OCR Error Correction for Unconstrained Vietnamese Handwritten Text

Quoc-Dung Nguyen
Van Lang University
Ho Chi Minh, Vietnam
Technical University of Ostrava
Ostrava-Poruba, Czech Republic
dungnq.vtrd@gmail.com

Duc-Anh Le
Center for Open Data in the Humanities
Tokyo, Japan
leducanh841988@gmail.com

Ivan Zelinka
Technical University of Ostrava
Ostrava-Poruba, Czech Republic
ivan.zelinka@vsb.cz

ABSTRACT

Post-processing is an essential step in detecting and correcting errors in OCR-generated texts. In this paper, we present an automatic OCR post-processing model which comprises both error detection and error correction phases for OCR output texts of unconstrained Vietnamese handwriting. We propose a hybrid approach of generating and scoring correction candidates for both non-syllable and real-syllable errors based on the linguistic features as well as the error characteristics of OCR outputs. We evaluate our proposed model on a Vietnamese benchmark database at the line level. The experimental results show that our model achieves 4.17% of character error rate (CER) and 9.82% of word error rate (WER), which helps improve both CER and WER of an attention-based encoder-decoder approach by 0.5% and 3.5% respectively on the VNOnDB-Line dataset of the Vietnamese online handwritten text recognition competition (VOHTR2018). These results outperform those obtained by various recognition systems in the VOHTR2018 competition.

CCS CONCEPTS

• **Computing methodologies** → *Language resources*; *Neural networks*; • **Applied computing** → *Optical character recognition.*

KEYWORDS

Unconstrained Vietnamese handwriting, OCR, Post-processing, Error detection, Error correction

ACM Reference Format:
Quoc-Dung Nguyen, Duc-Anh Le, and Ivan Zelinka. 2019. OCR Error Correction for Unconstrained Vietnamese Handwritten Text. In *The Tenth International Symposium on Information and Communication Technology (SoICT 2019), December 4–6, 2019, Hanoi - Ha Long Bay, Viet Nam.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3368926.3369686

1 INTRODUCTION

Optical Character Recognition (OCR) is the process of transforming typed, handwritten or printed text from scanned documents or images into digital text using various image processing and pattern recognition techniques [3, 5]. However, the OCR process often results in misspellings and linguistic errors in OCR-generated texts due to misrecognized characters, falsely identified text images as well as limitations of text recognition techniques. Post-processing is an important step to improve the quality of existing OCR output texts by detecting and cleaning the errors.

The increasing popularity of pen-based and touch-based devices has led to the crucial demand in processing so-called digital ink (a time sequence of pen/touch points) recently. The handwriting recognition systems have been developed with the ability to recognize, exchange and search for handwritten text from digital ink in order to meet various requirements and applications in education, entertainment, business and so on. Unconstrained handwritten text is written naturally in each writer's style without any restriction. Hence, the unconstrained handwritten text usually contains many variations in size, shape, slant, skew, and stroke order (see Fig. 1).

Unconstrained handwritten text recognition can be categorized into two main types: online and offline recognition. The first type of recognition makes use of spatial and temporal information of points of pen trajectory and strokes as input features to the recognition systems, for example, English [4], Chinese [20], Korean [8], etc. In the offline recognition type, only offline images (e.g. converted from online handwritten text) are available for image processing and recognition, such as English [2], Iranian [1], Indian [22], etc. In fact, the offline type of handwritten text recognition can be seen as a subtask of the OCR.

Vietnamese is a Latin script language. However, unlike other Latin script languages such as English, French or Spanish, Vietnamese contains a large amount of diacritic marks (DM), which are added to characters to transcribe all the sounds or to indicate variations in speech. They are placed over, under or through characters. For unconstrained handwritten text, the place and order of these DM strokes could be varied due to different writers or even different writing times of the same writer. In other words, DMs are often not positioned right where they should, and their sizes and shapes are also varied. They can be written after writing a few strokes of subsequent characters, or even after a sentence (called delayed DMs). These distorted and delayed DMs cause difficulties in recognizing unconstrained Vietnamese handwriting.

A few works have been proposed for Vietnamese online handwritten character recognition [17, 18, 21]. They came up with some solutions for solving the DM problem at character level. However, these approaches require pre-segmented handwritten text, they are not capable of dealing with cursively handwritten text in practice.

SoICT 2019, December 4–6, 2019, Hanoi - Ha Long Bay, Viet Nam

ACM ISBN 978-1-4503-7245-9/19/12…$15.00
https://doi.org/10.1145/3368926.3369686