Effective Handwritten Hangul Recognition Method Based on the Hierarchical Stroke Model Matching

Wontaek Seo and Beom-joon Cho*

Dept. of Computer Science, University of Maryland, College Park, MD 20742
Dept. of Computer Engineering, Chosun University, 375 Seosuk-dong,
Dong-gu, Gwangju, Korea
Tel.: +82-62-230-7103; Fax: +82-62-230-7381
wtseo@cs.umd.edu, bjcho@chosun.ac.kr

Abstract. This study defines three models based on the stroke for handwritten Hangul recognition. Those are trainable and not sensitive to variation which is frequently founded in handwritten Hangul. The first is stroke model which consists of 32 stroke models. It is a stochastic model of stroke which is fundamental of character. The second is grapheme model that is a stochastic model using composition of stroke models and the last is character model that is a stochastic model using relative locations between the grapheme models. This study also suggests a new stroke extraction method from a grapheme. This method does not need to define location of stroke, but it is effective in terms of numbers and kinds of stroke models extracted from graphemes of similar shape. The suggested models can be adapted to hierarchical bottom-up matching, that is the matching from stroke model to character model. As a result of experiment, we obtain 88.7% recognition rate of accuracy that is better than those of existing studies.

1 Introduction

A character is composed by joint of several strokes. The union and location of each stroke become very important information in recognizing a character. Besides, the other existed information in a character can be aware as noises which are occurred through a user or an input device. As this view, recognizing a handwritten character by union and location of each stroke is very common process and this is considered as structural method. The structural method can be completed under the hypothesis which the position of each stroke becomes a most important information of recognizing an independent character [1, 2].

The most methods, which have used strokes in the past, have expressed strokes and their relation by heuristic. For relation between each stroke, they used slope between a strokes and surrounding strokes [5, 6]. There was an approach which used symbolic way between each stroke. The types of stroke are divided as horizontal, vertical, left

* Corresponding author.

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 431 – 440, 2006.

diagonal, and right diagonal and the relationship of stroke is divided as L form, T joint, parallel and the others [7, 8].

Yet, this heuristic method is insufficient for practical uses because it is very sensitive with noise of input character. Furthermore, there is a limit which is difficult to be trained. In these days, statistic method has been introduced which uses graph modeling; a stroke is presented by probability of stroke slope and its length and the relationship of each stroke is presented by their relative location [3]. There is another method which uses a systematic relation between each stroke. In this method the relative information of each stroke are presented through statistic dependence [4].

In this research, a new stroke model will be presented and the composition method of grapheme models and character models will be introduced. And also, matching method of each model with statistic way and recognizing method of handwritten Hangul character by using characteristic of our own class composition will be introduced. In chapter 2, the characteristics of Hangul characters will be explained. Three new proposed models (stroke model, grapheme model, character model) and their composition and matching method will be explained in chapter 3, and experiment result will be shown in chapter 4, and the conclusion will be in chapter 5.

2 Characteristic of Hangul

2.1 Composition of Hangul

Hangul is a phonetic alphabet which one character has an independent sound, and it is constructed of consonant and vowel those are arranged on two dimensional spaces. Although there are only 24 vowels and consonants, the number of character which can be made through composing of them is 11,172. However, the practically being used characters are 2,350 and the commonly being used characters are only 520 of them.

Table 1. Shape and position of graphemes of Hangul

Groups	Grahpemes	Position
Basic consonant	ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅅ,ㅇ,ㅈ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ	FC, LC
Basic vowel	ㅏ,ㅑ,ㅓ,ㅕ,ㅣ	VV
	ㅗ,ㅛ,ㅜ,ㅠ,ㅡ	HV
Combined consonant	ㄲ,ㄸ,ㅃ,ㅆ,ㅉ	FC, LC
	ㄳ,ㄵ,ㄶ,ㄺ,ㄻ,ㄼ,ㄽ,ㄾ,ㄿ,ㅀ,ㅄ	LC
Combined vowel	ㅐ,ㅒ,ㅔ,ㅖ	VV
	ㅘ,ㅙ,ㅚ,ㅝ,ㅞ,ㅟ,ㅢ	HV + VV

Hangul has six important composition formats which consonant and vowel are arranged on 2 dimensional spaces. Every character is constructed by the six formats and there are rule which consonant and vowel are used for each format. Therefore, distinct recognition of formatting information will make us much easier to recognize a character.

