Perspectives for the Historical Information Retrieval with Digitized Japanese Classical Manuscripts

Hari Raghavacharya*1, Rafal Rzepka*2, Kenji Araki*2 and Toshimasa Miyazawa*1

*1Graduate School of Letters, Hokkaido University
*2Graduate School of Information Science and Technology, Hokkaido University

This paper outlines a system to retrieve and convert text matter to digital form from old manuscripts written in Classical Kana Characters.The system consists of five stages 1.Preparation of sample kana handwriting 2.Primary classification into sub categories 3.The Segmentation of Target Classical Manuscript 4.Feature Extraction 5.Recognition.The detailed aspect of preliminary stages like sample collection and primary classification will entail greater accuracy in recognition of a wide range of Kana manuscripts.

1. Introduction

Character recognition has come long way since it was conceived as an idea. Greater accuracy has become possible in extracting printed matter to a text form. Its potential is recognized widely for digitalizing the contents of text printed in an earlier era. However, handwriting recognition remains in its infancy despite its huge promise. There are several systems in existence for handwriting recognition of Latin based languages[1][2]. Some intensive research is seen in CJK version of handwriting recognition as well[3][4][5]. There are no existing systems for recognition of classical Japanese Kana. There is a large corpus of manuscripts written in Kana that includes some very famous works. Most of the famous works have been rendered into modern Japanese by going over entire manuscripts letter by letter. This process has covered a minuscule amount of texts and is not likely to take a vast majority of others unless some tool is made available to make the work simpler.

In this paper, we propose a recognition system for classical Kana characters. The proposed process can be divided into 3 stages. The first stage covers collection of sample handwriting spread over a few centuries. The second stage covers conversion of these samples into binaries and classifying them according to their sizes as a part of primary classification. In the final stage, an algorithm is used to recognize an unknown character as one of the primary classification groups.

The proposed system is being worked on as an interdisciplinary joint project.

This paper is organized in following sections. Section 2 deals with the background of Kana system of writing. Section 3 introduces the proposed system while section 4 deals with the possibilities of the proposed system. We present our conclusions in section 5 and discuss future work.

Contact: Language Media Laboratory, Research Group of Information Media Science and Technology, Division of Media and Network Technologies, Graduate School of Information Science and Technology, Hokkaido University, Kita-ku Kita 14 Nishi 9, 060-0814 Sapporo, Japan. TEL: (+81)(11)706-6535, FAX: (+81)(11)706-6277, {hari,kabura,araki}@media.eng.hokudai.ac.jp

2. Kana writing system

The modern Japanese writing system consists of three elements, *Hiragana* and *Katakana* syllabary and the Chinese ideograms called *Kanji*. It is typical for a sentence to have all the three elements, although the usage of *Katakana* is restricted to proper nouns of foreign origin.

Kana writing system is derived from Chinese ideograms that were used in early Nara period as phonetic symbols for Japanese syllabic sounds. These were called as *Manyogana* and had no meaning apart from the sound association. Handwritten *Manyogana* were abbreviated and simplified in actual usage, which gave rise to a new system of writing called *Sogana*. This *Sogana* or Kana system of writing remained popular throughout Japanese history till it was replaced by *Hiragana* syllabary during Meiji period. In Kana system, Kana characters derived from multiple Chinese ideograms coexisted until Hiragana syllabary fixed one Kana for one syllable. However, for a manuscript of a preceding era, there are no fixed number of Kana syllabary but the most commonly used Kana characters' number could be estimated around 200.

3. Ideas for Digital Transcription of Classical Kana Manuscripts

3.1 A typical Manuscript

Classical manuscripts have problems such as deteriorating paper, discoloration and worm holes. We use photographs of the original manuscript scanned at 300dpi, distortions and noise are removed using thresholding techniques[6]. Some manuscripts are written on papers with exquisite patterns which when photographed show as prominently as the text written with brush,making the reading of a particular portion inordinately difficult. The background of such a manuscript image needs to be converted to white to make the text legible. Most of the commonly available manuscripts, however, are on plain form of paper. There are some existing techniques for handling the discoloration and worm hole problems[7].

