2U-4

A Proposed Multi-Feature Extraction Method for Khmer OCR

Vanna KRUY vanna@ruri.waseda.jp
Wataru KAMEYAMA wataru@waseda.jp
Graduate School of Global Information and Telecommunication Studies, WASEDA University

1. INTRODUCTION

OCR technology has already matured. However, there is no reliable OCR system for Khmer Language. This is largely due to the lack of Khmer OCR research efforts and the complex nature of Khmer characters. Most of Khmer OCR researches focus on single feature with limited accuracy. We propose a method based on multi-features using Scale Invariant Feature Transform and Fourier Transform Descriptors. Our experiment has proved the accuracy of the proposed method as of 97.4%.

2. RELATED WORKS

There are only few research efforts on Khmer OCR. Two of that are Khmer OCR using Wavelet Descriptors and Khmer Printed Optical Recognition Using Lagendre Moment both by Chey et al which reaches 92.99% accuracy [1] and 92% accuracy [2], respectively. Ing L.I. experimented Khmer OCR for Limon R1 font, size 22 using Discrete Cosine Transform and Hidden Markov Model which reaches 98.88% accuracy [3]. [1] & [2] is experimented with several Khmer fonts, but get lower accuracies. [3] gets high accuracy but was limited to a fixed fontset and a fixed size.

3. PROPOSED METHOD

We would like the system to be able to deal with several fontsets and achieve high accuracy. We have chosen a multi-feature approach. We use Scale Invariant Feature Transform (SIFT) [4] and Fourier Descriptors [5] as the features.

3.1 SYSTEM OVERVIEW

The system is divided into two main modules—Training Module and Recognition Module. Each module depends on other sub-modules. Fig.1 shows all the modules in the system. The upper layers depend on the lower layers. Training Module is used to annotate Connected Component (CC) and extracts the character assembling rules. It relies on Rule Extraction, Connected Component Analysis (CCA), SIFT extraction, Fourier Descriptors (FD) extraction, Rule File and Annotated CCs Database modules. Recognition Module is used to test the system. It depends on Character Assembler (CA), Connected Component Recognition (CCR), Vertical Component Extraction, CCA, SIFT Extraction, FD Extraction, Rule File and Annotated CCs Database modules. Section 3.2 and 3.3 give detailed information about these processes.

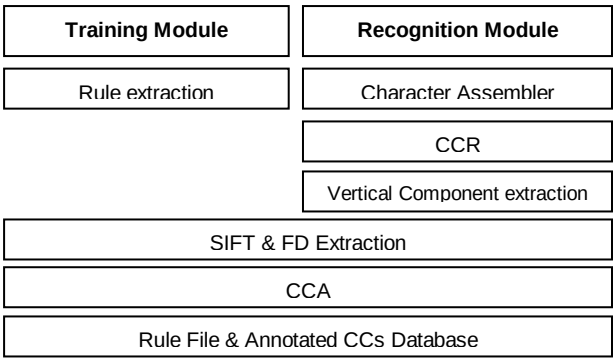

Fig.1 System Overview

3.2 TRAINING PROCESS

3.2.1 RULE EXTRACTION

Since several Khmer characters disjoint into parts. There is a need to assemble them back into a complete character. We use rule-based approach for the assembly. For every disjointed character, we extract the connected components and annotate them.

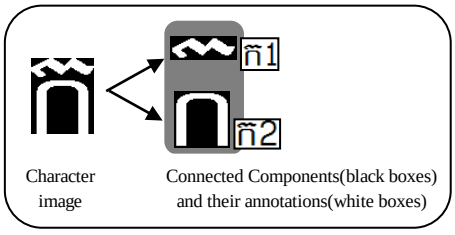

Fig.2 The Annotation Process

The annotated labels and their corresponding character label are then added to the Rule File (in flat text format) as shown in Fg.3. The left side of "=" sign are the connected components' labels, while the right side is the assembled character label. It will be used later by the Character Assembling module. See section 3.3.4.

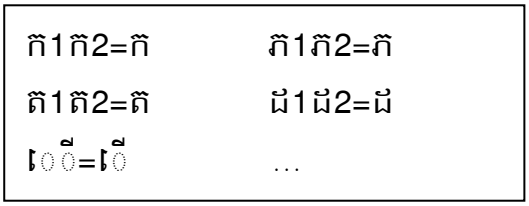

Fig.3 Character Assemble Rule File

3.3 RECOGNITION PROCESS

We have used CCA to segment CCs. To recognize CCs, we use SIFT and FD descriptor as the features. Finally we use Character Assembler to joint any disconnected character's parts.

Section 3.3.1 and 3.3.2 describes the process of SIFT extraction and FD extraction, respectively. Section 3.3.3 and 3.3.4 explains the fusion of SIFT and FD features, and the character assembling process.

