

A New Radical-Based Approach to Online Handwritten Chinese Character Recognition

Long-Long Ma, Cheng-Lin Liu

*National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, P. R. China
{longma, liucl}@nlpr.ia.ac.cn*

Abstract

This paper proposes a new radical-based approach for online handwritten Chinese character recognition. The approach is novel in three respects: statistical classification of radicals, over-segmentation of characters into candidate radicals, and lexicon-driven recognition of characters. Currently, we have applied the approach to Chinese characters of left-right structure and are extending to other structures. Preliminary results on a sample set of 4,284 characters consisting of 1,118 radicals demonstrate the superiority of the proposed approach.

1. Introduction

Online handwritten Chinese character recognition (OLCCR) is gaining renewed interest owing to the increase of new pen input devices and pen applications. In the last decades, many approaches have been proposed and the recognition performance has advanced constantly [1]. Researches for further improving the recognition accuracy or reducing the classifier complexity are underway.

The hierarchical nature of Chinese characters has inspired radical-based recognition methods. Using radical models instead of holistic character models can largely reduce the number of models, and training the models needs a smaller number of character samples.

The methods of OLCCR can be roughly grouped into two categories: statistical and structural. Statistical methods generally represent the holistic character shape as a feature vector and use a statistical classifier for classification. The feature vector representation, e.g., the directional feature density or so-called direction histogram feature [2][3], enables stroke-order and stroke-number free recognition. Statistical methods have yielded high accuracies but suffer from high complexity because of the large number of character classes. Structural methods are based on stroke analysis

and radical analysis. To tolerate stroke-order variations, a character or radical is often modeled as a relational graph, with strokes or sub-strokes as primitives. Hidden Markov models (HMMs) are frequently used to model strokes and radicals [4][5]. Discriminative training has been applied to decrease the HMM-based radical recognition error [6]. Since the HMM is stroke-order dependent, multiple models per radical/character are needed for stroke-order variations.

All radical-based methods encounter the difficulty of radical segmentation. Some rule-based methods use the prior knowledge of character structure and radical position for radical detection [7], but this cannot guarantee reliable detection in cases of large shape variation. In a HMM-based method, characters are modeled by a network of radical and ligature HMMs [5]. In recognition, radicals can be segmented by dynamically matching the radical models with sub-sequences of strokes. This approach, however, does not tolerate stroke-order variations. A method avoids radical segmentation by radical location detection and location-based radical classification using neural networks on whole images [8], but suffers from large number of location-dependent radical models and inferior radical classification accuracy.

To combine the advantages of statistical methods and radical-based structural methods, we propose a new radical-based approach. We model radicals using stroke-order free features and statistical classification, over-segment the character pattern into candidate radicals, and search for optimal radical segmentation in lexicon-driven recognition. The lexicon stores the radicals in a tree structure, with each path corresponding to a character. This approach is inspired from lexicon-driven character string recognition [9]. We have implemented the approach on Chinese characters of left-right structure. In experiments on 4,284 characters consisting of 1,118 radicals, the proposed approach yielded higher accuracies than a holistic statistical method.

2. Radical Analysis

A radical is a sub-structure shared by multiple characters. From linguistic viewpoint, each radical also has a semantic meaning (Fig. 1). Some linguistic radicals, however, are hard to separate from characters by computer algorithms. In our paper, we remove such radicals and define some new ones that appear frequently.



Figure 1. Radical decomposition of Chinese characters.

We obtained radical models by self-learning in two stages. First, each character class has a sample correctly segmented into radicals by human interaction. The remaining samples are matched with the correct radical sequence by dynamic programming to segment into radicals. The radical features of the training samples of a class are averaged to obtain the class-specific radical templates. In the second stage, the radical templates of all classes are clustered to obtain shared radical models. We used agglomerative clustering to obtain a hierarchy of radical partitions and then selected a partition by human judge. It is hard to determine an appropriate cluster number automatically, yet the hierarchical clustering followed by human selection of partition works effectively.

Currently, we have implemented the approach on Chinese characters of left-right structure. By clustering the radical templates of 4,284 characters, we obtained 1,118 shared radical models (cluster centers).

3. Radical-Based Character Recognition

The block diagram of our radical-based recognition system is shown in Fig. 2. The input pattern is a sequence of strokes, which are grouped into primitive segments in pre-segmentation according to the pen-up distance and overlap/crossing between strokes. Candidate radicals are generated by combining consecutive primitive segments and are recognized by the radical classifier in lexicon-driven matching. The optimal segmentation of radicals as well as their class labels (the radical classes in turn decide the character class) is given by the optimal path of maximum matching score.

The character-radical dictionary is stored in a trie (tree) structure, where each character is a string of

radicals. The strings with a common prefix, such as {口-可, 口-土, 口-力-口}, have a common path of parent nodes for the prefix. This can save both the storage of dictionary and the computation because the common prefix is stored only once and matched only once in search. The trie of 4,284 left-right characters has 4,949 nodes consisting 1,118 distinct radicals. A portion of the trie is shown in Fig. 3.