There are 14 of basic consonant, 5 of double consonant which makes strong sounds, and 11 of repeated consonant which is composed of two different basic consonant. The First Consonant (FC) decides early stage of pronunciation, and the Last Consonant (LC) does later stage of pronunciation. Vowel settles the middle parts of pronunciation. There are 10 basic vowel and 11 combined vowel. There are two types of vowel according to the shape, Horizontal Vowel (HV) and Vertical Vowel (VV). As above, each pronunciation and the meaning are decided by the composition of 2~4 of vowel and consonant, which is shown in table 1.

2.2 Hierarchical Decomposition of Hangul

Hangul shows hierarchical joint structure: several strokes joint together to make a form of grapheme and the grapheme joint each other to make an independent character on the 2 dimensional spaces. Therefore, we can divide a character using opposite way of jointing: a character is divided into vowel and consonant depending on the format type and those are divided again into strokes. In here, a stroke means basic stroke such as '—', ' | ' and composed stroke such as 'ㄱ', 'ㄴ' by Korean writing style.

3 Proposed Models

3.1 Stroke Model

Stoke model is suggested for effective modeling of stokes of lower parts in a point of top-down decomposition. The most common strokes in Hangul are horizontal, vertical, left diagonal, right diagonal and circle such as in 'ㅇ', 'ㅎ'. In this research, a model of jointed stroke rather than single one is proposed. Using four basic strokes, which are horizontal line, vertical line, left diagonal line and right diagonal line, we made 32 kinds of joint strokes which are shown in fig. 1. The circle stroke is excluded because those introduced strokes can make circle by jointing each other.

Each stroke models has parameters of edges and nodes. Each edge's probability distribution of directive angle is existed and the connect part node has probability distribution of connect angle of two edge.

A matching of stroke models is the first stage of matching process. After we set the ending point, connecting point and the bending point as a fixed node of the graph, as well as a attributed graph by extracting an edge between nodes, we find out a stroke model which is the best matching with the part of attributed graph. For this, production of sub graph which matches with stroke model is needed.

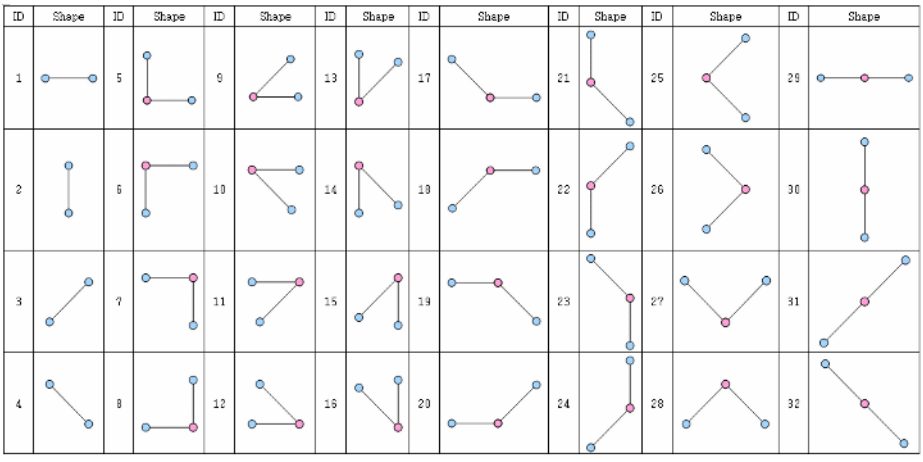

Fig. 1. Shape of proposed Stroke models

Fig. 2 shows an attributed graph decomposition method into sub graph for matching of stroke model with an example of grapheme '入'. In the set of edges, if two edges share a node, sub graph which has two edges and a node is extracted to match with stroke models. This method is very successful way because there is no need to define the positional relation by heuristic, and there is only one stroke model combination with a grapheme. Strokes that are shown as a single stroke also have two strokes which are same directions.

