

EE368 Project Proposal (Revised 10.26.2016)

Project Name: Recognition and Translation of Thai Characters and Words from Text

Project members:

- Nattapoom Asavareongchai
- Evan Giarta

Goal/Objective:

- Detect, segment, and translate Thai characters from images of printed text. (**Character classification**)
 - This includes characters that have been rotated, slanted, within noisy images, and in uneven lighting conditions.
 - Initially, we will test and work on detecting a single font of Thai, then add different fonts and sizes to improve algorithm robustness.
 - We will also just start with the 44 consonants and if we are able to characterize these, we will move on to the different vowels that exists on top, underneath, and on the side of consonants.
- We will **collect data** from Thai newspaper websites.
 - First our training data on one font would be done by copying text into google docs and printing them out as a jpeg images.
 - We will physically print them out and take photos of these document in different lightings and orientations. Use these as both training and test data.
 - We will then use images from actual newspapers that's been downloaded online as further test data with different fonts and sizes.
- Translate the text after language detection
- Reach Goal: Machine learning to derive feature model for vowel and consonant detection, primary handwritten Thai.
- We will **evaluate** our performances on the percentage of correct character classifications of each characters. The number of false positives and misclassifications. We will try to increase the correct characterization percentage and decrease misclassification and false positive percentages.

Methods:

1. Create database of Thai characters starting with one font then using multiple fonts.
 - a. Obtain computer type font screenshots of the 44 Thai consonants, its 15 vowel symbols, and the various vowel forms to use as character templates.
2. Generate character detection filters, such as hit-miss and minimum rank filters, and templates from various Thai fonts.
3. Process and extract features and characters from images via locally adaptive gray level thresholding, binarization, and region moments.
 - a. Resize and rotate images for proper matching and alignment
 - b. Reduce noise with median filtering, etc.

- c. Identify regions of neighboring text.
4. Detect characters using filters and template matching.
5. Translate words and sentences with Google
6. Additional: extract features of different Thai consonants and vowels using machine learning tools to create a model to detect handwritten Thai.

Dataset Samples (not exact sizes):

Thai Consonants:

1. ก	23. ท
2. ข	24. ฑ
3. ฃ	25. น
4. ค	26. บ
5. ฅ	27. ป
6. ฆ	28. ผ
7. ง	29. ฝ
8. จ	30. พ
9. ฉ	31. ฟ
10. ช	32. ภ
11. ฌ	33. ม
12. ฎ	34. ย
13. ญ	35. ร
14. ฎ	36. ล
15. ฏ	37. ว
16. ฐ	38. ศ
17. ฑ	39. ษ
18. ฒ	40. ส
19. ณ	41. ห
20. ด	42. ฬ
21. ต	43. อ
22. ถ	44. ฮ

Clean document sample:

อัศจรรย์นกเอี้ยงนับหมื่นหลบ พระบรมฉายาลักษณ์'ฟอหลวง'
อัศจรรย์! เมืองขุนแผน
ฝูงนกเอี้ยงนับหมื่นตัวหลบปายพระบรมฉายาลักษณ์"ฟอหลวง"ชาวบ้านเชื่อเป็นเพราะพระบารมีของพระ
องค์เคยก่อนหน้าไล่เท่าไรก็ไม่ไป กระทั่งนำพระบรมฉายาลักษณ์มาติดตั้งเพื่อแสดงความอาลัย
อังคารที่ 25 ตุลาคม 2559 เวลา 11.17 น.

เมื่อวันที่ 25 ต.ค. ที่ จ.สุพรรณบุรี ได้เกิดเหตุอัศจรรย์ขึ้น
เมื่อนกเอี้ยงนับหมื่นตัวได้พากันย้ายที่นอนจากที่เกาะอยู่บนสายไฟฟ้าแรงสูงบริเวณหน้าแขวงทาง
สุพรรณบุรีที่ 1 ไปอาศัยอยู่บนต้นไม้ที่ถนนแฉกร้าวห่างจุดเดิมประมาณ 500 เมตร

