Journal of the industrial Technology Research Institute,
Korea Maritime University, Vol. 10, 1992.

Korean character recognition using Directional Information of character contour

Hwang Seung - Wook, Kang Sun Mi

문자 윤곽선의 방향정보를 이용한 한글인식

황 승 욱 · 강 선 미

Abstract

In this paper, a new method is proposed to extract feature primitive and composing feature vector for character recognition, classification methods efficient to hierarchical structure of feature vector and making the similar character dictionary for correct recognition of similar character. By applying thinning process, which is used in preprocess to extract feature primitive for high speed character recognition, four directional information is used extracted by thinning template. Feature vector is structured hierarchically by nonary-tree, each node of the tree is composed by sum of each directional feature primitive of its nine subnodes. All the feature primitives extracted by template is given to the corresponding leaf and repeatedly accumulated to the higher node. By using this hierarchical structure of feature vector efficiently, we reduced the number of candidates gradually lst and 2nd generation feature vector in the classification process respectively, and tried to reduce the overall amount of calculations. For high recognition rate of similar character, produced by a little stroke of character in a vowel or consonant, which characteristic of Korean, we used the premade similar character dictionary using training character set. As a result of these, even if there are some difference between character sets, it is possible to recognize about 98%.

* Dept. of Electronic engineering, Korea university

1.Introduction

The technique of recognization of characters by processing document image is requested for a new computer input device of all kinds of document information in several fields. This technique is replacing the method of character inputting using key boards. And its usage is expected to be used widely, because the growth of information society and the continuous automation.

The character recognition is a process of encoding a character image, which gets rid of the noise, character region extraction, nomalization of chatacter size, and thinning process. A method to recognize characters from pre – processed image is structural analysis, which is a technique based on the structural relationship among strokes. The other one is patten matching method, which calculates the similarity between the standard pattern and input.

In this paper, the proposed recognition algorithm is based on the pattern matching method, which uses high speed feature primitive extraction, construction of hierarchical feature vectors, classification using the constructed feature vector, and discrimination using similar character dictionary. the

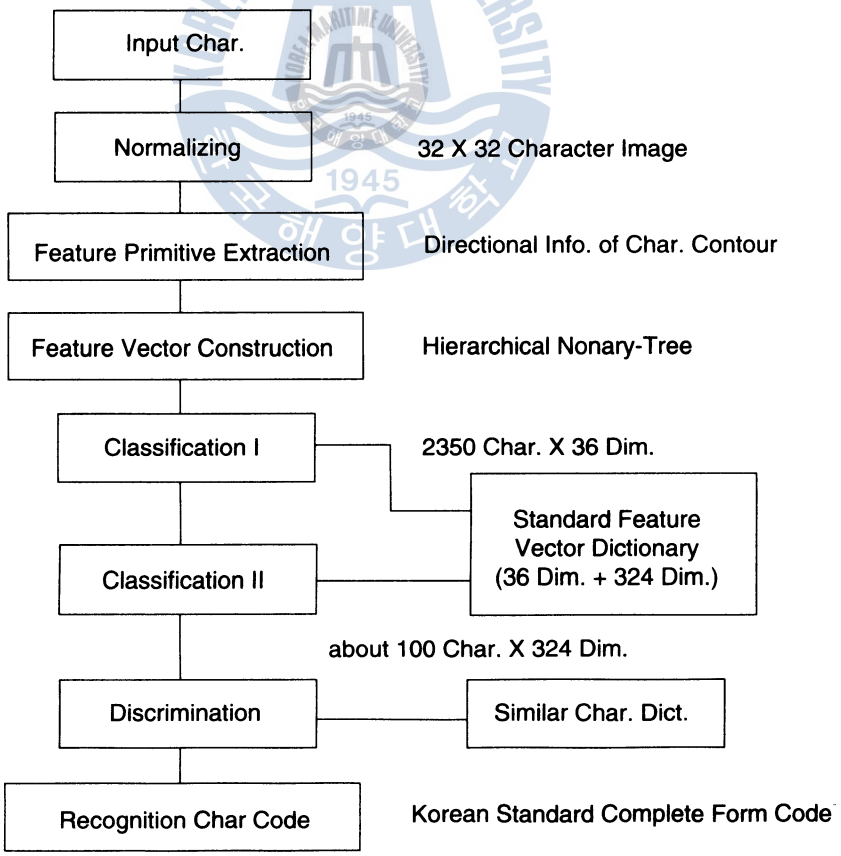


Fig. 1 Flow chart for the proposed algorithm

proposed overall algorithm is shown in Figure 1.

