




# An In-depth Analysis of OCR Errors for Unconstrained Vietnamese Handwriting

Quoc-Dung Nguyen<sup>1,4</sup> , Duc-Anh Le<sup>2,5</sup>, Nguyet-Minh Phan<sup>3</sup>,  
and Ivan Zelinka<sup>4</sup>

<sup>1</sup> Van Lang University, 45 Nguyen Khac Nhu, Co Giang ward, District 1,  
Ho Chi Minh City, Vietnam

`dung.nguyen@vlu.edu.vn`

<sup>2</sup> Center for Open Data in the Humanities, Tokyo, Japan

`leducanh841988@gmail.com`

<sup>3</sup> University of Information Technology, Quarter 6, Linh Trung Ward,  
Thu Duc District, Ho Chi Minh City, Vietnam

`minhpn@uit.edu.vn`

<sup>4</sup> Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava-Poruba,  
Czech Republic

`ivan.zelinka@vsb.cz`

<sup>5</sup> NTT Hi-Tech Institute, Nguyen Tat Thanh University, 300A Nguyen Tat Thanh,  
District 4, Ho Chi Minh City, Vietnam

**Abstract.** OCR post-processing is an essential step to improve the accuracy of OCR-generated texts by detecting and correcting OCR errors. In this paper, the OCR texts are resulted from an OCR engine which is based on the attention-based encoder-decoder model for unconstrained Vietnamese handwriting. We identify various kinds of Vietnamese OCR errors and their possible causes. Detailed statistics of Vietnamese OCR errors are provided and analyzed at both character level and syllable level, using typical OCR error characteristics such as error rate, error mapping/edit, frequency and error length. Furthermore, the statistical analyses are done on training and test sets of a benchmark database to infer whether the test set is the appropriate representative of the training set regarding the OCR error characteristics. We also discuss the choice of designing OCR post-processing approaches at character level or at syllable level relying on provided statistics of studied datasets.

**Keywords:** OCR errors · OCR post-processing · Vietnamese handwriting · Encoder · Decoder · Attention model

## 1 Introduction

Optical Character Recognition (OCR) is the process of transforming scanned document images into digital texts. However, the output texts of this process often contain errors due to many reasons such as poor quality of scanned documents, unusual font sizes and layouts. The erroneous OCR texts make the texts

© Springer Nature Switzerland AG 2020

T. K. Dang et al. (Eds.): FDSE 2020, LNCS 12466, pp. 448–461, 2020.

[https://doi.org/10.1007/978-3-030-63924-2\\_26](https://doi.org/10.1007/978-3-030-63924-2_26)