A Unified Network-based Approach for Recognition of Cursive Handwritings in Mixed Languages : A Case Study on Hangul and Roman Mixture

Jin H. Kim[1] and Jay J. Lee[1]

Introduction Recent development of pen computing technology enables handwriting to be one of the major modes of human and computer interaction. Although handwriting recognition capability is not sufficiently acceptable for practical use, several systems recognizing Roman characters are available in the market. On the other side of the globe, non-Roman character recognition systems are also under development to respond the pen computing technology. Japanese Kana and Korean Hangul recognition technologies are seemingly reaching a stage of practicality.

As English gains more popularity as a language for describing technologies, intermixed use of English words in their native non-Roman text is inevitable in most of the Asian countries. Figure 1 shows a Hangul text in which English words are embedded. Although, recognition of individual language is somewhat successful, no attempts has been seriously made to develop a recognizer to handle texts written in more than one languages. Since it is unable to handle handwritings in intermixed languages but able to recognize those written in a single language, awkward interface has to be introduced - such as clicking a button before writing - to indicate the language in use.

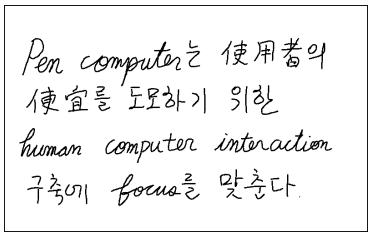


Figure 1: Typical Hangul Text with English words

Several approaches are possible to develope a system that may recognize handwritten text in intermixed languages. The first approach we may think about is to develop individual recognizers for each language and apply them parallel for every handwriting. The recognition trials are evaluated and one of them is selected as the recognition result. This approach may easily utilize individually developed recognizers, but it may not easy to make the recognizers produce recognition scores in a comparable scale. The second approach is to develop a preclassifier which classifies unknown handwritings into a language class by examining a certain set of features. Then an appropriate recognizer does its job. This approach seems efficient if the language class decision is reliable, but it is another difficult task to develop a robust preclassifier. The third approach which we used in this paper is developing a unified recognizer for the entire character set of more than one languages. This approach which obtains the character label as well as the language class at the same time allows smooth integration of the recognition result with language models for postprocessing. However, conquering the different features of different languages and unifying them into one recognizer is not an easy task.

[1]Computer Science Department & Center for AI Research, Korea Advanced Institute of Science and Technology

In this paper, we present a unified approach for recognizing freely handwritten text in more than one languages. We do not assume or put any constraint on writing styles. It can be either hand-printed or cursively connected by pen dragging between strokes. The approach is based on the network of hidden Markov models(HMM) producing a probability measure as its score of recognition [1]. Since the approach is based on HMM, it is basically stochastic. This means the approach is so general that any language combinations can be handled with proper model training. Since it is based on transition network, various knowledge such as language models and structural constraints can be easily augmented.

Network-based approach for handwritten word recognition Not only pen-down but also pen-up movements are coded into a chain of directional codes which is to be decoded as a sequence of characters with intermediate ligatures. Hidden Markov modelling technique is adopted to construct the character and ligature models. Viewing handwritten word as an alternating sequence of characters and ligatures, words in a language is modeled by a network of interconnected character and ligature models. For instance, Hangul syllables are modeled by a network called BongNet [2] and English words of indefinite length is modeled by circularly interconnected models of character and ligature [3]. Figure 2 shows the BongNet for Hangul syllable and Figure 3 shows the circular network for English word model.