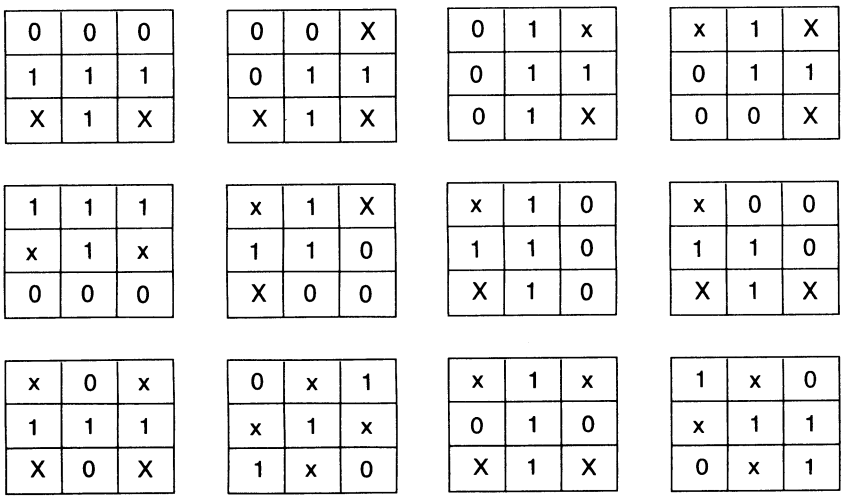
In general pattern matching method requires accurate and stable feature extraction[1]. Keeping this in mind, a new method is proposed, which uses high speed feature primitive extraction, construction of hierarchical feature vectors, classification using the constructed feature vector. By using similar character dictionary having partial information of Korean which has many similar characters, highly accurate recognition of similar character recognized faultily became possible. Since there isn't any standard in Korean, we made twelve sets of Korean image data base(KS 2350 characters) for research and used in the experiment. As a result, there are some variance between character sets, but we got 98% recognition rate.

2. Feature Extraction

2.1 Feature Primitive Extraction

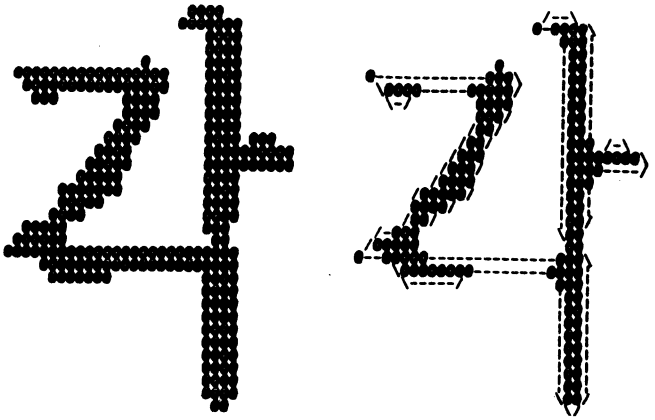
Based on the idea that a thinning template is usually extracted by direction, we use it as feature primitive for character recognition. According to the relative position of the neighboring pixel, the extracted pixel has directional and positional information. And the amount of information can be controled according to the thnning iteration.

Among the thinning algorithm[2,3], the thinning template proposed in the fast One – pass thinning algorithm was used to creat twelve feature primitive extraction template. Four directional element ('–', '/', ':', '\')are extracted as a feature primitive[4].



