Special Radical Detection by Statistical Classification for On-line Handwritten Chinese Character Recognition

Ma Long-Long, Adrien Delaye, Cheng-Lin Liu

Special Radical Detection by Statistical Classification for On-line Handwritten Chinese Character Recognition

Long-Long Ma[1], Adrien Delaye[2], Cheng-Lin Liu[1]
[1]*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
[2]*IRISA - INSA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France*
adrien.delaye@irisa.fr

Abstract

The hierarchical nature of Chinese characters has inspired radical-based recognition, but radical segmentation from characters remains a challenge. We previously proposed a radical-based approach for on-line handwritten Chinese character recognition, which incorporates character structure knowledge into integrated radical segmentation and recognition, and performs well on characters of left-right and up-down structures (non-special structures). In this paper, we propose a statistical-classification-based method for detecting special radicals from special-structure characters. We design 19 binary classifiers for classifying candidate radicals (groups of strokes) hypothesized from the input character. Characters with special radicals detected are recognized using special-structure models, while those without special radicals are recognized using the models for non-special structures. We applied the recognition framework to 6,763 character classes, and achieved promising recognition performance in experiments.

1. Introduction

As the increasing of digitizing tablets, tablet PCs, digital pens, pen-based PDAs and mobile phones, on-line handwritten Chinese character recognition (OLHCCR) is gaining renewed interest. In the last decades, many approaches have been proposed and the recognition performance has advanced constantly [1]. To be implemented in hand-held devices with limited computation and storage capability, researchers are working towards high accuracy recognition methods with lower complexity.

The hierarchical nature of Chinese characters and Hangul characters has inspired radical-based recognition methods, which model a much smaller number of radicals instead of characters. Hierarchical character representation has also been used in Hangul character recognition [2-4], where components (graphemes) and the relationships between them are statistically represented. Such hierarchical representation-based methods have three benefits. First, the model complexity is reduced by modeling radical shapes instead of holistic character shapes. Second, by focusing on radicals with simpler structures than characters, the recognition accuracy can be improved. Third, the classification of a small number of radicals needs a small set of training samples.

All radical-based recognition methods encounter the difficulty of radical segmentation, however. Rule-based radical detection using the prior knowledge of character structure and radical position [5] is likely to fail in cases of large shape variation. Based on a network representation of radical and ligature HMMs [6], radicals can be segmented by dynamically matching the radical models with sub-sequences of strokes. This approach does not tolerate stroke-order variations, however. A method avoids radical segmentation by radical location detection and location-dependent radical classification using neural networks on whole character images [7], but without radical segmentation, it suffers from the large number of location-dependent radical models and the low radical classification accuracy.

To overcome the difficulty of radical segmentation, we previously proposed a radical-based recognition approach for characters of left-right (horizontal) and up-down (vertical) structures [8][9], which takes advantage of appearance-based classification of radicals. The approach is similar to character string recognition in the senses of candidate radical segmentation and tree representation of character compositions [10]. It integrates appearance-based radical recognition and geometric context into a

978-0-7695-4221-8/10 $26.00 © 2010 IEEE
DOI 10.1109/ICFHR.2010.83

principled framework using a character-radical dictionary to guide radical segmentation and recognition during path research.

The above approach works well on horizontal and vertical structures, but is unable to detect special radicals that are not linearly separable from the character patterns. Examples of characters with special radicals are shown in Fig 1. We can see that the special radicals are not linearly separable from the remaining part of the characters, but the remaining part is usually a horizontal-vertical structure or single element.

To recognize special-structure characters, we introduce a special radical detection module with radical hypotheses and verification using statistical binary classifiers. When a special radical is detected from the input character pattern, the remaining part is recognized using our previous approach. Otherwise, the input character is treated as horizontal-vertical structure and is directly recognized using our previous approach. We design 19 binary classifiers for detecting special radicals of 19 classes. The overall recognition system has been evaluated on characters of 6,763 classes.

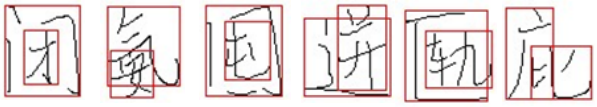


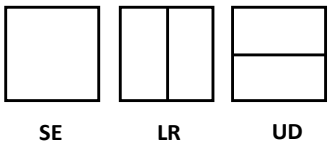
Figure 1. Characters with special radicals.