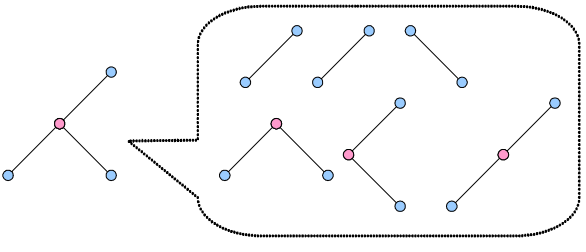

Fig. 2. Example of sub graph extraction from an attributed graph

The matching of output sub graph and the stroke models is calculated from multiplication of parameter of stroke models of direction d of sub graph's edge, and multiplication of parameter of stroke models of joint angle a of sub graph. The parameter of stroke models is probability distribution as explained above. Eq.(1) is the calculating the matching probability of stroke model.

$$P(M_S \mid X) = \prod_{x \in \forall E, k \in d,a} P(x_k) \tag{1}$$

X is a sub graph of attributed graph, and M_S is stroke model. As it is shown, matching probability of stroke model can be calculated from multiplication of probability of input.

3.2 Grapheme Model

Grapheme consists of combinations of several strokes, so it is presented as relationship of each stroke. As a result of using the extracting method, which is suggested previous, the kinds of stroke model and its frequency can be very good information. The result of extraction of stroke model from, '口' and '日' by using their matching method is shown at fig. 3. '日' seems it is added only two more lines to '口', but it's added 6 different strokes when we decompose to stroke model. Using this characteristic, grapheme model which has probability distribution of each stroke model frequency of occurrence is defined.

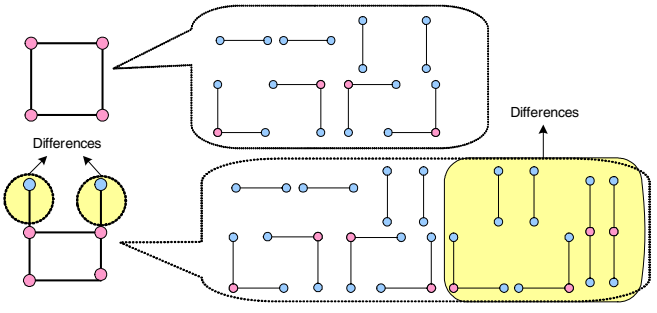

Fig. 3. Comparison of stroke models extracted from two graphemes of similar shape

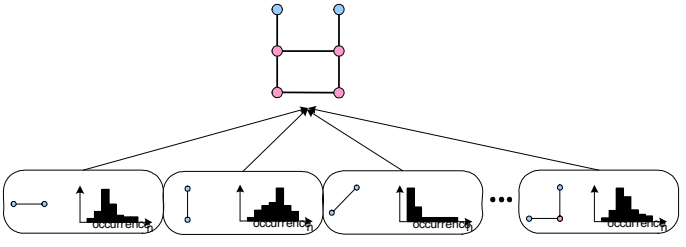

Fig. 4. Example of probability distribution of a grapheme model

There are 51 graphemes existed in Hangul as it is shown at table 1. Therefore 51 grapheme models are defined in this research.

Fig. 4 is an example of parameter of grapheme model of grapheme '日'. Because it is composed of horizontal stroke and vertical stroke, their probability of high

frequency is high, but the probability of high frequency of diagonal stroke's occurrence is low.

As shown in fig. 4, using probability distribution of occurrence frequency of 32 stroke models, 51 grapheme models are defined, and matching probability is defined by the average of occurrence frequency of 32 stroke models. The equation is shown at (2).

$$P(M_G \mid X) = \frac{1}{n} \times \sum_{i=1}^{n} P(O(S_n)) \tag{2}$$

X is the set of stroke model, M_G is a grapheme model, n is the number of model, and S_n is the nth stroke model, and $O(S_n)$ is the number of occurrence of S_n, and $P(O(S_n))$ is the probability of S_n.

3.3 Character Model

Character is made through an arrangement of several graphemes on 2 dimensional spaces, and meaning and pronunciation are concluded by their location and variety. In this research, character model is defined by using this characteristic. Probability distribution of relative position of each grapheme is used to present the character model. The position of each grapheme is already defined according to the character type, but it is available only after we recognize a character, therefore the grapheme information of location cannot be used. Relative location of grapheme is defined as horizontal, vertical, right diagonal, and left diagonal relationship. The example of positional relationship is shown in Fig 5. The relation between 'ㅅ' and 'ㅏ' is horizontal relation, and the relation between 'ㅅ' and 'ㄹ' is left diagonal relation.