3.3.1 SIFT SCORE

To compare the extracted CC with CCs in the database, we calculate the SIFT score with the following formula. SIFT parameters are shown in Fig.4.

$$score = \frac{2mp}{c_1kp + c_2kp} \quad (1)$$

mp: number of matched key points
$c_i$kp: number of key points in component i

UP_SCALE = true	STEP_PER_SCALE = 6
MIN_SIZE = 20	MAX_SIZE = 1024
FEATURE_DESCRIPTOR_SIZE = 4	INITIAL_SIGMA = 1.6
FEATURE_DESCRIPTOR_ORIENTATION_BINS = 8	

Fig.4 SIFT Parameters

3.3.2 FOURIER DESCRIPTORS

FD has been shown to be effective in shape discrimination [5]. It has translate, scale, and rotation invariant properties. To extract FDs, an appropriate shape signature needs to be obtained which in our case is the Centroid Distance. First we extract the contour from the input image, normalize it, and transform it to temporal domain using Centroid Distance. The fourier transform of the Centroid Distance is the FD. To compare two images using FD, Fourier distance is calculated using the least square distance.

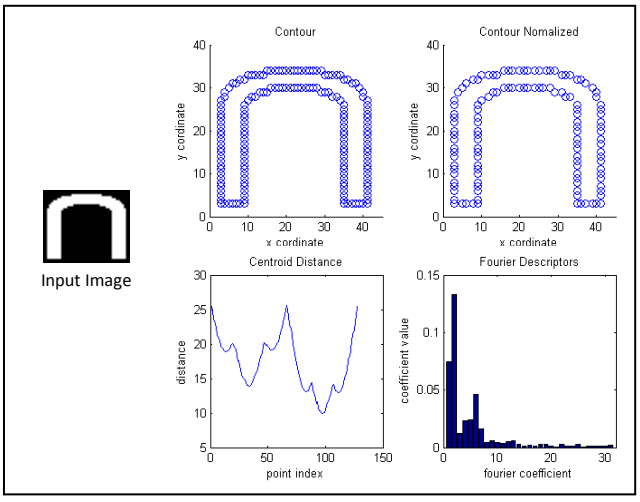

Fig.5 FD Extraction Process

3.3.3 FEATURES FUSION

To combine the SIFT score and Fourier distance, we use the following equation. For SIFT score, the higher value means the more similar the shapes are. In contrast, the lower the Fourier distance value, the more similar, as it is for Fused distance.

$$\text{Fused distance} = \frac{1 - \text{SIFT score} + \text{Fourier distance}}{2} \quad (2)$$

3.3.4 CHARACTER ASSEMBLING

During the CCR process, the annotation of the CC is added to the result string. This result string is not yet final since disjoint character parts are not yet combined. To assemble them back to their proper characters, we used the Rule File. The returned result string is searched if it matches any entry on the Rule File. Any matches found are replaced by the corresponding character. See Fig 6.

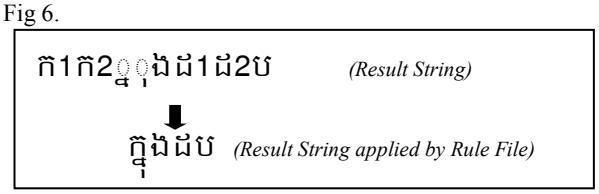

Fig.6 Character Assembling

4. EXPERIMENT SETTING AND RESULT

We have tested 2 documents taken from newspaper with 1104 words. The CC database contains 907 CCs. Five fontset samples are used in testing and database: Khmer OS System, Khmer OS Bokor, Khmer OS Freehand, Khmer OS Muol, and Khmer OS Muol Pali.

Table.1 Document Test Result

Precision	0.974
Recall	0.974
F-Measure	0.974

4.1 DISCUSSION AND FUTURE WORK

The method achieves high accuracy as we assume noise-less document image. We would like to know how the system performs in noisy document. This will be done in our future work. We would also like to test all the Khmer fonts currently in use. Moreover, we believe that features like holes and component positions would increase the system accuracy.

5. CONCLUSION

The proposed system outperforms several previous works. SIFT and Fourier Descriptors prove themselves as good features for Khmer OCR. Although our system performs slightly lower than the state-of-the-art. It has the advantage of being able to deal with multiple fonts.

REFERENCES

[1] C. CHEY, P. KUMHOM, and K. CHAMNONGTHAI. *Khmer Printed Character Recognition by using Wavelet Descriptors.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems Vol.14 NO.3. (2006) 337-350. Word Scientific Publishing Company.

[2] C. CHEY, P. KUMHOM, and K. CHAMNONGTHAI. *Khmer Printed Characters Recognition Using Lagendre Moment Descriptor*. Department Electronic and Telecommunication, King Mongkut's University of Technology Thomburi.

[3 I. LENG IENG, K. SOCHENDA and C. SOKHOUR , *Khmer OCR for Limon R1 Size 22 Report*, PAN Localization Cambodia (PLC) of IDRC. (2009). URL:http://www.panl10n.net/english/Outputs%20Phase%202/CCs/Cambodia/MoEYS/Papers/2009/KhmerOCRLimonR122.pdf visited: July 14th, 2010.

[4] D. G. LOWE, *Distinctive Image Features from Sclae-Invariant Keypoints*, International Journal of Computer Vision, 60, 2, pp. 91-110, 2004.

[5] E. PERSOONn and K. SUN FU, Shape Discrimination Using Fourier Descriptors. IEEE. 1976.