2. Chinese Character Structures

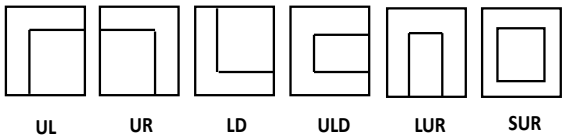
Chinese characters can be categorized into four rough structure types, single-element (SE), left-right (LR), up-down (UD) and special-structure. Special structures have been categorized in different ways: three types in [11], seven types in [12] and more detailed classes in [13]. We adopt the categorization of [12] but remove a structure type which involves very few character classes (we treat the characters of this type as single-element type). The nine types of structures considered in our system are shown in Fig. 2. For a character set of 6,763 Chinese characters (GB2312-80), the number of characters belonging to nine structure types, the percentage and some examples of the characters are shown in Table 1.

A radical is a sub-structure shared by multiple characters. Except single-element characters, each character consists of multiple radicals (like the radicals bounded by red boxes in Fig. 1). Most radicals have semantic meanings and often a radical is also a single character. It is beneficial to use radicals as the units of classification because the number of radical classes is much smaller than the number of characters and the radicals have simpler structures.

The characters of SE, LR and UD types can be well recognized by our previous system with nested (recursive) horizontal/vertical radical segmentation [8][9]. For the six special structure types UL, UR, LD, ULD, LUR and SUR, we design special radical detectors to segment the radicals and then recognize the remaining parts using the nested segmentation method. Particularly, we detect 19 special radicals for the six types of special structures.



(a) Non-special structures: single-element (SE), left-right (LR), up-down (UD).



(b) Special structures: up-left (UL), up-right (UR), left-down (LD), up-left-down (ULD), left-up-right (LUR), surrounding (SUR).

Figure 2. Nine structure types of Chinese characters.

Table 1. Statistics and examples of nine structure types.

Type	#Character	Percent (%)	Examples
SE	488	7.2	白本车册乘
LR	4284	63.3	败帐挣保知
UD	1489	22.0	曹罢恐息享
UL	240	3.6	疯雇层厄房
UR	28	0.4	氨甸氮旬氧
LD	141	2.1	迸趣毯魅魄
ULD	14	0.2	臣匿匹匣医
LUR	55	0.8	闭风冈闱闵
SUR	24	0.4	囤固国回困
Total	6,763	100	

3. Radical-Based Recognition Framework

The recognition framework with special radical detection is diagramed in Fig. 3. The input pattern is a sequence of strokes. At the special radical detection stage, radical-specific heuristics are used to group strokes for generating candidate radicals, which are classified using the special radical detector (binary classifier). If multiple candidates are accepted by a radical detector, the one of maximum similarity is retained. After special radical extraction, the remaining

part of the character is assumed to be SE/LR/UD structure and is recognized using the nested horizontal/vertical segmentation method. If multiple special radicals are detected by different detectors, each radical is respectively extracted from the input character and the remaining part is recognized by the corresponding SE/LR/UD module.

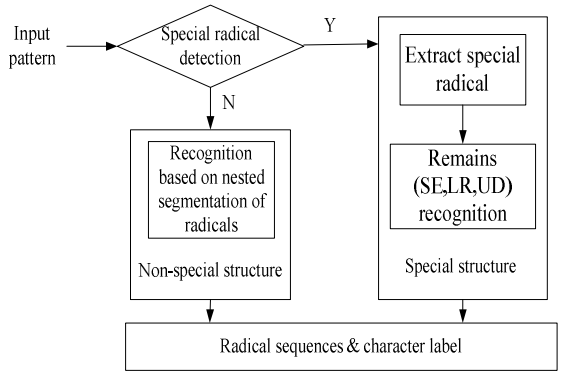


Figure 3. Framework of radical-based recognition.

When recognizing the remaining part after special radical extraction, if the matching distance of the remaining part (minimum over all the remaining part classes for a special radical) exceeds an empirical threshold, the detected special radical is rejected. Thus, the remaining part recognition performs radical verification. If exactly one special radical is accepted, the whole character recognition result is immediately obtained. If no special radical is detected from the input pattern or all the detected radicals are rejected, the input pattern is treated as a SE/LR/UD structure and recognized using the nested horizontal/vertical segmentation method. On the other hand, if multiple detected special radicals are accepted after remaining part recognition, each detected radical is given the whole character similarity by fusing the radical similarity and remaining part distance (transformed to sigmoidal confidence score), and the one of maximum similarity gives the character recognition result.

Note that there are 20 nested segmentation-recognition modules for non-special structures, one for SE/LR/UD whole characters and 19 for remaining parts of special radicals.