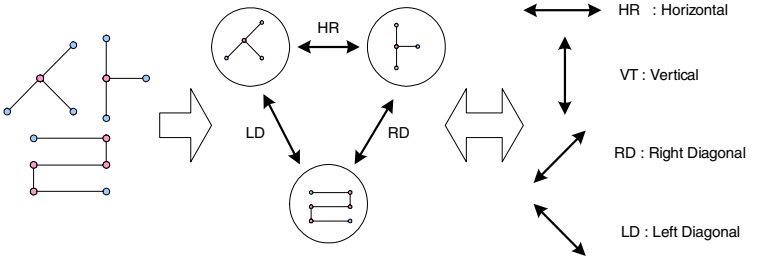

Fig. 5. Positional relationship between grapheme models

For probability matching with graphemes extracted from grapheme matching, character model has to constitute parameters. The parameters of character model are

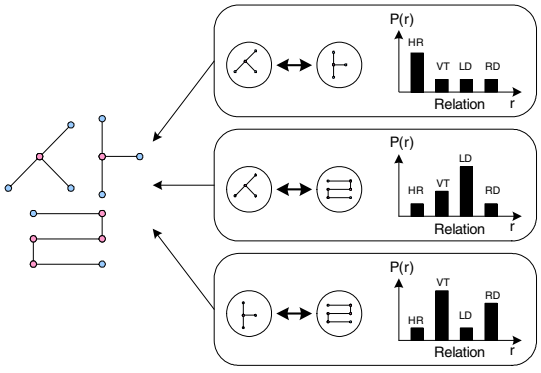

Fig. 6. Example of probability distribution of a character model

graphemes and probability distribution of their relative position. An example of the composition of character model is shown at Fig 6.

As graphemes and its positional relation which are existed in a character are defined by probability and compose an independent character model, the compositions work out for each of 2350 complete characters in Hangul.

By multiplication of parameters of every grapheme models extracted from grapheme matching stage in character model, matching probability of character is calculated. The formal equation is shown in Eq.(3).

$$P(M_C \mid X) = \prod_{x \in X} P_r(x \mid N(x)) \tag{3}$$

In above, X means outputs of matching stage of grapheme model, $N(x)$ means neighbor graph of x and $P_r(x \mid N(x))$ means the probability of relation of x and the neighbor x.

3.4 Recognition

Recognition in the statistic model means a process of deriving the maximum probability from a relation between data and a model. In this research, every graph model and its node and edge are models which are presented the distribution of probability, therefore calculating the maximum probability between a model and input is a matching method.

In this research, the method of calculating the recognition probability of character is shown in Eq.(4)..

$$P(M \mid X) = \prod_{i=1, x \subset X}^{n} P(M_{G_i} \mid x) \times P(M_C \mid X) \tag{4}$$

In the above, $P(M_G \mid x)$ means matching probability of model G_i to the set of input stroke x and $P(M_C \mid X)$ means matching probability of according character model. The character recognition probability is calculated when each grapheme multiplies $P(M_G \mid x)$ and $P(M_C \mid X)$. The character which produces the highest probability will be the matching character.

3.5 Hierarchical Bottom-Up Matching

In this research, the method of recognition through hierarchical bottom-up matching process using stroke, grapheme, and character model is proposed. First, stroke and its model need to be matched, and then, perform matching with grapheme's model using the extracted stroke model. In this process, for every condition of possible stroke model's set, we must perform the matching with grapheme's model. At last, the set which can compose the character at grapheme model's set has to be extracted and need to calculate the matching probability. The character model from the process and each grapheme that attends the matching are multiplied to get recognition probability and the character which has the highest probability can be the result of recognition.

4 Experimental Results

The experiment in this research used the database of common handwritten Hangul database PE92. PE92 consists of each 100 sets of character image to the total 2350 of Hangul character; 40 sets for training and 60 sets for test.