(–)Direction (/) Direction (:)Direction (\) Direction

Fig. 2 Proposed templates

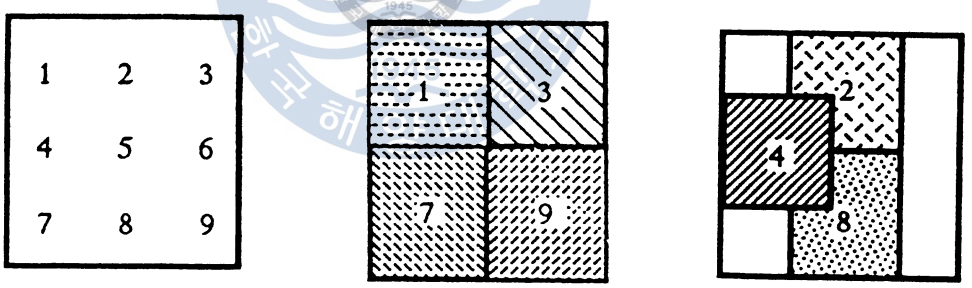


(a) Input image (b) Extracted feature primitives

Fig. 3 Feature primitives extracted by proposed templates

2.2 Construction of Hierarchical Feature Vector

The hierarchical feature structure methode using Nonary – tree is as follows. When the input character size becomes $2^N \times 2^N$ (in the experiment N=5) after normalization. By superposing 2^{N-2} in a column and in a row, we get sub – region of nine, which size is $2^{N-1} \times 2^{N-1}$ as Figure 4.

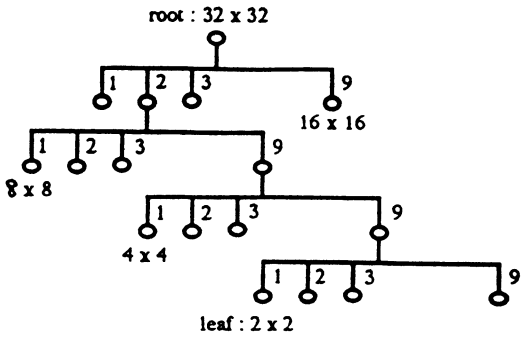


(a) Corresponding region of nonary – tree (b) Sub – region 1, 3, 7, 9 (c) 2, 4, 8

Fig. 4 Examples of subregion

Again superposing 2^{N-3} in a colum and in a row of each sub – region we can make nine sub – region ($2^{N-2} \times 2^{N-2}$). If repeat this process until it is possible to split to sub – region, and match character region to the nonary – tree, then the Root will be the whole normalized character region, and the Leaf will be the sub – region of $2^1 \times 2^1$. all nodes of nonary – tree have four directional features. Feature vector of each generation is as figure 5.

Since the feature vectors of a node is addition of nine subnodes' feature vector, the feature primitives collected at fourth generation will have weights 1, 2 and 4 to the higher node (third generation) depending on its position. When a feature primitive is extracted from a point (x, y) of normalized character region, the effect on the first and second generation nodes get weights as Figure 6. Since



(a) Nonary – Tree

Gener ation	Nodes	Feature Vectors
0(Root)	9^0	4
1	9^1	36
2	9^2	324
3	9^3	2,916
4(Leaf)	9^4	26,244

(b) Feature Vectors of Each Generations

Fig. 5. Nonary tree and feature vectors of each generations

hierarchically divided subregion gets information (i. e. partial information) of regional character pattern, we get accurate information of subregion. This feature vector construction is similar to human recognition process, which finds the overall characteristics first, then th the details. Therefore, the most suitable feature used in recognition can be applied step by step from the feature of each generation, considering the complexity of pattern, processing time and degree of recognition. Since the feature is constructed of Nonarv tree, emphasized in the center of region and weakened in the outskirts. It is stable for the displacement of strokes.

(a) Second Generation (8×8)

(b) First Generation (16×16)

Fig. 6. Weight of feature vectors for each generations of nonary – tree

3. Classification Process

According to the general classification method, the input pattern is compared with the standard pattern for each character. In Korean and Chinese character recognition, Which has complex strokes and variety of characters, requires many feature vectors. thus, much processing time is consumed. By using the hierarchical strusture of feature vectors, we can improve the processing speed. We are going to show a simple classification method.

3.1 Classification Experiment

Korean character sets used as a training and test for recognition experiment are 2350 characters of KS complets form. We made twelve sets of character, using four kinds of laser printers, from which we printed out three sizes of font. Aming them, we used eight sets as training sets and the rest four sets as test sets.

We have verified the effectiveness of feature primitives and feature vector construction by classification experiment of comparing the distance between standard feature vector extracted from traning sets and the one extracted from testing sets. The distance calculation method used in classification is city – block distance, which considers feature vectors as n – dimensional vectors, and adds the absolute distance between the standard pattern and the test pattern of each generation' s feature vectors.