3.1. Non-Special Structure Recognition

For non-special structures (SE/LR/UD) of whole characters or remaining parts, we use our previously proposed nested horizontal/vertical radical segmentation and integrated segmentation-recognition method for recognition. The method is outlined as follows, and more details can be found in [8][9].

Character models (or remaining parts) are represented as nested horizontal/vertical strings of radicals (common sub-structures) in up to three hierarchies (horizontal-vertical-horizontal). Each radical is represented as a feature vector template or Gaussian density model (statistical radical model).

In recognition, we first split the input strokes at corner points of higher curvature to overcome the stroke connection between radicals. Candidate radicals are hypothesized from input pattern by grouping strokes according to the horizontal/vertical (depending on the hierarchy) overlapping degree. The candidate radicals are classified by statistical radical models to give matching scores. The combination of candidate radicals with highest score (minimum distance) gives the result of radical segmentation and character recognition. Radical hypothesis and classification are guided by a lexicon-driven tree search strategy, similar to lexicon-driven character string recognition [10].

4. Special Radical Detection

The special radicals cannot be segmented from the input character by horizontal/vertical grouping of strokes, so we design binary classifiers to detect them by classification on hypothesized candidate radicals. From six types of special-structure characters, we extract 19 special radicals, as shown in Table 2. Fig. 4 shows an example of special radical detection.

Table 2. Special radical models.

Type	# Special radical	Radical models
UL	6	疒 广 厂 疒 户 尸
LD	6	辶 走 毛 九 鬼 风
LUR	3	门 几 门
UR	2	气 勹
SUR	1	冂
ULD	1	匚

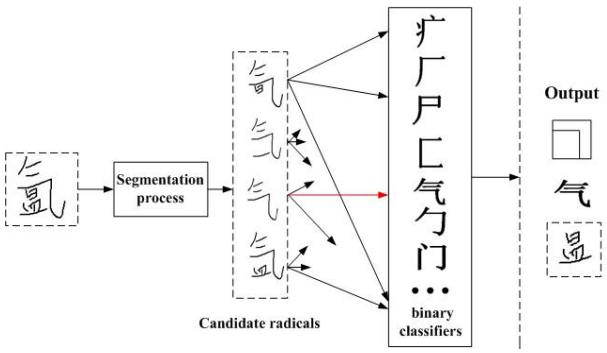


Figure 4. An example of special radical detection.

4.1. Special Radical Hypothesis

A radical is a sequence of temporally adjacent strokes, and special radicals are usually located at the beginning, end or both the beginning and end of stroke sequence. From this prior knowledge, we consider the character as a circular sequence of strokes as shown in Fig. 5.

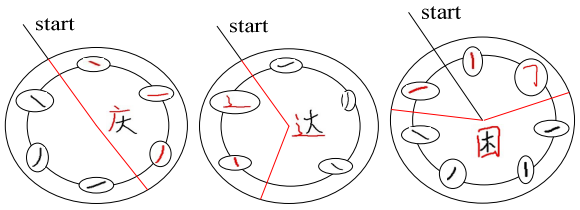


Figure 5. Circle sequence of strokes partitioned into candidate radicals. The red strokes form special radicals.

By inspecting the special-structure characters, we have more prior knowledge on the segmentation points and geometrics of special radicals. Some important geometric properties are below:

(1) The pen-up distance from special radical to the remaining part is often larger than those within radicals.

(2) The bounding box of special radical is large enough compared to that of the whole character.

(3) The longer side of the bounding box of the remaining part always overlaps with the special radical.

Further, special radicals are rarely connected with the remaining parts because of the large pen-up distance. Hence, we do not split strokes at corner points for special radical detection.

To hypothesize all candidate special radicals in an input character, we use two cutting points c_1 and c_2 (indices of strokes, $0 \leqslant c_1 < c_2 \leqslant n-1$, n is the number of strokes) to partition the strokes into two parts: $S_1 = s_0 \ldots s_{c_1-1} s_{c_2+1} \ldots s_{n-1}$, $S_2 = s_{c_1} \ldots s_{c_2}$. If $c_1=0$, S_1 consists of the end strokes and S_2 consists of the beginning strokes, and both S_1 and S_2 are candidate radicals; if $c_1>0$, only S_1 is candidate radical because it consists of both beginning and end strokes.

