Bayesian Network Modeling of Hangul Characters for On-line Handwriting Recognition

Sung-Jung Cho and Jin H. Kim
CS Div., EECS Dept., KAIST,
373-1 Kusong-dong, Yousong-ku, Daejon,
305-701, KOREA {sjcho, jkim}@ai.kaist.ac.kr

Abstract

In this paper, we propose a Bayesian network framework for explicitly modeling components and their relationships of Korean Hangul characters. A Hangul character is modeled with hierarchical components: a syllable model, grapheme models, stroke models and point models. Each model is constructed with subcomponents and their relationships except a point model, the primitive one, which is represented by a 2-D Gaussian for X-Y coordinates of point instances. Relationships between components are modeled with their positional dependencies. For on-line handwritten Hangul characters, the proposed system shows higher recognition rates than the HMM system with chain code features: 95.7% vs 92.9% on average.

1 Introduction

For highly accurate character recognition, it is important to model character structures as realistically as possible. In this paper, the character structure is defined as the set of hierarchical components and their relationships; each component is hierarchically composed of subcomponents at a lower level and their relationships.

In the case of Korean Hangul characters, four levels of model hierarchy are found: syllable characters, graphemes, strokes and points. A syllable character is structurally composed of two or three graphemes: one first consonant (19 classes, denoted as *C*), one vowel (21 classes, *J*) and one optional last consonant (27 classes, *Z*). Theoretically, 11,772 syllable characters are constructed by the combination of graphemes, but only 2,350 ones among them are practically used. A grapheme, which corresponds to an alphabet in English words and a digit in digit strings, is composed of strokes. Here, a stroke is composed of points and defined as a nearly straight trace that has distinct directions from connected traces in writing order. Finally, a point is the primitive component and represented by (x, y) coordinate. Figure 1 shows an example of those components.

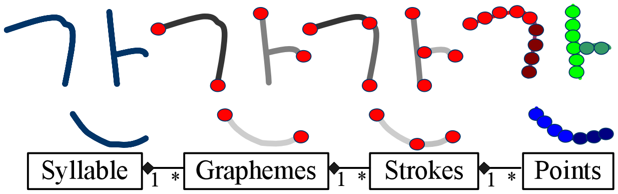


Figure 1. Components of a Hangul character.

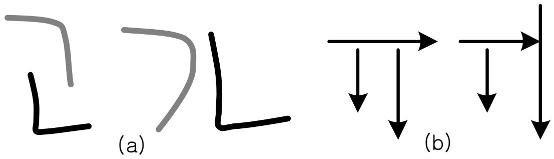


Figure 2. Importance of relationships (a) between graphemes. (b) between strokes.

Relationships between components are also important to identify and discriminate characters. Some different characters have same components so that only relationships between components discriminate them. Fig. 2 (a) and (b) shows examples that only relationships between graphemes and between strokes can discriminate different characters and graphemes respectively.

In spite of the importance, components and their relationships have not been actively adopted in previous studies. Hidden Markov models (HMMs) [6, 8], segmental HMMs [7] and time delay neural networks [3] are ones of the most popular frameworks. They are based on the assumption that a handwriting input is composed of independent local feature inputs. Therefore, they do not have parameters for modeling long-time relationships between feature inputs.

We proposed a Bayesian network framework for modeling components and their relationships of single digits and got promising results [2]. A digit model is composed of stroke models and point models in turn. Their relationships are modeled as conditional Gaussian distributions [5]. The proposed system outperformed an HMM system with chain code features and a neural network system in terms of recognition rates.

Our research goal is to extend the previous Bayesian network modeling framework to Hangul characters in three aspects. First, the large character classes (2,350) are modeled by sharing grapheme and ligature models. Second, the relationship models are extended invariant to scaling and translation variations of graphemes. Third, inter-grapheme relationships are modeled by position dependencies between bounding boxes of graphemes.

The rest of this paper is organized as follows. Section 2 briefly introduces Bayesian networks. Section 3 introduces the Bayesian network models for graphemes. Section 4 presents the Bayesian network models for Hangul syllables. Section 5 gives experimental results and Section 6 concludes this paper. Recognition and training algorithms of the proposed system will not be described in this paper.