3.2 Problems in Digital Transcription of Classical Kana Manuscripts

Most of the classical texts have annotations and revisions and reviews scribbled alongside the main text matter.Corrections by the author exist side by side with notes of a reviewer which makes it extremely difficult to identify the original text.

A classical manuscript composed in Kana does not mean that it is entirely written in just one system of syllabary. There are frequent occurrences of Chinese ideograms or Kanji which require a separate handling apart from Kana when it comes to digital transcription of the documents.

3.3 The proposed method

The proposed method consists of five different stages. The first one is Pre-processing that includes preparation of sample handwriting of different people spread over a few centuries. A primary classification of samples into select sub categories based on average occurrence of a letter with reference to provided reference lines.

The next stage deals with the segmentation of the Target text followed by feature extraction. The final stage is recognition.

3.31 Preparation of sample Kana handwritingIn order to make this experimental system work on a wide range of Kana texts, a corpus of Kana handwriting samples will be built. Beginning with sample handwriting from the early Kana literary texts, a wide range of handwriting samples will be collected spread over a few centuries. All forms and variations of one Kana will be categorized into one Kana class

3.32 Digital Transcription of Classical Kana Manuscripts Each sample will be scanned at a specified resolution and segmented.In the experimental stage the resolution of 300 dpi is considered suitable. Segmentation size of each sample is proposed at 32x32 pixels. The samples thus rendered will be stored as 8 bit gray-scale images. Each image is then further broken down to 4x4 pixel zones for which the pixel density is calculated. An initial feature vector extraction captures the pixel density of each zone for each of the images.

3.33 Primary classification into sub categoriesThree vertical reference lines parallel to each other are drawn for every column of Kana writing. The position of each Kana is then calculated with reference to these lines. For each column of Kana, the position of central reference line is estimated upon the measurement of the width of the first Kana character of that column, so that it occurs precisely at the center. The Right and Left reference lines are then estimated after the calculation of the average character width of that particular column.

The Kana characters tend to occur in a set pattern with reference to the first character that is written at the top of the column. There is a delicate system of character balance, the equilibrium of which is maintained throughout the document. The position of individual Kana with reference to the three lines can be said to be almost constant. Each Kana can be placed within a pre-classified group based on its where it occurs in reference to the three lines. The pre-classified groups can be determined based on the four parameters.

(1) The Kana character touches only the Central Reference line

(1) The Kana character touches or protrudes all three reference lines

(1) The Kana character protrudes RRL but does not touch the LRL

(1) The Kana character protrudes LRL but does not touch RRL

3.34 The Segmentation of Target Classical ManuscriptThe system of Kana writing seen in classical works is entirely vertical. Since all of them are composed with brush, there is a tendency to write without lifting the brush till the ink runs dry. This results in a set of Kana characters that are connected without any demarcation among individual letters. This causes some problems in segmenting the text along the traditional methods.

To overcome this problem, it is proposed that all joined Kana characters are treated as one unit apart from the separate Kana characters which constitute a single unit. Each such unit is referred to an N-gram database to check the existence of a particular combination of characters to suit the context or alternate segmenting positions are considered till the appropriate combination is rendered[10].

3.35 Feature ExtractionOnce the Target Text is segmented, feature vector is extracted for each of the segmented area. The pixel density is calculated in each zone where each zone is identical to the zone size in the sample images already prepared.

3.36 RecognitionRecognition of Kana characters is made by an algorithm to identify similarity.The positon of Kana character with reference to the Reference lines determine which subcategory of Kana the unknown character belongs to. Inside the sub category, to which Kana class the unknown character belongs to is determined on the basis of most number of matches. The three best matches are then processed using N-gram database to identify the most appropriate character.

