A Structural Analysis Based Feature Extraction Method for OCR System For Myanmar Printed Document Images

Htwe Pa Pa Win, University of Computer Studies-Yangon, Myanmar
Phyo Thu Thu Khine, University of Computer Studies-Yangon, Myanmar
Khin Nwe Ni Tun, University of Computer Studies-Yangon, Myanmar

ABSTRACT

This paper proposes a new feature extraction method for off-line recognition of Myanmar printed documents. One of the most important factors to achieve high recognition performance in Optical Character Recognition (OCR) system is the selection of the feature extraction methods. Different types of existing OCR systems used various feature extraction methods because of the diversity of the scripts’ natures. One major contribution of the work in this paper is the design of logically rigorous coding based features. To show the effectiveness of the proposed method, this paper assumed the documents are successfully segmented into characters and extracted features from these isolated Myanmar characters. These features are extracted using structural analysis of the Myanmar scripts. The experimental results have been carried out using the Support Vector Machine (SVM) classifier and compare the previous proposed feature extraction method.

Keyword: Feature Extraction, Myanmar, Optical Character Recognition (OCR) System, Structural Analysis, Support Vector Machine (SVM)

1. INTRODUCTION

Optical Character Recognition, OCR system plays an important role in Document Image Analysis System. The reason for this is that the primary goal of document analysis and recognition is to transform a paper document into a digital file with as less information loss as possible. To get the high performance of the OCR system, selection of the effective feature extraction methods is one of the most important facts that need to be done.

Features extraction is one of the most important factors in achieving high recognition performance in character recognition systems. The extracted features must be invariant to the expected distortions and variations that character may have in a specific application. Also, the number of features must be kept reasonably
small if a statistical classifier is to be used (curse of dimensionality) (Duda, Hart, & Stork, 2001).

A feature extraction method that proves to be successful in one application domain may turn out not to be very useful in another domain. Furthermore, the type of features extracted must match the requirements of the chosen classifier.

There are both many built-in methods and various invention techniques for Feature Extraction of the different OCR systems. All of these methods can be categorized into two types of features: statistical, derived from statistical distribution of points and structural. The most common statistical features used for character representation are: zoning, projections and crossings and distances. Structural features are based on topological and geometrical properties of the character (Vamvakas, Gatos, & Perantonis, 2009). Furthermore, global transformations techniques such as Gabor and Hough Transformations are now becoming popular in some recognition systems. All different types of feature extraction methods are surveyed in Trier, Jain, and Taxt (1996) for the previous century and illustrated that the best method for each application domain can’t be the best for all applications.

The successful usage of statistical features are described in Rajashekararadhya and Ranjan (2005) and Kumar (2010) the former used 50 features and the later used considerable large amounts of features. For the printed Gurmukhi Script (Jindal, Sharma, & Sharma, 2008) and historical documents (Vamvakas, Gatos, & Perantonis, 2009) structural features are used depending on the nature of scripts and by subdividing the character image. They also show their best results. Features from Gabor filters are used mostly in scrip identification in bilingual documents and in Ramanathan, Ponmathavan, Thaneshwaran, Nair, Valliappan, and Soman (2009), Borji and Hamidi (2007), and Ramanathan Nair, Thaneshwaran, Ponmathavan, Valliappan, and Soman (2009), they also selected for their OCR systems.

When analyzing the previous literatures for Myanmar scripts, the OCR system for printed documents Swe and Tin (2006) used all the pixel values as the features. The group of Myanmar Intelligent Character Recognition (MICR) Thein and Yee (2010) used combination of Statistical and usual Structural features such as number of pixels, stroke counts, number of loop and open direction and can only show the good accuracy for normal alphabets and not yet done for all the compound words. The rule based feature extraction method is used in Than, Aung, Yi, and Win (2006) and they only showed the five alphabets of handwritten characters. To the best of our knowledge, the extensive researches for OCR system for Myanmar scripts are needed to be done. The reason for this may be that there are a large number of characters in Myanmar script and because of the complex nature of this script. In Win, Khine, and Tun (2011b) zoning and projection profile based feature extraction method for printed documents is proposed, but this method can correctly recognize only the trained data such as trained font types and sizes. Although, in the machine printed documents, shape discrepancy among characters belonging to same class is sometimes quite large because of the degradations of the document images. Therefore, it is required to select features, which can adapt the shape variations due to touching noise blobs. In fact the main problem in OCR system is the large variation in shapes within a class of character (Borji & Hamidi, 2007). Moreover, large amount of trained data can prevent the high performance of the OCR system. Hence, in South-East Asian Scripts, including Myanmar, one-stage discrimination does not generally suffice and two-stage classification (coarse and fine) should use. The aim of coarse classification is to cluster similar-looking characters into groups and then perform fine classification to extract the right class (Agrawal & Doermann, 2008; Kato, Suzuki, Omachi, Aso, & Nemoto, 1999). Therefore, in this paper, the characters are logically clustered and propose a new feature extraction method using structural features depending on the nature of writing style of the Myanmar Scripts in order to get the more accuracy of the OCR system.

One major contribution of the work presented in this paper is the design of mathematically...
Related Content

A Reflexion on Implementation Version for Active Appearance Model
[www.irma-international.org/article/a-reflexion-on-implementation-version-for-active-appearance-model/95968/](www.irma-international.org/article/a-reflexion-on-implementation-version-for-active-appearance-model/95968/)

Introduction to 3D Imaging
[www.irma-international.org/chapter/introduction-imaging/60257/](www.irma-international.org/chapter/introduction-imaging/60257/)

On Applying the Farey Sequence for Shape Representation in Z2
[www.irma-international.org/chapter/applying-farey-sequence-shape-representation/65059/](www.irma-international.org/chapter/applying-farey-sequence-shape-representation/65059/)

Image Compression Technique for Low Bit Rate Transmission
[www.irma-international.org/chapter/image-compression-technique-low-bit/77042/](www.irma-international.org/chapter/image-compression-technique-low-bit/77042/)

Stereoscopic Vision for Off-Road Intelligent Vehicles
[www.irma-international.org/chapter/stereoscopic-vision-off-road-intelligent/60270/](www.irma-international.org/chapter/stereoscopic-vision-off-road-intelligent/60270/)