2 Introduction to Bayesian networks

A Bayesian network [4] is a graph with probabilities for representing random variables and their dependencies. It efficiently encodes the joint probability distribution of a large set of variables. Its nodes represent random variables and its arcs represent dependencies between random variables with conditional probabilities at nodes. It is a directed acyclic graph (DAG) so that all edges are directed and there is no cycle when edge directions are followed.

The joint probability of random variables $\{X_1, \ldots, X_n\}$ in a Bayesian network S is calculated by the multiplication of local conditional probabilities of all the nodes. Let a node X_i in S denote the random variable X_i, and pa_i denote the parent nodes of X_i, from which dependency arcs come to the node X_i. Then, the joint probability of $\{X_1, \ldots, X_n\}$ is given as follows:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | pa_i) \qquad (1)$$

It is not a simple task to get the exact conditional probability distribution when random variables have continuous values and high order dependencies. A conditional probability table is not adequate for this because continuous values should be quantized and the table size grows exponentially with the dependency orders.

For these reasons, we adopt conditional Gaussian distributions [5]. The mean of a random variable is assumed to be determined from the linear weight sum of dependent variable values. The difference between the mean and the random variable value is assumed to be Gaussian. When a multivariate random variable X depends on $X_1, \ldots, X_n$, the conditional probability distribution is given as follows:

$$\begin{aligned} & P(X = \mathbf{x} | X_1 = \mathbf{x_1}, \ldots, X_n = \mathbf{x_n}) && (2) \\ = \quad & (2\pi)^{-\frac{d}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)] \end{aligned}$$

The mean μ is determined from the dependant variable values $\mathbf{Z} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T, 1]$ as follows:

$$\mu = \mathbf{W}\mathbf{Z}^T \qquad (3)$$

where $\mathbf{W}$ is a $d \times k$ linear regression matrix, d is the dimension of X, and k is the dimension of $\mathbf{Z}^T$.

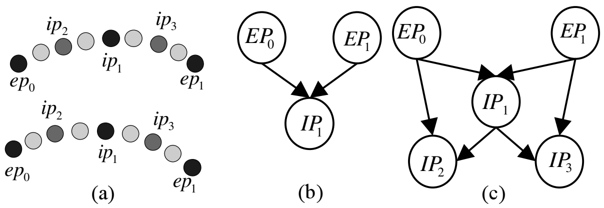


Figure 3. Recursive construction of a stroke model.

3 Bayesian network framework for graphemes (digits) [2]

3.1 Point model

A point instance has the attribute of (x,y) position on the 2-D plane. So, a point model has 2-D Gaussian distribution for modeling 2-D point positions. It is represented by one node in Bayesian networks.

When a point $P = (x, y)$ depends on other points $P_1 = (x_1, y_1), \ldots, P_n = (x_n, y_n)$, its matching probability $P(x, y | x_1, \ldots, y_n)$ is given from the conditional Gaussian distribution (Eq. (2), (3)) by setting $X = (x, y)$ and $\mathbf{Z} = [x_1, y_1, \ldots, x_n, y_n, 1]$.

3.2 Stroke model

A stroke instance is composed of points. Therefore, a stroke model is composed of point models with their relationships, called within-stroke relationships (WSRs).

A stroke model is constructed by recursively adding mid point models and specifying WSRs. At the mid point, the lengths of the left and the right partial strokes are equal. A WSR is defined as the dependency of a mid point from two end points of a stroke. Fig. 3 shows the recursive construction example of a stroke model. Fig. 3 (a) shows an example of stroke instances. At the first recursion ($d = 1$), IP_1 is added for modeling ip_1's with the WSR from EP_0 and EP_1 (Fig. 3 (b)). At $d = 2$, IP_2 and IP_3 are added for the left and the right partial strokes (Fig. 3 (c)). This recursion stops when the covariances of newly added point models become smaller than some threshold.