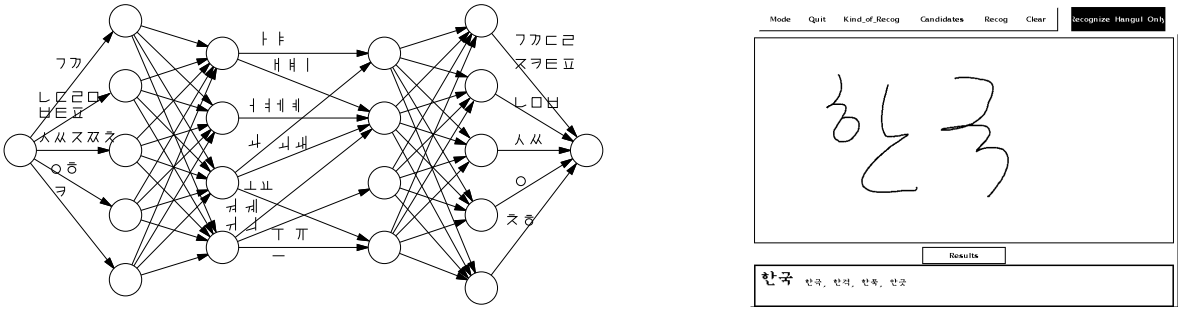
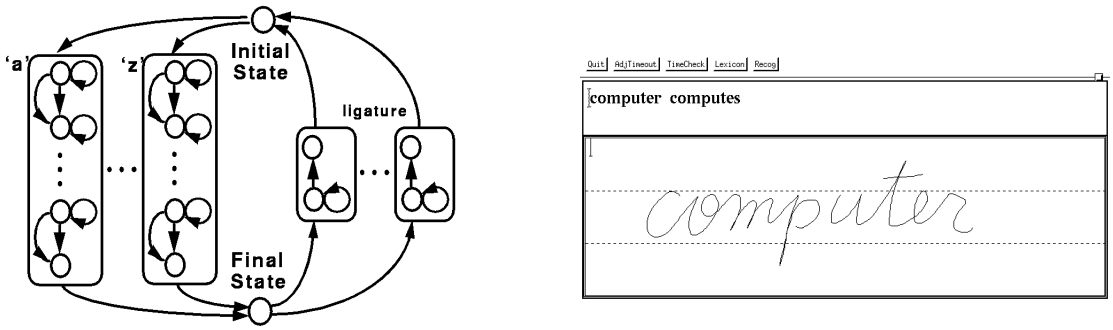


Figure 2: (a) Hangul Syllable Network: BongNet (b) Recognition Example



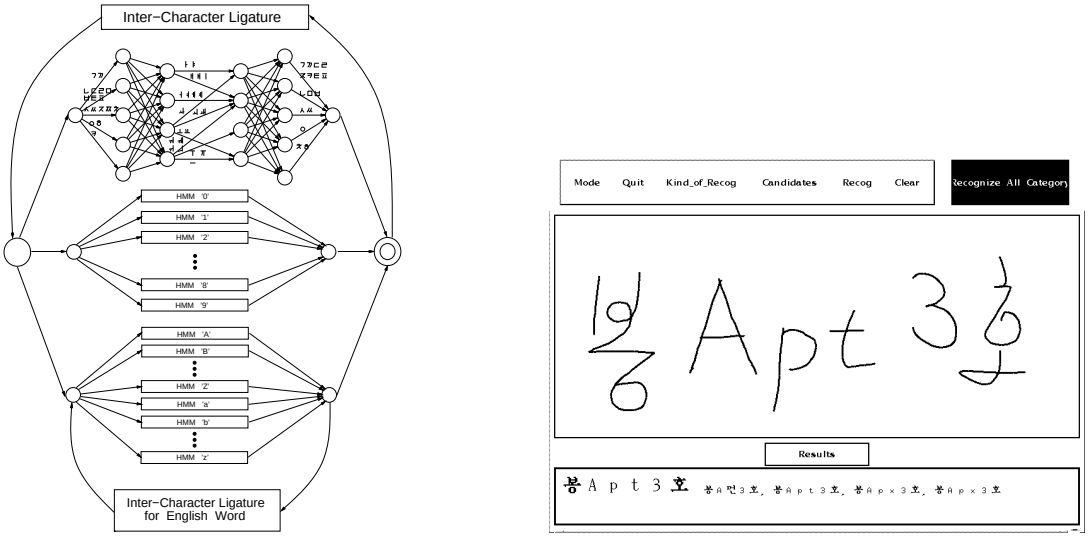
(a) (b)