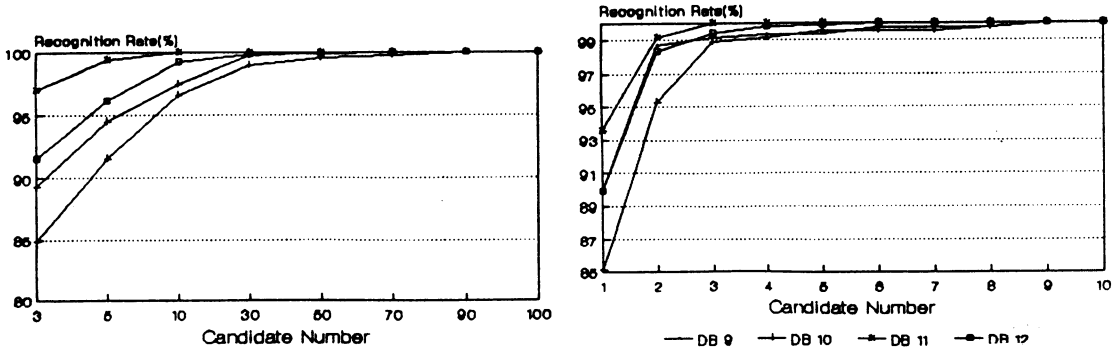
$$D(X, s) = \sum_{i=1}^{N} |X_i - S_i| \qquad \cdots\cdots\cdots\cdots (Eq - 1)$$

Where, X_i : Feature Vector of Input Character

S_i : Standard Feature Vector of Training Set

N : Dimension of Feature Vectors

The result of classification using the feature vectors of first generation node(36 dimensions) and the second generation(324 dimensions) in a Nonary – tree is as Figure 7. When the feature vector of 324 – dimension was applied, the probability to be classified within ten candidates is highly accurate as 99.9 percents. When the feature vector of 36 – dimension was applied, we could get the same classification rate within one – hundred candidates. From this, we could observe that it could be appropriately used in the delopment of recognition algorithm of applying the proposed feature vector to classification and discrimination.



(a) first generation (36 dimensions) (b) second generation (324 dimensions)

Fig. 7. Results of classification experiments

3.2 The classification using the Hierarchical Structure

The basic concept of classification for character recognition is efficient collection of the similar characters to the unknown input character. At this time the same character as input character must be included among the colecter candidate characters. If we use the second generation feature vector, about ninety percent were included in the first rand and 99% could be considered within the third rank candidate.

To improve the correct recognition rate, we have to use the second generation feature vectors for recognition but there exists a problem in processing speed to select the candidate characters from

Table 1. Comparison amount of calculation

2nd Generation	Using Hierarchical Construct	
2350 Char × 324 Dim.	2350 Char × 36 Dim.	Classification Ⅰ
	100 Char × 324 Dim.	Classification Ⅱ

Korean character(2350 characters of KS Complete Form), which requires a lot of calculation. Observe the Figre 7, we can find the characteristics of feature vector's hierarchical sturcture. When we use the first generation feature vector of thirty – six dimension, we could get 99.99 percent classification rate within the 100th rank of candidate. Based on this idea, we reduced the number of candidates to one – hundred using the firsr generation featue vectors. From these candidates we select one character by calculating the city – block distance from the second generation feature vectors of 324 – dimension. As a result the overall calculation time was reduce about one – sixth of the previous method.

4. Descrimination Using Similar Character Dictionary

Implementing the previous method, the result is shown in Table 2. If we analyze the incorrectly recognized characters of the Table 2, usually the difference is one stroke or two short strokes. If there is much noise in the character image and/or thick strokes are connected together during the preprocess. It is hard to make correct feature vector extraction. 'ㅃ' or 'ㄸ' and 'ㅍ' etc. are the examples. Also, '옷' and '웃', '옹' and '웅', '웁' and '움', and '쎈' and '쏀' could be recognized as same characters on some fonts of printers. It could be distionushied only by the meaning of the sentence. Most cases, however, in the change of vowels, i. e. 'ㅐ' and 'ㅒ' or 'ㅏ' and 'ㅑ' etc., are making trouble in correct recognition. So, by calculating the similatity on those subregions, it is posible to descriminate what are hard to recognize.