The matching probability of a stroke model S of the recursion depth d and a stroke instance of length t, O_1^t, is obtained from those of point models. At first, O_1^t is recursively resampled into $2^d - 1$ mid points $(ip_1, ip_2, \ldots, ip_{2^d-1})$. Then, the matching probability is calculated as follows:

$$\begin{aligned} & P(S = O_1^t) && (4) \\ = \quad & P(EP_0 = O_1) P(EP_1 = O_t) \prod_{i=1}^{2^d-1} P(IP_i = ip_i | pa(IP_i)) \end{aligned}$$

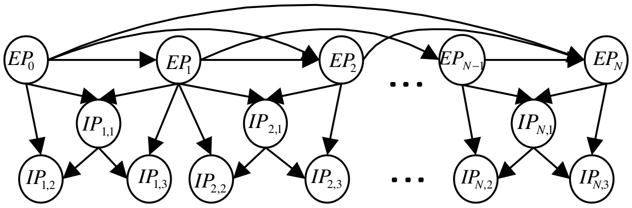


Figure 4. Bayesian network representation of a grapheme model.

3.3 Grapheme (digit or alphabet) model

A grapheme instance consists of strokes which have relationships. Therefore, a grapheme model consists of stroke models with inter-stroke relationships (ISRs).

ISRs are represented with dependencies of stroke end points. Ideally, a stroke gets influence from all the points of other strokes. However, representing all the relationships is too complex and redundant. So, we encapsulate them as relationships of stroke end points. A grapheme model is constructed by concatenating stroke models according to their writing order and specifying ISRs.

Fig. 4 shows a Bayesian network based grapheme model with N strokes and the stroke recursion depth $d = 2$. EP_i's are the stroke end point models and $IP_{i,j}$'s are the internal point models in the i-th stroke. The right end point of the previous stroke is shared with the left one of the following stroke. ISRs are represented by the arcs between EP_i's, and WSRs are represented by the incoming arcs to $IP_{i,j}$'s.

The model likelihood of a grapheme is calculated by enumerating all the possible stroke segmentations. Let's assume that a grapheme model G has N stroke models and a character input has T points O_1^T. Because many different stroke segmentations are possible, let one stroke segmentation instance be denoted as $\gamma = (t_0, t_1, \ldots, t_N)$, $t_0{=}1{<}t_1{<}\ldots{<}t_N{=}T$, and the whole set as Γ. Then the grapheme model likelihood is given as follows:

$$P(O_1^T|G) {=} \sum_{\gamma\in\Gamma} \prod_{i=1}^{N} P(S_i = O_{t_{i-1}}^{t_i} | O_{t_0}, \ldots, O_{t_{i-1}}) \quad (5)$$

4 Extension of Bayesian network framework to Hangul characters

4.1 Scaling- and translation-invariant relationship modeling

Graphemes have scaling and translation variations in Hangul syllable bounding boxes (Fig. 5). The aspect ratios and positions of a grapheme are quite various according to Hangul character classes and writing styles.

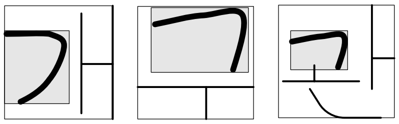


Figure 5. scaling and translation variations of the grapheme ⁄.

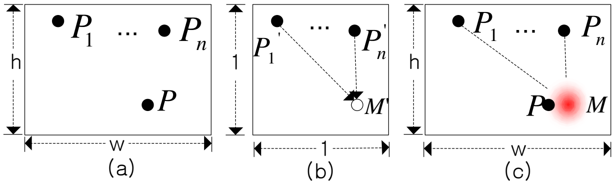


Figure 6. Translation and scaling invariant relationship modeling.

The conditional Gaussian distribution (Eq. (2) (3)) is not robust against these variations. When points with dependencies are transformed geometrically, the mean position predicted from transformed points is different from the transformed position of the mean predicted from the original points.

One simple solution is to normalize coordinates of points with their bounding box. However, this normalization distorts handwriting shapes severely when the aspect ratio is very large or small such as the numeric character 1.