The probability of recognition is being compared while the numbers of characters are being limited, in this experiment, the 520 of practical character and 2350 of character can be found at table 2. As it shown in the table, we could get 90.5% of accuracy about 520 of characters and 88.7% for 2350 of characters. Also, as a result of considering the 5th recognition candidates, we could get 95.5% of accuracy for the 520 characters and 95.2% for the 2350 characters. In here, those are recognized as similar forms, but not often as different character.

Table 2. Recognition result of proposed method

Number of category (char.)		520	2350
Recognition rate(%)	1st	90.5	88.7
	2nd	92.3	90.3
	5th	95.5	95.2

When we compare the result of this research to the previous researches [3,4], we can claim that the method, which records the increasing 1% rate of accuracy with the

standard of perfect characters, is much more excellent than other methods. [3] composes three kinds of models, primitive stroke model, grapheme model and character model using random graph modeling, but the definition of the models is quite different to those of models suggested in this paper. There are some problems which the grapheme modeling use stroke model extracted from the other stroke groups and it takes more time to match all different strokes. Oppositely, the method which is suggested in this paper uses only the stroke models which are extracted from identical stroke group of grapheme model matching by decomposing the stroke group. Therefore, this method brings the solution to the previous problems. In the study [4], the condition of limitation about grapheme production prevent the production of it in unusual spaces, yet the way of establishing the condition rely too much on heuristic. However, in our research, by using the relative probability in decomposing and composing of stroke group, we can present every process in a correct range of calculation of probability.

Table 3. Performance comparison to the previous works

method	hierarchical random graph[3]	stochastic relationship[4]	proposed
Rec. rate(%)	86.3	87.7	88.7

5 Conclusion

In this paper, we proposed the new stroke based models and matching methods for an effective Hangul recognition system. The stroke model is defined using 3 probability distributions which two are direction of edges and one is angle between two edges. 32 stroke models are defined based on the composition of 4 basic strokes. The grapheme model is defined based on the new stroke extraction method. The frequency of each stroke extracted from a grapheme is modeled as grapheme model. The character model is defined using relative position between each grapheme. Hierarchical bottom-up matching can be adapted to these three models, because the concept of model definition is started from the structural characteristic of Hangul. As a result of experiment, we could get the high performance of recognition of 88.7% compare to previous research as well as better result in the periodic problems.

The model introduced in this paper will be successful for not only for Hangul recognition, but also for the other characters such as Chinese which has also very complicated structure.

Acknowledgement

This study was supported in part by research funds from Chosun University, 2004.

References

1. A. K. C. Wong, D. E. Ghahraman, “Random Graphs: structural-contextual dichotomy”, IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 2, No. 4, pp. 341-348, 1980.
2. A. K. C. Wong and M. You, “Entropy and distance of random graphs with application to structural pattern recognition”, IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 7, No. 5, pp. 599-609, 1985.
3. Ho-Yon Kim, Jin H. Kim, “Hierarchical random graph representation of handwritten characters and its application to Hangul recognition”, Pattern Recognition, 34, pp.187-201, 2001.
4. Kyung-won Kang, Jin H. Kim, “Utilization of Hierarchical, Stochastic Relationship Modeling for Hangul Character Recognition”, IEEE PAMI. Vol. 26, No. 9, pp. 1185-1196, 2004.
5. F. H. Cheng, “Multi-stroke Relaxation Matching Method for handwritten Chinese Character Recognition”, Pattern Recognition, Vol. 31, No. 4, pp. 401-410, 1998.
6. H. j. Lee, B. Chen, “Recognition of handwritten Chinese Characters via Short Line Segments”, Pattern Recognition, Vol. 25, No. 5, pp. 543-552, 1992.
7. X. Zhang, Y. Xia, “The Automatic Recognition of Handprinted Chinese Characters – A Method of Extracting an Order Sequence of Strokes”, Pattern Recognition Letters, Vol. 1, No. 4, pp. 259-265, 1983.
8. C. L. Liu, I. -J. Kim, J. H. Kim, “Model-based Stroke Extraction and matching for Handwritten Chinese Character Recognition”, Pattern Recognition, Vol. 34, No. 12, pp. 2339-2352, 2001.