Table 2. Example of incorrect recognition

Character Set	Input Character	Candidate Character			
		1st	2nd	3rd	4th
DB1	넘	님	넘		
DB2	뜰	똘	뚤	뜰	
DB3	쳉	셍	쟁	셍	쳉
DB4	룽	룡	룽	룽	
DB5	몌	메	몌		
DB6	뼝	뼝	뼝		
DB7	욥	옴	욥		
DB8	봇	못	봇		

4.1 Method to Making Similar Character Dictionary

The similar character dictionary was made from the Korean training character set(DB 1 - 8) as the following method.

1)If it is not recognized in the first rank using the method explained in the chap. 3.2, the incorrectly recognized characters are registered in the same similar character group. In case of correct recognition, if the distance of second and/or third randed character is very close to the first ranked character is also registered.

2)Once the similar characters are grouped, we put the first ranked characted of each group to the all candidates, which are ranked first in their group.

3)The basis for deciding the recognition region is the amount of variance of characters' feature vectors in a group. Through the adoptability test with the sorted variance and training character set, we decide the region. Also, this region information is recorded in the similar character dictionary.

Table 3. Exmaple of Dictionary

Candidates	Recognition Region(Index of Feature vector
간 갼	64,68,132,148,185
갓 걋	132,240
객 책	75,103,107,159,169,171,187,197,199,200
걸 길	108,136,140,192,220,224
검 겸	236
…	

It is essential in high accuracy recognition of Korean character, even if it requires extra calculation of similarity according to the dictionary. Since, the dictionary may not be complete for all character sets, we made it possible to update the dictionaty partially by training.

If the number of candidates are more than

five, it is hard to decide the recognition region. Thus, we made limitation in the number of candidates to be five and the number of region to be fifty.

4.2 The Result of Discrimination

The result of applying the discrimination for the test character sets is shown in table 4. In the table, you can see the recognition rate is about ninety – eight percents depending on the character sets. If we improve the adoptability for the second candidate in the dictionary, we can expect above ninety – nine percents of recogition rate. But, as the number of candidates increases in a group, it is getting harder to decide recognition region. So, there is trade – off between the number of candidates and deciding recognition region. There could be further research in utilizing the rejected candidates.

Table 4. Comparison of recognition rate before and after discrimination

Test Character Set	After Discrimination	Before Discrimination		
		1st	2nd	3rd
DB 9	98.60	89.09	97.81	99.23
DB10	96.62	85.04	95.26	98.87
DB11	98.47	93.40	98.94	99.79
DB12	99.30	94.04	98.94	100

5. Conclusion

We proposed a new method, to extract featract feature primitives and hierarchical feature vector in high speed, to recognize a character based on the pattern matching and proved that it is resonable through a simple classification experiment. In the implementation of recognition system, we used the first and second generation's feature vectors in the classification. By reducing the number of candidates, reduction of calculation amount became possible. High accuracy recognition is also possible by using the similat chatacter dictionary(Korean has a lot of similar characters), which is the partial feature vectors of the second generation. We are planing to continue our research in improved non – linear normalization, classification method which emphasizes characteristics of characters, and finding more efficient distance calculation method.

6. References

1) N. SUN, T. Tabara, H. Aso and M. Kimura, "Printed Character Recognition Using Directional Element Feature", IEIC of Japan D – Ⅱ vol. J74 – D – Ⅱ No. 3 pp. 330~339 March 1991

2) D. L. Stover and R. D. Iverson, "A One – Pass Thinning Algorithm and Its Parallel Implementation", Computer Vision, Graphics, and Image Processing 40, pp. 30~40, 1987

3) T. Y. Zhang and C. Y. Suen, "A fast parallel Algorithm for Thinning Digital Patterns", Image Processing and Computer Vision Commun. ACM March Vol. 27 Number 3 1984

4) S. M. Kang, S. W. Hwang, Y. M. Yang and D. J Kim, "Extraction of Feature Primitives and Construction of Feature vectors Using Direction Information of Character Contour" KITE Conference Fall '91

5) T. Ejima, Y. Katsuyama and M. Kimura, "Rough Classification of Handwritten Character by Divide and Unify Method", IEIC of Japan '87/2 Vol. J70 - D No.2