Our solution is to modify the conditional Gaussian distribution (Eq. (2), (3)) so that the mean is predicted in the normalized coordinate space and the covariance is calculated in the input space. Fig. 6 shows this procedure. P depends on $P_1, \ldots, P_n$ (Fig. 6 (a)). P_i''s are normalized points of P_i's with a bounding box. The mean position M' of P is predicted from $P_1', \ldots, P_n'$ in the normalized space (Fig. 6 (b)). Then, M' is mapped into M in the input space. From the covariance and M in the input space, the matching probability is calculated (Fig. 6 (c)). Compared to Eq. (3), the mean is changed as follows:

$$\mu = \mathbf{S}\mathbf{W}(\mathbf{Z}')^T + \mathbf{T} \quad (6)$$

where $\mathbf{S}$ and $\mathbf{T}$ are scaling and translation matrix for the input space, and $\mathbf{Z}'$ is the set of normalized coordinates P_i''s.

4.2 Inter-grapheme relationship modeling

Inter-grapheme relationships (IGRs) are important to identify graphemes. Because graphemes have many strokes and points, IGRs need to be represented abstractly. Bounding boxes are used for abstractly representing grapheme shapes because it easy to calculate them from points and graphemes usually have rectangular shapes.

IGRs are grouped into six types ($T = \{1, \ldots, 6\}$) according to the positions of graphemes (Fig. 7 (a)). So,

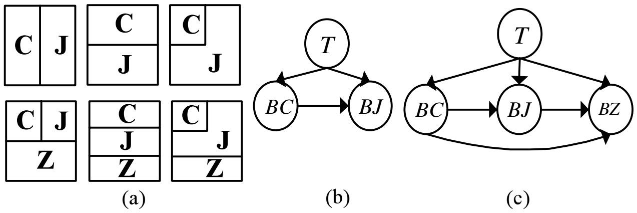


Figure 7. Inter-grapheme relationship model.

the bounding box of a grapheme depends on the type T and the positions of previous graphemes (Fig. 7 (b), (c)). BC, BJ, BZ denote the bounding boxes of C, J and Z.

The matching probability of a grapheme G to a grapheme instance, a sequence of points, O_k^l when its dependent grapheme C is matched to O_i^j, is calculated by combining the IGR probability and the grapheme shape matching probability. Let $B(O_k^l)$ and $B(O_i^j)$ denote the bounding boxes of O_k^l and O_i^j respectively. Then, the matching probability is represented as follows:

$$
\begin{aligned}
& P(G = O_k^l | C = O_i^j) & (7)\\
= \;& P_G(O_k^l, B(O_k^l) | O_i^j) \quad (\because P_G(B(O_k^l)|O_k^l) = 1) & (8)\\
= \;& P_G(B(O_k^l)|O_i^j) \cdot P_G(O_k^l | B(O_k^l), O_i^j) & (9)\\
\approx \;& P_G(B(O_k^l)|B(O_i^j)) \cdot P_G(O_k^l | B(O_k^l)) & (10)
\end{aligned}
$$

In Eq. (8), there is no uncertainty in $B(O_k^l)$ when O_k^l is given. In Eq. (10), points in a grapheme are assumed to depend only on the current bounding box, and the dependency from other grapheme is encapsulated with that of grapheme bounding boxes.

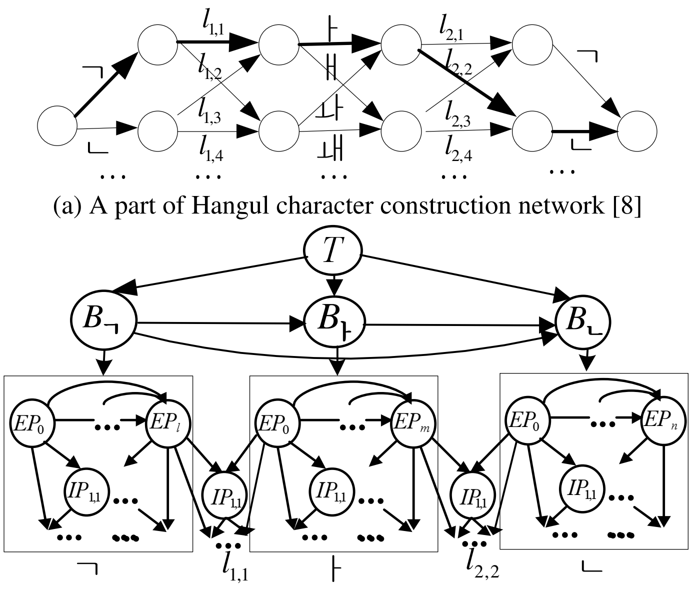
4.3 Bayesian network representation

In terms of practical application, it is infeasible to model 2,350 Hangul characters because of large model complexity and the limited amount of training data. So, Hangul character models should share grapheme and ligature models.