Figure 3: (a) Interconnected HMMs for English Word Modelling (b) Recognition Example

Figure 4: Hangul: the most scientific writing system

Here a brief introduction should be given to Hangul, which is invented by the King Sejong about five hundred years ago. Hangul consists of 24 phonetic symbols: 10 vowels and 14 consonants. Each 'block' of Hangul is formed by combining phonetic symbols to represent single spoken syllable. The example in Figure 4 shows the word for 'Korea'(HAN-KUK). The numbers indicate the order of pronunciation of each phonene within each syllable block, which is the same as the order of normal penmanship writing. Note that the vowels always appear in position #2. Combining the best features of an alphabet and a syllabary, Hangul is the most scientific writing system among those ever actively used.

The first layer of arcs in BongNet models first consonents in Hangul syllable and the second layer represents ligatures between first consonents and vowels. The third layer represents vowels, and so on. In such word networks above, a path spanning from the initial node to the terminal node, possibly with a number of cycles on the way, represents a word. The recognition problem is then formulated into that of finding the most likely path from the initial node to the terminal node, consuming all the input sequence of directional codes. From the maximal probability path which consists of character and ligature models, optimal character and ligature segmentation and associated character labels are obtained simultaneously.



(a) (b)

Figure 5: Unified Network for Intermixed Use of Roman, Digit and Hangul Syllable

Since such network produces a probability associated with a path, more than one networks can be easily combined under the probability framework. For example, BongNet which models Hangul syllables and the interconnected HMMs which model English words are combined to recognize intermixed use of the two languages. The network for digit recognition can be easily added. As a result, a unified network is constructed for recognition of intermixed use of Hangul, English and digits as shown in Figure 5-(a).

Ligatures among Hangul syllables, English words and digits, which are mostly pen-up movement, are modeled as a circular path from the global terminal node to the global initial node. As done in individual networks for single language recognition, paths in the unified network from the global initial node to the global terminal node are competing for the label of the input sequence of directional codes.

Various kinds of knowledge can be easily integrated with such network based approach. Knowledge from language models is integrated by checking bigram and trigram at the entrance of each character HMM. Structural constraints, either within character or intercharacter, are also confirmed when exiting from each character HMM. Subtle differences such as the difference in printed 'h' and 'n' is emphasized by a pairwise discrimination at the end of the character HMMs.

In order to obtain acceptable recognition accuracy, several mechanisms are incorporated. These include a mechanism to handle delayed strokes in writing English words, pairwise discrimination between some digits and Roman alphabets. An efficient search method is also devised to exploit the regularity of Hangul construction rule.

Conclusions We have proposed a unified network based approach for recognizing freely handwritten text in more than one languages. We believe this approach can be applied to any combination of phonetic writing systems including Arabic, Tai and Japanese. Although intensive evaluation yet to come, initial implementation for Hangul and Roman mixture with digits yields promising results which cannot be easily surpassed by other approaches. By combining component languages, recognition accuracy drops negligibly little, but speed is slowed substantially. More powerful search methods and machines are in demand to use such approach in practice.

Acknowledgement This research is supported by Korea Science and Engineering Foundation through the center of AI Research program and the Notepad consortium formed by DaeWoo Communication, Hyundai Electronics, Korea Computer, PosData, Samsung Electronics, and Trigem Computer, Inc.

Reference

[1] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of the IEEE*, v.77 ,n.2, pp.257–286, 1989.

[2] B. K. Sin, J. H. Kim, "A Statistical Approach with HMMs for On-line Cursive Hangul (Korean Script) Recognition," *Proceedings of the Second International Conference on Document Analysis and Recognition, Zukuba, Japan*, pp147–150, Oct. 1993.

[3] J. Y. Ha, S. C. Oh, J. H. Kim, Y. B. Kwon, "Unconstrained Handwritten Word Recognition with Interconnected Hidden Markov Models," *Third International Workshop on Frontiers in Handwriting Recognition, Buffalo New York, USA*, pp 455–460, Sep. 1991.