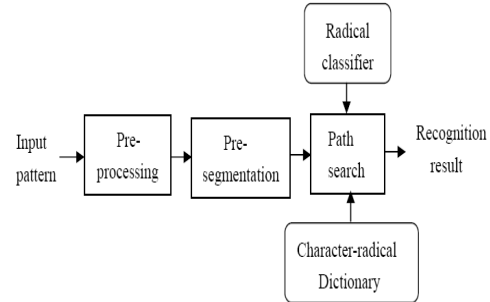


Figure 2. Block diagram of radical-based recognition.

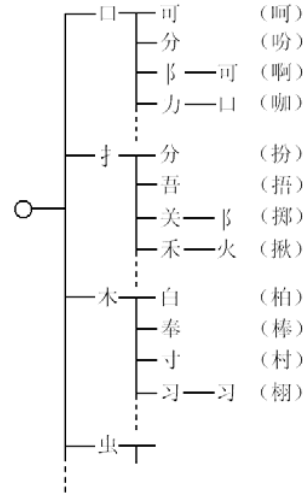


Figure 3. A portion of the trie structure.

3.1. Pre-segmentation

The strokes of a character pattern are iteratively grouped into primitive segments. Initially, each stroke is viewed as a component s_i . Two adjacent components are calculated a normalized overlap degree $novlp$ from their bounding boxes [9]. The components are grouped in following steps.

Step 1: Iteratively merge two temporally adjacent components s_i and s_j if $novlp(s_i, s_{i+1}) > T_1$ (empirically selected threshold), until the overlap condition is not

met. The merged components are ordered spatially according to the left boundary.

Step 2: Iteratively merge two spatially adjacent components s_i and s_j if $novlp(s_i, s_j) > T_2$ ($T_2 > T_1$) until the condition is not met.

Step 3: Merge small components to their left or right neighbor according to the horizontal distance.

3.2. Path Search

After pre-segmentation, the input character is represented as a sequence of primitive segments ordered by the left boundary. In lexicon-driven matching, candidate radical patterns are formed by combining consecutive (at most six) primitive segments and are matched with radical classes corresponding to the offspring nodes of a node in the trie. The search algorithm is radical-synchronous, namely, the increment of depth in the search space corresponds to the extension of string by one radical.

We use a beam search strategy to find matched radical strings. In the search space, a node stands for a pattern-radical pair, each pair is given a dissimilarity measure by the radical recognizer (classifier). For a radical string (a path from the root node), the costs of the constituent radicals are accumulated. During search, the accumulated cost is used to evaluate a partial string (at the same depth of search space, only the nodes with small partial costs are retained for extension), while for complete strings (characters), the average cost is used to decide whether to accept the result string or not.

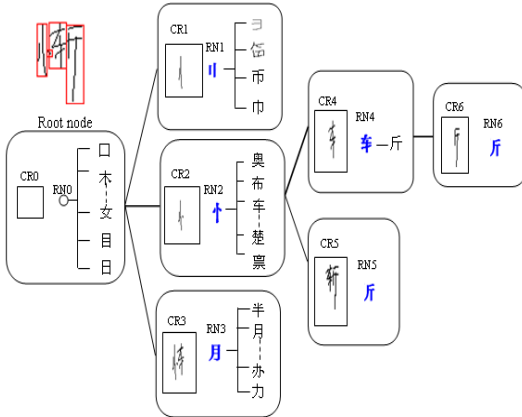


Fig. 4. Expansion of nodes in the search space of lexicon-driven matching.

Fig. 4 shows an illustrative example of node expansion in search space (part of nodes). CR denotes candidate radical patterns, and RN denotes radical

nodes in the trie. The path corresponding to the most plausible segmentation is denoted by a thick line.

3.3. Radical Classification

Each candidate radical pattern is assigned to a number of radical classes with their dissimilarity measures $f(\mathbf{x}, \omega_i)$, which are used in the path score.

$$f(\mathbf{x}, \omega_i) \propto -\log p(\mathbf{x} | \omega_i), \quad (1)$$

where $p(\mathbf{x} | \omega_i)$ is the class-conditional probability.

The radical pattern undergoes the same procedures of normalization, feature extraction and classification as done in a holistic character recognition method [10]. Specifically, a moment normalization method is used to normalize the coordinates of pen trajectory points, and direction histogram features are extracted directly from pen trajectory using a normalization-cooperated feature extraction (NCFE) method [3]. The resulting 512-dimensional feature vector is reduced to 160 by Fisher linear discriminant analysis (LDA), and a modified quadratic discriminant function (MQDF) classifier [11] is used to assign the radical pattern to 10 top-rank radical classes. The MQDF classifier uses 20 principal eigenvectors per class and the constant minor eigenvalue is selected by cross validation.

