Problems and Approaches for Oriental Document Analysis

Jin H. Kim
Department of Computer Science, KAIST
373-1 Kusong-dong Yusong-gu, Taejon 305-701, South Korea
Fax: +82-42-869-3510, E-mail: jkim@ai.kaist.ac.kr

Machine understanding of handwritten documents in China, Japan, and Korea requires not only general solutions of document analysis but also ability to handle peculiarities of the Oriental languages. As expected, handwritten Chinese character recognition is the major task for it. In addition, Japanese Kana, Korean Hangul, Roman alphabet as well as numerals are targets of recognition.

The main difficulties of Oriental character recognition originate from their large character sets. A practical system should be able to handle at least 5000 classes from possibly 50000 over classes. For Hangul, 11720 classes can be made in theory. The difficulty closely depends on writing styles. Oriental script is generally classified into regular, fluent, cursive style. Needless to say, cursive style is deformed most seriously and, therefore, most difficult to recognize. Regular style writing is often attacked successfully by feature matching and statistical analysis, while fluent style is now actively under investigation by stroke analyses and structural matching. Cursive style recognition is seldom found even in research papers.

Since Chinese and Hangul characters are intrinsically hierarchical, often hierarchical analysis has been applied. A Hangul character, which corresponds a syllable, is formed by 2 to 5 basic graphemes, drawn from 24 classes, deploying them in two dimensional way. Recognizing component graphemes is, we believe, the viable approach to handle the large set of Hangul classes. Therefore, segmentation into graphemes, which is another difficult task, is the key for hierarchical recognition.

For robust recognition of fluent to cursive style, the following research directions are suggested.

1. Robust preprocessing / Stroke extraction / Character segmentation
   - Reliable stroke extraction knowing that the images are handwritten characters
   - Segmentation into characters from cursive writing
3. Hierarchical description and matching
4. Training from sample data
5. Linguistic post-processing for similar shape discrimination.
6. Multi-lingual document analysis