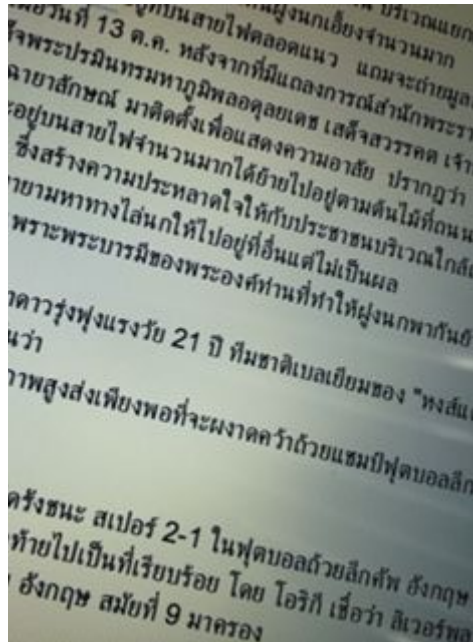
ทั้งนี้นายกิตติ โพธิ์เตียน อาสาสมัครมูลนิธิเสมอกันกู้ภัยสุพรรณบุรี กล่าวว่า
ตนพร้อมเพื่อนอาสาสมัครมูลนิธิเสมอกันกู้ภัยสุพรรณบุรี
ได้พากันมาคอยช่วยเหลือประชาชนที่ถนนมาลัยแมน บริเวณแยกแขวงทางสุพรรณบุรีที่ 1 เขต
อ.เมืองสุพรรณบุรี มาหลายปี และจะเห็นฝูงนกเอี้ยงจำนวนมาก
นับหมื่นตัวมาอาศัยเกาะอยู่ที่บนสายไฟตลอดแนว แฉมจะถ่ายมูลลงพื้นจนดูสกปรก
กระทั่งต่อมาเมื่อวันที่ 13 ต.ค. หลังจากที่มีแถลงการณ์สำนักพระราชวังว่า
พระบาทสมเด็จพระปรมินทรมหาภูมิพลอดุลยเดช เสด็จสวรรคต เจ้าหน้าที่เทศบาลเมืองสุพรรณบุรี
ได้นำพระบรมฉายาลักษณ์ มาติดตั้งเพื่อแสดงความอาลัย ปรากฏว่า
ฝูงนกที่เคยเกาะอยู่บนสายไฟจำนวนมากได้ย้ายไปอยู่อาศัยตามต้นไม้ที่ถนนแฉกร้าวห่างกันประมาณ 500
เมตรจนไม่เหลือ ซึ่งสร้างความประหลาดใจให้กับประชาชนบริเวณใกล้เคียงเป็นอย่างมาก
เพราะที่ผ่านมาพยายามหาทางไล่นกให้ไปอยู่ที่อื่นแต่ไม่เป็นผล
จึงเชื่อว่าจะเป็นเพราะพระบารมีของพระองค์ท่านที่ทำให้ฝูงนกพากันย้ายไปอยู่ที่อื่น.... อ่านต่อที่

ดิวอด โอริกิ กองหน้าดาวรุ่งพุ่งแรงวัย 21 ปี ทีมชาติเบลเยียมของ "หงส์แดง" ลิเวอร์พูล
ยอดทีมแดนผู้ดี เชื่อกันว่า
ต้นสังกัดของเขามีศักยภาพสูงส่งเพียงพอที่จะผงาดคว้าถ้วยแชมป์ฟุตบอลศึก ฟุตบอลชิงแชมป์แห่งชาติยุโรป 2016-17 มาครอง

ผลงานล่าสุด ลิเวอร์พูล เปิดรังชนะ สเปน 2-1 ในฟุตบอลถ้วยลีกคัพ อังกฤษ รอบ 4 ทำให้ "หงส์แดง"
ตีปีกบินลี้วเข้ารอบ 8 ทีมสุดท้ายไปเป็นที่เรียบร้อย โดย โอริกิ เชื่อว่า ลิเวอร์พูล
มีโอกาสคว้าถ้วยแชมป์ลีกคัพ อังกฤษ สมัยที่ 9 มาครอง

"เรายังอยู่ในเส้นทางการเล่นแชมป์ และเราต้องการไปให้สุดทาง มันไม่ง่าย
แต่เราแสดงให้เห็นถึงคาแรคเตอร์นักเตะของทีม เราเล่นด้วยหัวใจที่แข็งแกร่งในทุก ๆ เกม
และผมมั่นใจว่า เรามีคุณภาพที่จะคว้าแชมป์มาครอง".... อ่านต่อที่

Testing data:



Newspaper sample:

เดลินิวส์

ฉบับที่ 23,169 วันพุธที่ 20 มีนาคม พ.ศ. 2556 อ่านความจริง อ่านเดลินิวส์

www.dailynews.co.th ราคา 10 บาท

คุณภาพ

หมุนบันทึกอินเทอร์เน็ต

'เพิ่มคุณภาพ'

'โมเดล' อ-โรฮงยัด?