Ligatures are the traces connecting graphemes, which are usually pen-up movements. Their shapes are simple and nearly straight. Therefore, a ligature model is made of one stroke model.

Hangul character models are constructed from a Hangul character construction network [8] (Fig. 8 (a)). The nodes and arcs of the construction network have different meaning from Bayesian networks; a node stores all the possible paths reaching it from the start node, and arcs contain grapheme or ligature models. A path is converted into a Hangul character model by concatenating grapheme and ligature models in its arcs and specifying inter-grapheme relationships. All the paths from the left-most node to the right-most one lead to 2,350 complete Hangul characters.

Fig. 8 (b) shows the Bayesian network model corresponding to a path 'ㄱ-$l_{1,1}$-ㅏ-$l_{2,2}$-ㄴ' (Hangul character 간). T(=4) denotes the type of IGR. B's denote bounding boxes of the first consonant (ㄱ), the vowel (ㅏ) and the last consonant (ㄴ) respectively and have dependencies from T. Their dependencies are modeled as the conditional Gaussian distribution (Eq. (2),(3)). The dependencies from grapheme bounding boxes to grapheme models are modeled by translation- and scaling-invariant conditional Gaussian distributions (Eq. (6)).



(a) A part of Hangul character construction network [8]

(b) Hangul character model corresponding to the path marked by thick arcs in (a)

Figure 8. Hangul character model.

When a Hangul character model $H = C \cdot L_1 \cdot J$ whose type is T matches a point sequence O_1^M, the matching probability is calculated by finding the most probable grapheme segmentation. Let $t_0 = 1 \le t_1 \le t_2 \le t_3 = M$ denote a segment instance of O_1^M such that C is matched to $O_1^{t_1}$, L_1 to $O_{t_1}^{t_2}$, J to $O_{t_2}^{t_3}$. Then the matching probability is as follows:

$$
\begin{aligned}
& P(O_1^M | C \cdot L_1 \cdot J) & (11)\\
= \;& P_C(B(O_1^{t_1})|T) P_{J|C}(B(O_{t_2}^{t_3})|B(O_1^{t_1}), T) & \\
& P_C(O_1^{t_1}|B(O_1^{t_1})) P_{L_1}(O_{t_1}^{t_2}) P_J(O_{t_2}^{t_3}|B(O_{t_2}^{t_3})) &
\end{aligned}
$$

In the case of a Hangul character with the ligature L_2 and the final consonant Z, O_1^M is divided into five segments. Then $P_{L_1}(O_{t_3}^{t_4})$, $P_Z(O_{t_4}^M|B(O_{t_4}^M))$ and $P_{Z|C,J}(B(O_{t_4}^M)|B(O_1^{t_1}), B(O_{t_2}^{t_3}))$ are multiplied to Eq. (11).

5 Experimental results

5.1 Data sets

The data for the experiment were collected from high school and college students. There was no restriction or

Table 1. Properties of the test databases.

Database	*Nation*	*High School*	*Jungang*
Writers	9	9	39
Characters	3,127	15,250	16,427
Writing variations	small	large	medium

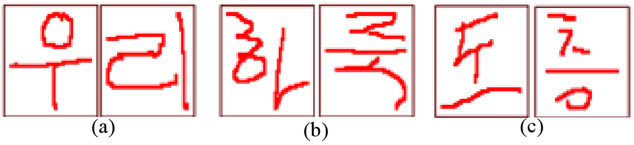


Figure 9. Sample Hangul syllable characters from (a)*Nation* (b) *High School* (c) *Jungang* sets.

guidance in the writing styles. As a result, cursive writing style as well as run-on style were found in the data.

