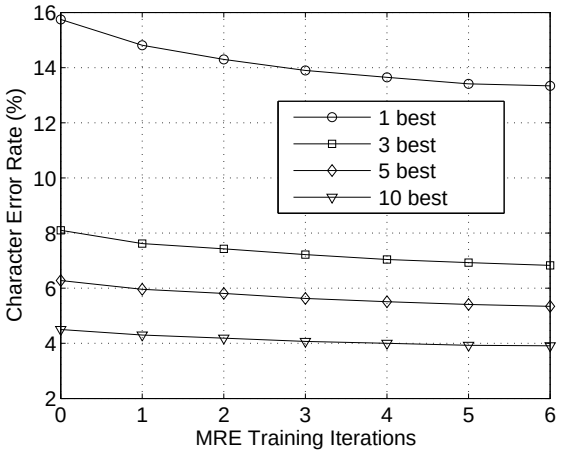


Figure 5. MRE Training Results

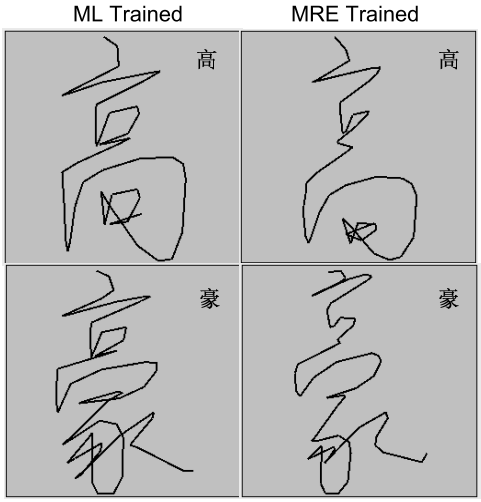


Figure 6. Radical Comparison

Table 2. Result(Character Error Rate)

	1-best	3-best	5-best	10-best
Baseline	18.79%	10.36%	8.23%	6.18%
3rd MMI	15.75%	8.10%	6.28%	4.50%
Impr.(rel)	16.18%	21.81%	23.69%	27.45%
6th MRE	13.34%	6.83%	5.34%	3.91%
Impr.(rel)	29.00%	34.07%	35.12%	36.73%

tions. The result illustrated in Figure 5 shows that we further reduce the CER from 15.75% to 13.34%, with a relative reduction of 15.30%. The reduction proves that MRE training criterion gives a more accurate measurement of the similarity between a hypothesis recognition result and the reference class than MMI training criterion, which better improves the discrimination among confusable radical models.

In sum, after two stages of discriminative training, a CER reduction of 29.00% is achieved , as listed in Table 2. When considering N-best results, the improvement is even larger, which shows that the models become more reasonable.

