

ชื่อเรื่องการค้นคว้าแบบอิสระ การรู้จำอักขระไทยด้วยแสงโดยวิธีวัดระยะทาง
แบบเฮาส์คอฟฟ์และโครงข่ายประสาทเทียม

ผู้เขียน นายบรรจง ปิยศทิพย์

ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

อาจารย์ที่ปรึกษาการค้นคว้าแบบอิสระ
อาจารย์ ดร.เสมอแหะ สมหอม

บทคัดย่อ

การค้นคว้าแบบอิสระเชิงวิทยานิพนธ์นี้ มีวัตถุประสงค์หลักเพื่อศึกษาการนำโครงข่ายประสาทเทียมและการวัดระยะทางแบบเฮาส์คอฟฟ์ มาใช้ในการรู้จำอักขระไทย โดยทำการออกแบบและพัฒนาโปรแกรมด้วยภาษาเดลไฟ แล้วจึงทำการฝึกสอนโครงข่ายประสาทเทียมให้รู้จำอักขระไทยด้วยจำนวนตัวอย่าง 14,550 ตัว ประกอบด้วย รูปแบบตัวอักษร 6 รูปแบบ คือ อังสนานิวคอร์เดียนิว คิลเลนเนียยูพีซี ยูโครเซียยูพีซี ฟรีเซียยูพีซี และไอริสยูพีซี แต่ละรูปแบบมีขนาด 12 14 16 18 และ 20 พอยต์

ในขั้นตอนการรู้จำอักขระ ได้ทำการแยกอักขระจากรูปภาพโดยอาศัยฮีสโตรแกรมในแนวนอนและแนวตั้ง และทำการหาระดับของอักขระในแต่ละบรรทัด โดยได้ทำการแบ่งบรรทัดออกเป็น 4 ระดับ นอกจากนี้ยังนำวิธีวัดระยะทางแบบเฮาส์คอฟฟ์มาใช้ในการแยกสระบนออกจากพยัญชนะในกรณีที่มีการซ้อนทับกัน จากนั้นจึงทำการวิเคราะห์อักขระโดยใช้โครงข่ายประสาทเทียมด้วยค่าน้ำหนักที่ได้จากการฝึกสอน

ผลจากการทดลองพบว่า เมื่อทำการทดลองฝึกสอนโครงข่ายประสาทเทียมด้วยอักขระจำนวน 6 รูปแบบๆ ละ 5 ขนาด และกำหนดขนาดของอินพุตเมตริกเท่ากับ 30×23 จำนวนนิวรอนในชั้นซ่อนเท่ากับ 300 โหนด ค่าอัตราการเรียนรู้เท่ากับ 0.4 ค่าความผิดพลาดเฉลี่ยเท่ากับ 0.0001 แล้วนำมาทดสอบกับบทความที่มีรูปแบบตัวอักษร 6 รูปแบบๆ ละ 5 ขนาดเช่นเดียวกับข้อมูลที่ใช้ในการฝึกสอน ซึ่งภายในบทความจะมีจำนวนอักขระอยู่ 1,903 ตัว ผลการทดสอบพบว่า โปรแกรมสามารถวิเคราะห์อักขระได้ถูกต้องเฉลี่ยร้อยละ 95.35

Independent Study Title Thai Optical Character Recognition Using Hausdorff Distance
and Artificial Neural Network

Author Mr.Banjong Piyottip

Degree Master of Science (Computer Science)

Independent Study Advisor Lecturer Dr.Samerkae Somhom

ABSTRACT

This independent study has the main purpose to study the application of the Artificial Neural Network (ANN) and the Hausdorff Distance measurement to recognize the Thai characters. Delphi is used in order to design and develop the program, then we train the ANN to recognize 14,550 Thai sample characters. The samples consist of six different font types; Angsana New, Cordia New, DilleniaUPC, EucrosiaUPC, FreesiaUPC and IrisUPC, with font sizes varies from 12, 14, 16, 18 and 20 points.

In the recognition step, a Histogram technique is used to extract characters from pictures vertically and horizontally. Each line is divided into four levels and read the character according to each line level. Besides, in case those characters are overlapped, they are separated by the Hausdorff Distance technique. Finally, the analyze character is with the taken weight from the test.

The result shows that when train the ANN with six fonts, five sizes each and set the input matrix size to 30x23 with 300 nodes in hidden layer the learning rate is 0.4 with average error of 0.0001. When test the program with 1903 characters article which has the same font type and size as the samples. It shows that the program can analyze correctly at the average of 95.35 percent.