4. Experiments

We evaluated the performance of the proposed radical-based recognition approach on a database of online handwritten Chinese character database of 6,763 classes (the characters in GB2312-80), each class with 60 samples produced by 60 writers. We selected the samples of 4,284 classes of left-right structures for our experiments. We used 50 samples per class for training classifiers for radical and character recognition, and the remaining 10 samples per class for evaluating the performance.

We implemented three versions of radical-based recognition. Algorithm A does not use lexicon (i.e., lexicon-free) in radical segmentation and radical string recognition, and gives a correct recognition when the searched optimal radical string matches a legal character. Algorithm B is also lexicon-free in searching the segmentation path but multiple radical strings are obtained to match with the lexicon. Algorithm C is the proposed lexicon-driven radical string matching method. The test accuracies of three radical-based methods as well as the holistic MQDF method are listed in Table 1.

Table 1 shows that using lexicon in radical string matching (Algorithm C) is beneficial for the character

recognition accuracy. Compared to holistic character recognition, the lexicon-driven radical-based method yields higher accuracy. The radical-based method has another advantage that the radical dictionary consumes a smaller storage (about one quarter) than that of holistic characters.

The remaining errors of radical-based recognition are mainly attributed to two factors: failure of radical pre-segmentation (Fig. 5) and radical misclassification (Fig. 6). Pre-segmentation failure is due to stroke connection or high overlap between radicals. Such errors cannot be corrected in the subsequent path search. So, pre-segmentation is a very important step in our recognition system.

Table 1. Test accuracies of radical-based and holistic methods.

| Method | | #Class | Correct (%) |
|---------------|---|--------|--------------|
| Radical-based | A | 1,118 | 91.58 |
| | B | | 97.49 |
| | C | | 97.71 |
| Holistic | | 4,284 | 97.07 |

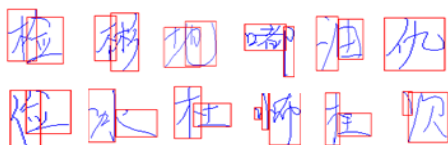


Figure 5. Examples of radical pre-segmentation failure.

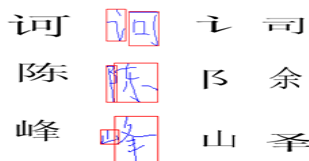


Figure 6. Examples of radical misclassification.

5. Conclusion

We presented a novel radical-based online handwritten Chinese recognition approach, which combines statistical classification with radical-based structural recognition. The difficulty of radical segmentation is overcome by radical over-segmentation and lexicon-driven radical string matching. In experiments on characters of left-right structures, the proposed method outperforms a state-of-

the-art holistic statistical recognition method. We are extending the method to characters of other structures.

Acknowledgements

This work was supported in part by Microsoft Research Asia under the program of Mobile Computing in Education and the National Natural Science Foundation of China (NSFC) under grant no.60775004.

Reference

- [1] C.-L. Liu, S. Jaeger, M. Nakagawa, Online handwritten Chinese character recognition: The state of the art, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(2): 198-213, 2004.
- [2] A. Kawamura, K. Yura, T. Hayama, Y. Hidai, T. Minamikawa, A. Tanaka, S. Masuda, On-line recognition of freely handwritten Japanese characters using directional feature densities, *Proc. 11th ICPR*, Hague, 1992, Vol.2, pp.183-186.
- [3] M. Hamanaka, K. Yamada, J. Tsukumo, On-line Japanese character recognition experiments by an off-line method based on normalized-cooperated feature extraction, *Proc. 2nd ICDAR*, Japan, 1993, pp.204-207.
- [4] H.J. Kim, K.H. Kim, S.K. Kim, F.T.-P. Lee, On-line recognition of handwritten Chinese characters based on hidden Markov models, *Pattern Recognition*, 30(9): 1489-1499, 1997.
- [5] M. Nakai, N. Akira, H. Shimodaira, S. Sagayama, Substroke Approach to HMM-Based On-Line Kanji Handwriting Recognition, *Prof. 6th ICDAR*, Seattle, WA, 2001, pp.491-495.
- [6] Y.D. Zhang, P. Liu, F.K. Soong, Minimum error discriminative training for radical-based online Chinese handwriting recognition, *Proc. 9th ICDAR*, Brazil, 2007, pp.53-57.
- [7] Y.J. Liu, L.Q. Zhang, J.W. Dai, A new approach to on-line handwriting Chinese character recognition, *Proc. 2nd ICDAR*, 1993, pp.192-195.
- [8] K. Chellapilla, P. Simard, A new radical based approach to offline handwritten East-Asian character recognition, *Proc. 10th IWFHR*, 2006, pp.261-266.
- [9] C.-L. Liu, M. Koga, H. Fujisawa, Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(11): 1425-1437, 2002.
- [10] C.-L. Liu, X.-D. Zhou, Online Japanese character recognition using trajectory-based normalization and direction feature extraction, *Proc. 10th IWFHR*, La Baule, France, 2006, pp.217-222.
- [11] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to Chinese character recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9(1): 149-153, 1987.