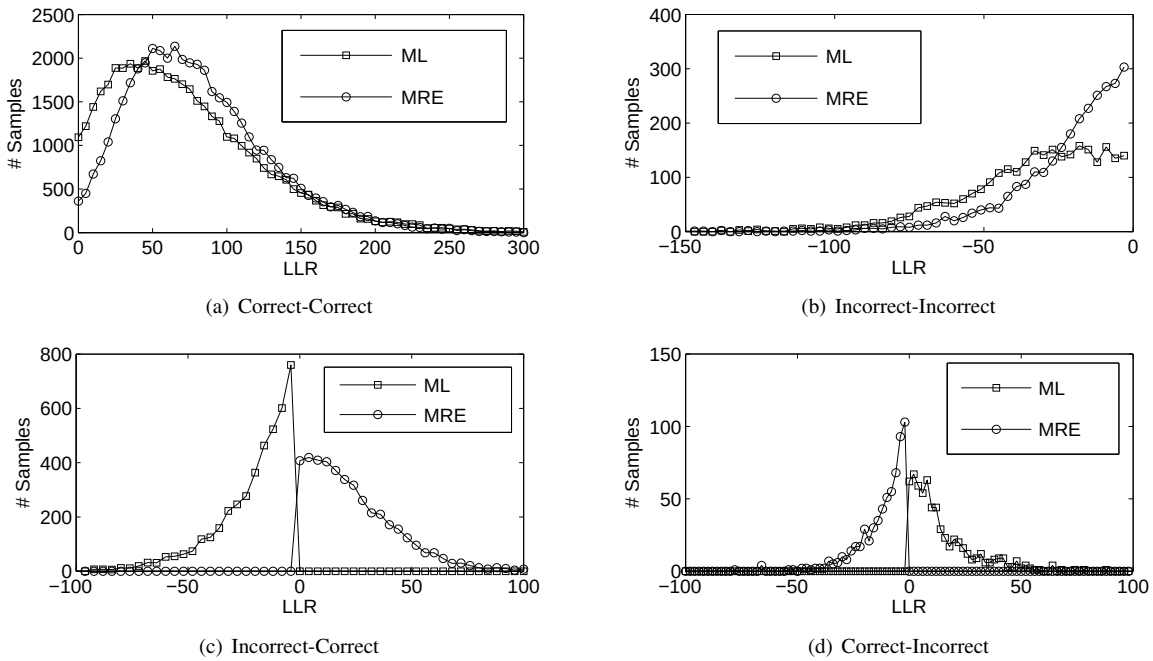
4.5. Result Analysis

To illustrate the difference between the ML trained model and the MRE refined model, we visualize them for comparison. An ideal sample according to a HMM is constructed by connecting stroke pieces corresponding to each state in the HMM: The stroke piece runs along the directions indicated by the model mean of $(\cos\theta, \sin\theta)$, and the lengths is determined by the expected duration of the state. Two comparison pairs are shown in Figure 6. In general, we observe that MRE refined model is more smooth and pays more attention to those distinguishing segments, e.g. the start, the end of a radical and transitions between stokes within a radical. As a result, the model becomes more robust and powerful in dealing with the variety of writing styles from different writers.

To further reveal how discriminative training improve the recognition performance, we analyze the recognition results and the corresponding discriminative models. We observe the Log-Likelihood Ratio (LLR) on the testing set. The measure is defined as:

$$LLR = \log \frac{P(\mathcal{O}|C^{\text{ref}})}{\max_{C \neq C^{\text{ref}}} P(\mathcal{O}|C)}$$

We compute the LLR of each test sample using both ML trained model and the MRE refined model. The test set is divided to four sub-sets and their histograms of LLR are shown in Figure 7. The "Correct-Correct" set is composed of those correct recognized samples both on ML model and MRE model, while the "Incorrect-Correct" set represents those mis-recognized samples on ML model but they are corrected after MRE refinement, and so on. A promising way of optimizing a model is to let the correct samples keep ahead and continuously increase the distance from its competitors, and to let the falsely recognized samples catch up the leader as early as possible. Accordingly, sub-figure (a) and (b) demonstrate us these trends. Sub-figure (c) and (d) illustrate why the recognition performance is improved after discriminative training. It is easy to find that much more samples are corrected (c) while only a few ones which get the correct recognition result on ML model are mis-recognized after MRE refinement (d).



(a) Correct-Correct

(b) Incorrect-Incorrect

(c) Incorrect-Correct

(d) Correct-Incorrect

Figure 7. Log-Likelihood Distribution

5. Conclusions

In this paper, we present a novel Minimum Radical Error discriminative training criterion and apply it to a radical-based online Chinese handwriting recognition system. The results show that by using the new MRE training criterion, the recognition performance improves by 29.0% compared with the maximum likelihood trained baseline system. By visualizing models and investigating the Log-Likelihood Ratio changes on each test sample, we provide a clear analysis of how MRE training improves the recognition accuracy. The effective HMMs to characterize pen trajectory. In future work, we plan to apply our MRE training on a larger character database and do more comparison with other discriminative training criteria. Moreover, we try to define errors in a more higher resolution, such as on sub-radical or stroke level, to further refine the models.

References

[1] Hidden markov model toolkit. *http://htk.eng.cam.ac.uk/*.

[2] J.-W. Chen and S.-Y. Lee. On-line handwriting recognition via a representation of spatial relationships between strokes. *IEEE Trans. Pattern Recognition and Artifical Intelligence*, 11(3):329–357, 1997.

[3] J. Du, P. Liu, F. Soong, J.-L. Zhou, and R.-H. Wang. Minimum divergence based discriminative training. In *Proceedings of ICSLP'06*, pages 2410–2413, 2006.

[4] B. Feng, X.-Q. Ding, and Y.-S. Wu. Chinese handwriting recognition using hidden markov models. In *Proceedings of ICPR'02*, pages 212–215, 2002.

[5] B. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Trans. Signal Processing*, 40(12):3043–3054, 1992.

[6] H.-J. Kim, K.-H. Kim, S.-K. Kin, and F.-T. Lee. On-line recognition of handwritten chinese characters based on hidden markov models. *Pattern Recognition*, 30(9):1489–1499, 1997.

[7] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[8] C.-L. Liu, S. Jaeger, and M. Nakagawa. Online recognition of chinese characters: The state-of-the-art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):198–213, 2004.

[9] H.-L. Liu and X.-Q. Ding. Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes. In *Proceedings of ICDAR'05*, pages 19–25, 2005.

[10] Y. Normandin, R. Lacouture, and R. Cardin. Mmie training for large vocabulary continuous speech recognition. In *Proceedings of ICASLP'94*, pages 1367–1371, 1994.

[11] D. Povey. Discrimative training for large vocabulary speech recognition. *Ph.D. Thesis*, University of Cambridge.

[12] D. Shi, R. I. Damper, and S. R. Gunn. Offline handwritten chinese character recognition by radical decomposition. *ACM Trans. Asian Language Information Processing*, 2(1):27–48, 2003.