สรุปหน้า 1 เดลินิวส์
หน้า 1

แ มีประเทศไหนที่จะมีการพัฒนาด้านอุตสาหกรรมกันมานานแล้ว แต่ปัจจุบันตลาดก็ยังคงเป็นประเทศเป็นหลัก และไทยก็เป็นตลาดของธุรกิจจีนที่เริ่มจากธุรกิจโลกตะวันออกและตะวันออก ซึ่งสำหรับธุรกิจจีนที่เริ่มจากโลกตะวันออก ก็อย่างที่ว่ากัน ๆ กันว่าระลอก ๆ เกาหลีได้ได้แต่แรก และก็เป็นกรณีศึกษาด้วย

"โมเดลไทยอินเทอร์เน็ต" ถูกถกมานานแล้ว แต่ในทางปฏิบัติจริง ยังไม่ได้เห็นผลอะไรสักอย่าง ซึ่งหากถามว่าเพราะอะไร? ก็คือสิ่งที่ในหลวงที่เสด็จมาทรงพระกรุณาโปรดเกล้าฯ ให้ตั้ง

"ที่ในภาคนี้ซึ่งสามารถสร้างผลงานอันเป็นที่ภาคภูมิใจขึ้นมาจากปัจจัยสำคัญคือ คุณภาพของอินเทอร์เน็ต" ...นี่เป็นการระบุของ พันธุ์พันธุ์ วัฒนาธการกิจ ประธานกรรมการกลุ่มบริษัท ออเอสอาร์ อินเทอร์เน็ต กรุ๊ป กลุ่มบริษัทที่ร่วมกับมหาวิทยาลัยอาเซีย เปิด "วิทยาลัยอินเทอร์เน็ตเอเชีย (Superstar College of Asia)" เปิดหลักสูตรสำหรับปริญญาตรีที่โปรแกรมศาสตร์เพื่อผลิตบุคลากรสำหรับวงการนี้

ทั้งนี้ พันธุ์พันธุ์ยังตั้งข้อสังเกตว่า การพัฒนาอุตสาหกรรมด้านอินเทอร์เน็ตที่ไทยมีศักยภาพได้บ้างช่วงก่อนหน้านี้... ปีนี้คือเรื่อง "อินเทอร์" ด้วยนโยบายของกระทรวงการคลังที่นำงบฯ มาช่วยเหลือและขยายวงกว้าง รัฐบาลเกาหลีใต้ได้ตั้งกองทุนเพื่อสนับสนุนกิจการด้านอินเทอร์เน็ต ทำให้บริษัทเน็ตเวิร์กของเกาหลีใต้ ซึ่งแม้จะมีขนาดใหญ่ แต่ก็สามารถผลิตออกมาคุณภาพเทียบเท่ากับเกาหลีใต้ และส่งออกสู่ตลาดได้ ขณะเดียวกันเรื่องอุตสาหกรรมที่ไทยได้ดำเนินการได้แก่ การตั้งกองทุนเพื่อสนับสนุนผู้ประกอบการด้านอินเทอร์เน็ต

"ในเชิงของเงินทุน จุดต่างระหว่างไทยกับเกาหลีใต้คือ รัฐบาลไทยมีนโยบายช่วยเหลือผู้ประกอบการด้านอินเทอร์เน็ตให้เป็นอย่างดี ขณะที่รัฐบาลเกาหลีใต้มีนโยบายที่ชัดเจนกว่า รัฐบาลไทยจะพัฒนาถึงขั้นสนับสนุนผู้ประกอบการด้านอินเทอร์เน็ต

หมุนบันทึกอินเทอร์เน็ต 'เพิ่มคุณภาพ' 'โมเดล' อ-โรฮงยัด?

จะผลิตขึ้นได้ยาก แต่เทคโนโลยีภาคเอกชนของวงการเน็ตเวิร์กผู้เข้ามาในธุรกิจที่ค่อนข้างเข้มแข็งทางด้านการเงิน ทายาทบริษัทจีนเข้มแข็งกว่าบริษัทเกาหลีใต้หลายเท่า ทำให้พอจะสามารถช่วยตัวเองในด้านการลงทุนได้ ที่นี้ก็ต้องมององอีกปัจจัยหนึ่งว่า ในเมื่อเงินทุนพร้อม แล้วทำไมไทยยังไม่ถึงปลายทางอย่างเกาหลีใต้บ้าง คำตอบคือ คุณภาพอินเทอร์เน็ตยังไม่ดีมากรวมทั้ง... พันธุ์พันธุ์ระบุ พร้อมทั้งชี้ถึงปัจจัยเรื่อง "คุณภาพอินเทอร์เน็ต" ว่า...