By scanning the strokes in two nested cycles $c_1=0$ to $n-1$ and $c_2=c_1+1$ to $n-1$, all the possible candidate radicals can be generated. The candidate radicals are filtered using three rules: (1) the eligibility of cutting points, (2) the eligibility of special radical class and (3) the separation between the radical and the remaining part. A candidate radical rejected by any one of the three rules is not passed to binary classification. A cutting point (c_1 or c_2) is valid if the pen-up distance preceeding the stroke is large enough. The eligibility of special radical class and the separation are evaluated as follows.

4.1.1. Eligibility of Special Radical Class

A special radical has some class-specific geometric rules: the number of strokes, the position in the sequence of strokes and the position of bounding box. Candidate radicals that do not pass these rules are rejected. Table 3 lists the rules for 19 special radicals. #Stroke denotes the number of strokes in regular writing. In our experiments, we filter out the candidate radicals with the number of strokes over #Stroke+1. Position code 0 denotes beginning strokes, 1 denotes end strokes, and 2 denotes both beginning and end strokes. Bounding box rules specify that the bounding box of radical coincides with one or more boundaries of the character bounding box.

Table 3. Geometric rules of special radicals.

Special radical	Rules		
	#Stroke	Position	Bounding box
0	5	0/2	left, top
1	3	0	left, top
2	2	0	left, top
3	3	0	left, top
4	4	0	left, top
5	4	0	left, top
6	3	0/1	left, bottom
7	7	0	left, bottom
8	4	0	left, bottom
9	2	0	left, bottom
10	9	0	left, bottom
11	4	0	left, bottom
12	3	0	left, right, top
13	2	0	left, right, top
14	2	0	left, right, top
15	4	0	top, right
16	2	0	top, right
17	2	0/2	top, left, bottom
18	3	0/2	left,right,top,bottom

4.1.2. Separation between Radical and Remaining

A special radical is assumed to be spatially apart from the remaining part. We evaluate the spatial separation using a function J defined as the ratio of within-radical to between-radical weights, similar to linear discriminant analysis (LDA). Consider a

candidate radical S_1 and the remaining part S_2, we denote the intra-radical weight and inter-radical weight by $w(S)$ and $B(S_1,S_2)$, respectively,

$$w(S)=\frac{1}{N}\sum_{i,j=m}^{n} dw(s_i,s_j),\quad S=s_m\cdots s_n,$$
$$B(S_1,S_2)=db(c_1)+db(c_2), \qquad (1)$$

where N is the number of stroke pairs, $dw(s_i,s_j)$ is the minimum distance between two strokes s_i and s_j, $db(c_i)$ is the pen-up distance preceding the cutting stroke c_i. The function J is computed by

$$J=\frac{w(S_1)+w(S_2)}{B(S_1,S_2)}, \qquad (2)$$

By computing the value of function J for all the cutting hypotheses, the top K hypotheses of maximum J are retained to pass to binary classification (K was set as 10 in our experiments).

4.2. Special Radical Decision

Each candidate radical retained in special radical hypothesis is classified by 19 binary classifiers (each for a special radical class) to decide whether it is a special class or not. We use the support vector machine (SVM) for binary classification on statistical feature vector representation of candidate radicals.

For a special radical class, a SVM is trained with positive radical samples and negative samples (of different classes or non-radicals). Negative samples are collected using the bootstrap strategy: first train a base classifier using positive samples and negative samples segmented from special-structure characters. The base classifier is used to classify candidate radicals segmented from non-special-structure characters, and those with high classification score are collected as additional negative samples for retraining the classifier. In this way, the classifier is trained iterative for three times. For each binary classify, the number of positive samples ranges from 200 to about 5,000, and the ratio of positive samples to negative samples is kept about 1 to 20.

5. Experimental Results

We evaluated the performance of the proposed special radical detection method by statistical classification on a dataset of online handwritten Chinese characters of 6,763 classes (in the standard GB2312-80), each class with 60 samples produced by 60 writers. We used 50 samples per class for training classifiers, and the remaining 10 samples per class for evaluating the recognition performance.

Each candidate radical undergoes the same procedures of trajectory normalization and direction feature extraction as done for holistic character recognition [14]. Specifically, a moment normalization method is used to normalize the coordinates of pen trajectory points, and 8-direction histogram features (512D) are extracted directly from pen trajectory. For radicals of non-special-structure characters, the 512D feature vector is reduced to 160D by Fisher linear discriminant analysis (LDA) for accelerating classification; while for special radical detection, the 512D feature vector is input to the SVM for classification.

