THAI TYPE STYLES RECOGNITION

LIBRARY AND INFORMATION CENTER
MAHIDOL UNIVERSITY

SUTAT SAETANG

With compliments
of
MAHIDOL UNIVERSITY

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
1998

ISBN 974-661-136-4

COPYRIGHT OF MAHIDOL UNIVERSITY

3837400 SCCS/M : MAJOR : COMPUTER SCIENCE; M.Sc. (COMPUTER SCIENCE)
KEY WORD : THAI TYPE STYLE / THAI OCR / NEURAL NETWORK / BACK-PROPAGATION / THINNING ALGORITHMS

SUTAT SAETANG THAI TYPE STYLES RECOGNITION. RESEARCH PROJECT ADVISOR : CHULARAT TANPRASERT Ph.D., DAMRUS WONGSAWANG Ph.D., 59 p. ISBN 974-661-136-4

Thai printed character recognition has been a very popular research topic in Thailand. There are three commercial Thai OCR softwares available to the public at the present. None of them can preserve the type styles of the original document image such as normal, bold, italics, and bold & italics styles into the output text file. Therefore, users who need to maintain a document's original character type styles have to modify the document by themselves which takes more time and more labor on a tedious job.

This research presents the technique for preserving the specified Thai type styles by applying a specific preprocessing with a supervised neural networks learning algorithm. Only four type styles of Thai typed characters are considered. They are normal, bold, italics and bold & italics. Therefore, there are two main features to extract for these four Thai type styles : the thickness and the inclination of character.

This research designed and experimented several types of templates to extract these two features from the raw bit-map character images. The best template preserves the two main characteristics and gives an average recognition at 95.85% with the unseen testing patterns. Therefore, the results confirm that the proposed technique effectively preserves the type styles of Thai typed fonts from the original document image into the output text file.

Copyright by Mahidol University

3837400 SCCS/M สาขาวิชา · วิทยาการคอมพิวเตอร์; วท.ม. (วิทยาการคอมพิวเตอร์)

สุทัศน์ แซ่ตั้ง · การรู้จำลักษณะพิเศษของตัวอักษรไทย (Thai Type Styles Recognition) คณะกรรมการควบคุมโครงการวิจัย จุฬารัตน์ ตันประเสริฐ Ph.D., ดำรัส วงศ์สว่าง Ph.D , 59 หน้า
ISBN 974-661-136-4

การรู้จำตัวพิมพ์อักษรไทยเป็นหัวข้อวิจัยที่กำลังนิยมมากหัวข้อหนึ่งในประเทศไทย ปัจจุบันมีซอฟต์แวร์ทางด้านนี้ในท้องตลาดของประเทศไทย 3 ซอฟต์แวร์ด้วยกัน แต่ไม่มีซอฟต์แวร์ตัวใดเลย ที่สามารถรู้จำรูปแบบตัวอักษรของเอกสารต้นฉบับเช่น ตัวปกติ ตัวหนา ตัวเอียง และตัวหนาเอียงได้เลย ด้วยเหตุนี้ผู้ใช้ที่ต้องการจะได้รูปแบบตัวอักษรของเอกสารต้นฉบับในแฟ้มข้อความผลลัพธ์จึงจำเป็นต้องแก้ไขด้วยตนเองภายหลัง ซึ่งเป็นการเสียเวลาและเป็นงานที่น่าเบื่อหน่าย

งานวิจัยฉบับนี้ จะแสดงถึงเทคนิคในการรู้จำรูปแบบตัวอักษรไทยโดยอาศัยโครงข่ายประสาทเทียมแบบมีผู้สอนช่วยในการรู้จำ โดยจะรู้จำรูปแบบทั้งหมด 4 รูปแบบคือ ตัวปกติ ตัวหนา ตัวเอียง และตัวหนาเอียง ซึ่งสามารถแบ่งเป็นลักษณะของรูปแบบหลัก ๆ ของตัวอักษรภาษาไทยได้ 2 รูปแบบด้วยกัน คือ รูปแบบความหนา และรูปแบบความเอียงของตัวอักษร

จากการวิจัยได้ออกแบบและทดสอบกับหลาย ๆ แผ่นแบบ (template) ที่พัฒนาขึ้น เพื่อให้สามารถดึงรูปแบบหลักทั้ง 2 รูปแบบจากภาพลักษณ์ตัวอักษรไทยได้ โดยแผ่นแบบที่ดีที่สุดสำหรับการดึงรูปแบบหลักทั้ง 2 รูปแบบดังกล่าวมีอัตราการรู้จำที่ 95.85% กับข้อมูลทดสอบ ดังนั้นสามารถสรุปได้ว่าเทคนิคที่พัฒนาขึ้นสามารถรู้จำรูปแบบตัวอักษรภาษาไทยจากภาพลักษณ์ตัวอักษรได้อย่างมีประสิทธิภาพ

Copyright by Mahidol University