การที่จะผลิตอินเทอร์เน็ต ต้องเริ่มมาจากวัตถุดิบที่ คือ "ชิป" โดยการผลิตชิปที่มีคุณภาพ และระยะเวลาการผลิตที่ยาวนานเพียงพอ เพราะชิปเป็นต้นทุน ไม่ใช่เรื่องจู้จี้ คุณภาพของชิปขึ้นอยู่กับขนาดชิปที่นักวิจัยคิดค้นอย่างยากลำบากและใช้เวลาพัฒนาเพราะชิปที่เพิ่งพอ ซึ่งความสำคัญด้านเน็ตเวิร์กของเกาหลีใต้ โดยคุณภาพชิปเป็นส่วนสำคัญส่วนหนึ่งมาจาก **โมเดลวิทยาลัยอินเทอร์เน็ต** เช่น วิทยาลัยอินเทอร์เน็ตที่มหาวิทยาลัยอัสยาเปิดขึ้นเป็นแห่งแรกเมื่อ 12 ปีก่อน และมีส่วนในการสร้างความสัมพันธ์กับพันธมิตรให้เกิดขึ้นกับเกาหลีใต้ เช่น เวบสำหรับในประเทศไทย ที่ผ่านมากับเน็ตเวิร์กเน็ตเวิร์ก



เรื่องการเรียนของชิปเป็นสิ่งที่สำคัญ เช่น บางงานเน็ตเวิร์ก แต่ไปไม่ถึงฝั่งเพราะไม่สามารถหา-ไม่สามารถหาชิปได้ ซึ่งเมื่อชิปเป็นต้นทุนก่อนก็อาจจะต้องพึ่งงบการเงินบ้างหรือไม่ หรือชิปในบางงานต้องใช้เวลาเรียนกันถึงวันละแปดชั่วโมง ที่เหนื่อย และเพื่อเวลาที่ซื้อชิป การแข่งขันกันเอง ซึ่ง การขยายแบบ "จับปลาเหนือ" ย่อมประสบความสำเร็จได้ยากขึ้น

จากจุดนี้ พันธุ์พันธุ์ระบุว่า... วิทยาลัยผลิตชิปหนักสุดในระดับปริญญาตรี คือค่าสอนสำหรับทุกฝ่าย ทั้งสำหรับพ่อแม่ที่อยากให้ลูกสามารถเรียนจบปริญญาตรี สำหรับผู้ที่ต้องการเป็นชิปเป็น ซึ่งการเรียนก็เป็นการที่เรียนการแต่ง หลานการแต่ง หรือผลงานของ ที่คือการส่งไปรษณีย์ให้อาจารย์ และยังเป็นค่าสอนสำหรับค่ายันท์ซึ่งสามารถลดต้นทุนและเวลาในการผลิตชิปในเชิงธุรกิจ เพราะสามารถเลือกนักศึกษาไปใช้งานได้ หรือสามารถส่งชิปไปเชิงธุรกิจพัฒนาคุณภาพชิปในวิทยาลัย โดยที่มหาวิทยาลัยที่ไปสร้างผลงานได้

Superstar College ๖ ปีเปิดสอน วิทยาลัย (Superstar Asia) สร้างชิปเกาหลี, ญี่ปุ่น

โดย AEC (Association of East Asian Nations) คนอื่น และคิดเรื่องการแต่ง ก็อาจเรียน มีการแข่งขันสูงเป็นรายตัว สร้างชิปเป็นของมหาวิทยาลัยที่ชิปของญี่ปุ่น และกำลังจะ

เปิดว่า... **คุณภาพเน็ตเวิร์กไทย** วิทยาลัยผลิตชิปเกาหลีใต้

เพียงสองเดือนอายุ 10-15 ปี เพื่อหาหลักสูตรที่ผลิตชิปให้ใช้ในอนาคต 6 ปี ถึงเวลานี้มหาวิทยาลัยได้ตั้งใจ จึงต้องมีการตั้งกองทุน ชูตัวนี้ผ่านเน็ตเวิร์กเน็ตเวิร์กไทยไปใช้สักพัก นี้เป็นสิ่งที่เน็ตเวิร์กเน็ตเวิร์กที่ไทยไปใช้สักพัก ซึ่งก็พบว่าไทยไม่เก่งด้านนี้ ส่วนการร่วมมือกับของต่างชาติและวิทยาลัยผลิตชิปเอเชีย ไทยน่าจะดี "ไทยและเกาหลีใต้ (Thailand Model)" เปรียบอย่างเกาหลีใต้

อย่าที่เกาหลี... มาเกาหลี... และภาคีความร่วมมือเน็ตเวิร์ก... เพื่อเน็ตเวิร์ก

References:

1. Jin, Michelle, Ling Xiao Wang, and Boyang Zhang. Poster: "Text to Image Translation for Restaurant Menus." EE 368/CS 232, Department of Electrical Engineering, Spring 2014.
2. Phokharatkul, Pisit, and Chom Kimpan. "Recognition of handprinted Thai characters using the cavity features of character based on neural network." *Circuits and Systems, 1998. IEEE APCCAS 1998. The 1998 IEEE Asia-Pacific Conference on*. IEEE, 1998.
3. Hochberg, Judith, et al. "Automatic script identification from images using cluster-based templates." *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. Vol. 1. IEEE, 1995.