4. Possibilities

The proposed system can contribute to accurate digitalization of classical Japanese manuscripts.This in turn opens up new vistas for Japanologists and Japanese Philologists. A classical manuscript in Kana is useful as resource to Japanologists who are trained to read the Classical Japanese Language and Kana. However, digitally transcripted texts can be made available easily for use by Japanologists improving prospects for new research.

Japanese Philologists can re-look at Japanese language and grammar synchronically.Different copies of the same manuscript exist simultaneously in most cases making it necessary to determine the text genealogy, especially in cases

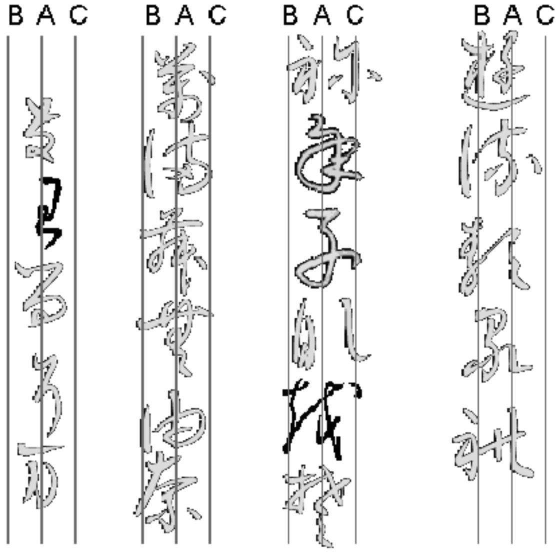

Figure 1: The Primary Classification based on three reference lines

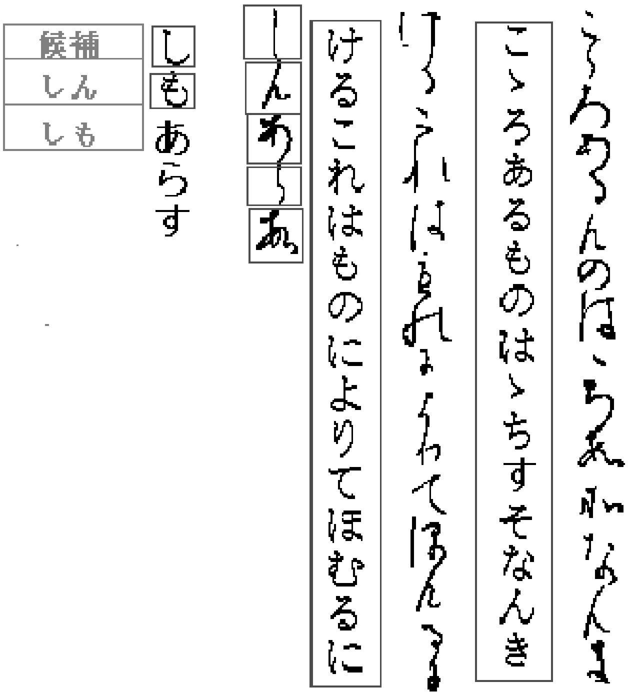
こゝろあるものはゝちすそなんき
けるこれはものによりてほむるに
しもあらす
候補
しん
しも