The training data have 49,049 characters written by 48 writers. The test data consists of three sets: *Nation, High School* and *Jungang*. Different writing styles are observed in the data sets as shown in Fig. 9. By subjective judgment, their writing variations are evaluated as shown in Table 1.

5.2 Recognition performance

To evaluate the performance of the proposed system, we recognized Hangul graphemes and syllable characters with the proposed system and compared the results with those of the HMM recognition system. **BN** is the proposed system; WSRs are modeled until the recursion depth three, and ISRs are fully modeled. **HMM** is the discrete HMM based recognizer with chaincode feature [8]. Neural network verifiers are augmented for postprocessing HMM's outputs [1].

For grapheme data, the recognition rate by **BN** is 98.6% on average. On the other hand, that of **HMM** is 91.7% (Table 2). **BN** reduced the recognition errors of vowels (J) much largely because they are more structural than consonants, and inter-grapheme relationships are also stronger.

Because graphemes are more correctly classified, the recognition rates of **BN** on Hangul characters also become higher as shown in Table 3. Its recognition rates are higher across all the data sets. Its average recognition rate is 95.7%, which is higher than 92.9% of **HMM**.

6 Conclusions

In this paper, we propose a Bayesian network framework for explicitly modeling components and their relationships of Korean Hangul characters. A Hangul character is modeled with hierarchical components: a syllable model, grapheme models, stroke models and point models. Each model is constructed with subcomponents and their relationships. A point model is represented by a 2-D Gaussian for point positions on X-Y plane. Relationships are modeled with position dependencies between components.

Compared to the previous Bayesian network model framework for single digits, the proposed model is extended in three aspects. First, the large character classes (2,350) are modeled by sharing grapheme and ligature models. Second, the relationship models are extended invariant to scaling and translation variations of graphemes. Third, inter-grapheme relationships are abstractly represented by position dependencies between bounding boxes of graphemes.

The recognition rates of the proposed system on online handwritten Hangul databases (three sets, 57 writers, 34,804 characters) are higher than those of the HMM system with chain code features. For graphemes, it shows 98.6% on average but the HMM system does 91.7 %. For Hangul characters, it shows 95.7% on average and the HMM system does 92.9%. Its higher recognition rates come from its enhanced modeling capacity of relationships between components.

Table 2. Recognition rates of graphemes.

	C	J	Z	Average
BN	98.31%	98.97%	98.56%	98.61%
HMM	93.19%	86.28%	95.62%	91.70%

Table 3. Recog. rates of Hangul characters.

	Nation	*High School*	*JungAng*	Average
BN	98.27%	94.04%	94.84%	95.72%
HMM	95.59%	91.12%	92.00%	92.90%

References

[1] S. Cho and J. Kim. Verification of graphemes using neural networks in hmm-based online korean handwriting recognition. *Proc. of 7th IWFHR, Amsterdam*, pages 219–228, Sep. 2000.

[2] S. Cho and J. Kim. Bayesian network modeling of strokes and their relationships for on-line handwriting recognition. *Proc. of the sixth ICDAR, Seattle, WA*, pages 86–90, Sep. 2001.

[3] S. Jaeger, S. Manke, and A. Waibel. Npen++: an on-line handwriting recognition system. *Proc. of the Seventh IWFHR, Amsterdam, The Netherlands*, pages 249–260, Sep. 2000.

[4] F. Jensen. *An Introduction to Bayesian Networks*. Springer Verlag, New York, 1996.

[5] K. Murphy. Inference and learning in hybrid bayesian networks. *T.R. 990, U.C.Berkeley, Dept. Comp. Sci*, 1998.

[6] K. Nathan, H. Beigi, J. Subrahmonia, G. Clary, and H. Maruyama. Real-time on-line unconstrained handwriting recognition using statistical methods. *Proc. of IEEE ICASSP, Detroit, USA*, 4:2619–2622, May 1995.

[7] M. Ostendorf. From hmm's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4(5):360–378, Sep. 1996.

[8] B.-K. Sin and J. Kim. Ligature modeling for online cursive script recognition. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 19(6):623–633, Jun. 1997.