For radical classification of non-special (SE/LR/UD) structures, we used a modified quadratic discriminant function (MQDF) classifier [15] with 20 principal eigenvectors per class. For special radical detection, we tested SVM classifiers with two types of kernels: polynomial kernel (SVM-poly, order 4) and Gaussian kernel (SVM-rbf). In training SVMs, the upper bounds of multipliers were set to 10.

We evaluate the performance of special radical detection in terms of the rates of Recall (R) and Precision (P), which are defined as

R=number of correctly detected radicals /number of true radicals (3)

P= number of correctly detected radicals /number of detected radicals (4)

Table 4 gives the effect of special radical detection. We can see the detection rate (Recall) is high enough such that special-structure characters will be passed to the modules of special-structures. For characters of non-special structures, some candidate radicals (non-special radicals) in the stroke sequences are similar to one of special radicals in shape, and consequently, these characters are misclassified as special structures. The misclassification rate is about 2%. Fig. 6 shows some characters that are detected special radicals. Such mis-detected special radicals can be rejected in the subsequent recognition of remaining parts.

Table 4. Performance of special radical detection.

Type	#Character	svc-poly(%)		svc-rbf(%)	
		R	P	R	P
Special	502	99.83	97.34	99.65	97.29

Figure 6. Examples of misdetection of special radicals.

Based on special radical detection, recognition of the remaining part and non-special-structure character, we obtained recognition results on all the 6,763 classes. Using SVM-poly and SVM-rbf for special radical detection, the whole-character recognition accuracy on the test samples are 95.73% and 95.71%, respectively. This accuracy is relatively low compared to that of special radical detection. We observed that the confusion between special-structure characters and non-special-structure characters is considerable. This is because we did not elaborate the fusion of radical similarity scores to give a proper character similarity. Refining the whole-character similarity should help to reduce some recognition errors.

6. Conclusion

We presented a special radical detection method for online handwritten Chinese character recognition so as to realize a system for recognizing characters of both special structures and non-special structures. We use some heuristic but effective rules of geometrics to hypothesize candidate special radicals from the input character, and the candidate radicals are classified by binary SVM classifiers, one for each special radical class. Our experiments demonstrate that the proposed method yields very high Recall rate of special radical detection, and mis-detected radicals can be verified by remaining part recognition. The whole-character recognition rate is not sufficiently high but can be improved by elaborating the fusion of radical similarity scores in the future.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under grants no.60775004 and 60825301.

References

[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] J. Kwon, B. Sin, J.H. Kim, Recognition of on-line cursive Korean characters combining statistical and structural methods, *Pattern Recognition*, 30(8): 1255-1263, 1997.

[3] K.-W. Kang, J.H. Kim, Utilization of hierarchical stochastic relationship modeling for Hangul character recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(9): 1185-1196, 2004.

[4] S. H. Lee, J.H. Kim, Complementary combination of holistic and component analysis for recognition of low-resolution video character images, *Pattern Recognition Letters*, 29: 383-391, 2008.

[5] Y.J. Liu, L.Q. Zhang, J.W. Dai, A new approach to on-line handwriting Chinese character recognition, *Proc. 2nd ICDAR*, Tsukuba, Japan, 1993, pp.192-195.

[6] M. Nakai, N. Akira, H. Shimodaira, S. Sagayama, Substroke approach to HMM-based on-line Kanji handwriting recognition, *Proc. 6th ICDAR*, Seattle, WA, 2001, pp.491-495.

[7] K. Chellapilla, P. Simard, A new radical based approach to offline handwritten East-Asian character recognition, *Proc. 10th IWFHR*, 2006, La Baule, France, pp.261-266.

[8] L.-L. Ma, C.-L. Liu, A new radical-based approach to online handwritten Chinese character recognition, *Proc. 19th ICPR*, Tampa, FL, 2008.

[9] L.-L. Ma, C.-L. Liu, On-line handwritten Chinese character recognition based on nested segmentation of radicals, *Proc of 2009 CCPR & First CJKPR*, Nanjing, China, 2009, pp.929-933.

[10] 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.

[11] Z.-T. Lin, K.-C. Fan, Coarse classification of on-line Chinese characters via structure feature-based method, *Pattern Recognition*, 27(10): 1365-1377, 1994.

[12] A.-B. Wang, K.-C. Fan, W.-H. Wu, Recursive hierarchical radical extraction for handwritten Chinese characters, *Pattern Recognition*, 30(7): 1213-1227, 1997.

[13] H. Cao, A.C. Kot, Online structure based Chinese character pre-classification, *Proc. 17th ICPR*, Cambridge, UK, 2004, pp.395-398.

[14] 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.

[15] 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.