Figure 2: A hypothetical image showing the recognition process

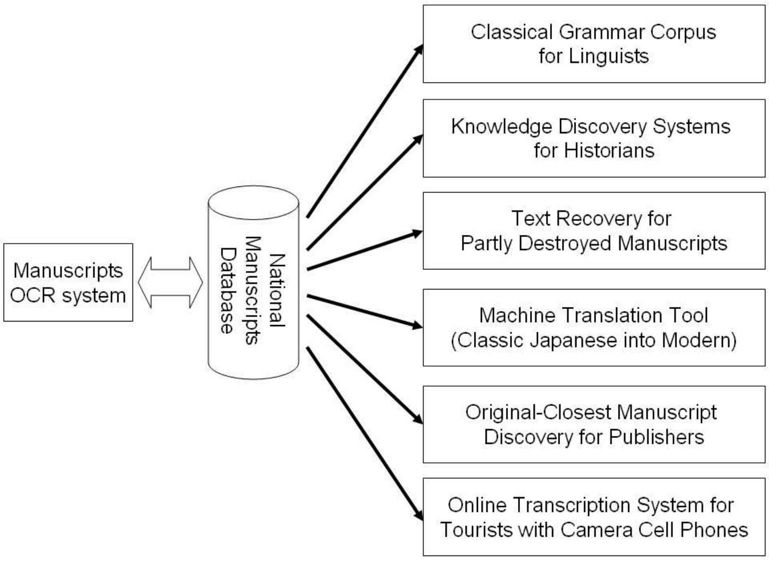
Manuscripts OCR system
National Manuscripts Database
Classical Grammar Corpus for Linguists
Knowledge Discovery Systems for Historians
Text Recovery for Partly Destroyed Manuscripts
Machine Translation Tool (Classic Japanese into Modern)
Original-Closest Manuscript Discovery for Publishers
Online Transcription System for Tourists with Camera Cell Phones

Figure 3: Possibilities for IR of historical manuscripts

where the ‘original‘ manuscript does not exist. Comparison of different versions become simpler and the differences among the text can be demonstrated instantly. Using the methods of textual genealogy, the manuscript that is closest to the ‘original‘ in content can be determined, a process that takes enormous time visually going over each text can be accomplished in a reasonable amount of time with accuracy.

5. CONCLUSIONS

Digitally transcripted manuscripts of Classical Japanese Kana available openly is likely to spur a renewd interest in Classical Japanese language and Literature.Established interpretations of Classical works will be revisited with the possibility of many new ones emerging. For many Japanologists of Pre-modern studies the world over, it will eliminate the need for visiting Japan for consulting first hand resources.

References

[1] R.M.Bozinovic and S.N.Srihari, **Off-line cursive script word recognition**, IEEE Trans. on Pattern Anal.Mach.Intell.,vol.11,no.1,pp.68 83, Jan.1989

[2] Hu, M.K.Brown and W.Turin, **HMM based online handwriting recognition** IEEE Trans. on Pattern Anal.Mach.Intell.,vol.18,no.10,1039 1045, Oct.1996

[3] D.Deng, K.P.Chan and Y.Yu, **Handwritten Chinese Character Recognition using spatial Gabor filters and self-organizing feature maps**, Proc.IEEE Inter.Confer.on Image Processing,vol.3,pp. 940 944, Austin TX, June 1994

[4] C.H Chang, **Simulated annealing clustering of Chinese words for contextual recognition**, Pattern Recognition Letters,vol.17,no.1,pp.57 66,1996

[5] H.Yamada, K.Yamamoto, and T.Saito,**A non-linear normalization method for handprinted Kanji character recognition-line density equalization**, Pattern Recognition Letters,vol.23,no.9,pp.1023 1029, 1990

[6] G.Leedham, S.Varma, A.Patankar,V.Govindaraju, **Separating Text Background in Degraded Document Images - A Comparison of Global Thresholding Techniques for Multi-Stage Thresholding** Proceedings Eighth International on Frontiers of Handwriting Recognition,pp.244 249, September 2002

[7] S.Hara, **OCR for Classical CJK Texts - Preliminary Examination** Digital Libraries and Digital Collection in the Global Community, Pacific Neighborhood Consortium 2004

[8] N.Otsu, **A threshold selection method from grey level histogram** IEEE Trans. Syst. Man Cybern., vol. 9 no. 1, pp. 62-66,1979

[9] R.C.Gonzalez, R.E.Woods, Digital Image Processing, Addison Wesley publishers, 1993.

[10] J. J. Hull and S. N. Srihari, **Experiments in text recognition with binary n-gram and Viterbi algorithms** IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-4(5):520-530. September 1982