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. ธ
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. ซ	33. ม
12. ฌ	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'พ่อหลวง'
อัศจรรย์! เมืองขุนแผน
ฝูงนกเอี้ยงนับหมื่นตัวหลบป้ายพระบรมฉายาลักษณ์"พ่อหลวง"ชาวบ้านเชื่อเป็นเพราะพระบารมีของพระองค์เผยก่อนหน้าไล่เท่าไรก็ไม่ไป กระทั่งนำพระบรมฉายาลักษณ์มาติดตั้งเพื่อแสดงความอาลัย
อังคารที่ 25 ตุลาคม 2559 เวลา 11.17 น.

เมื่อวันที่ 25 ต.ค. ที่ จ.สุพรรณบุรี ได้เกิดเหตุอัศจรรย์ขึ้น
เมื่อฝูงนกเอี้ยงนับหมื่นตัวได้พากันย้ายที่นอนจากที่เกาะอยู่บนสายไฟฟ้าแรงสูงบริเวณหน้าแขวงการทางสุพรรณบุรีที่ 1ไปอาศัยอยู่บนต้นไม้ที่ถนนเณรแก้วห่างจุดเดิมประมาณ 500 เมตร

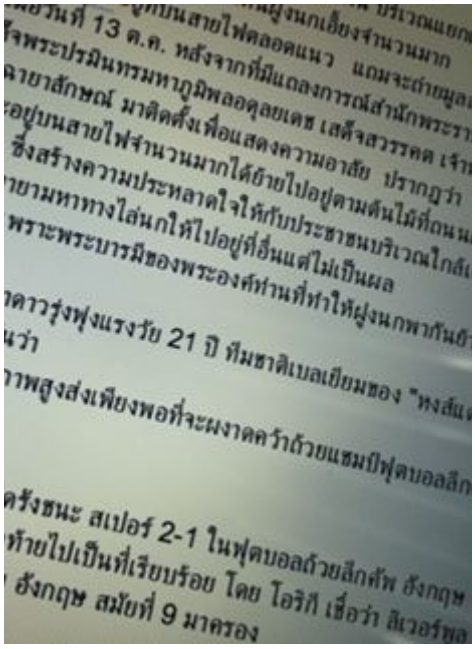
ทั้งนี้นายกิตติ โพธิ์เตียน อาสาสมัครมูลนิธิเสมอกันกู้ภัยสุพรรณบุรี กล่าวว่า
ตนพร้อมเพื่อนอาสาสมัครมูลนิธิเสมอกันกู้ภัยสุพรรณบุรี
ได้พากันมาคอยช่วยเหลือประชาชนที่ถนนมาลัยแมน บริเวณแยกแขวงการทางสุพรรณบุรีที่1 เขต อ.เมืองสุพรรณบุรี มาหลายปี และจะเห็นฝูงนกเอี้ยงจำนวนมาก
นับหมื่นตัวมาอาศัยเกาะอยู่ที่บนสายไฟตลอดแนว แถมจะถ่ายมูลลงพื้นจนดูสกปรก
กระทั่งต่อมาเมื่อวันที่ 13 ต.ค. หลังจากที่มีแถลงการณ์สำนักพระราชวังว่า
พระบาทสมเด็จพระปรมินทรมหาภูมิพลอดุลยเดช เสด็จสวรรคต เจ้าหน้าที่เทศบาลเมืองสุพรรณบุรี
ได้นำพระบรมฉายาลักษณ์ มาติดตั้งเพื่อแสดงความอาลัย ปรากฏว่า
ฝูงนกที่เคยเกาะอยู่บนสายไฟจำนวนมากได้ย้ายไปอยู่ตามต้นไม้ที่ถนนเณรแก้วห่างกันประมาณ500 เมตรจนไม่เหลือ ซึ่งสร้างความประหลาดใจให้กับประชาชนบริเวณใกล้เคียงเป็นอย่างมาก
เพราะที่ผ่านมาพยายามหาทางไล่นกให้ไปอยู่ที่อื่นแต่ไม่เป็นผล
จึงเชื่อว่าน่าจะเป็นเพราะพระบารมีของพระองค์ท่านที่ทำให้ฝูงนกพากันย้ายไปอยู่ที่อื่น.... อ่านต่อที่

ดิวอค โอริกี กองหน้าดาวรุ่งพุ่งแรงวัย 21 ปี ทีมชาติเบลเยียมของ "หงส์แดง" ลิเวอร์พูล
ยอดทีมแดนผู้ดี เชื่อมั่นว่า
ต้นสังกัดของเขามีศักยภาพสูงส่งเพียงพอที่จะผงาดคว้าถ้วยแชมป์ฟุตบอลลีกคัพ อังกฤษ ฤดูกาล 2016-17 มาครอง

ผลงานล่าสุด ลิเวอร์พูล เปิดรังชนะ สเปอร์ 2-1 ในฟุตบอลถ้วยลีกคัพ อังกฤษ รอบ 4 ทำให้ "หงส์แดง" ตีปีกบินลิ่วเข้ารอบ 8 ทีมสุดท้ายไปเป็นที่เรียบร้อย โดย โอริกี เชื่อว่า ลิเวอร์พูล
มีโอกาสคว้าถ้วยแชมป์ลีกคัพ อังกฤษ สมัยที่ 9 มาครอง

"เรายังอยู่ในเส้นทางการลุ้นแชมป์ และเราต้องการไปให้สุดทาง มันไม่ง่าย
แต่เราแสดงให้เห็นถึงคาแรกเตอร์นักสู้ของทีม เราลงเล่นด้วยหัวจิตหัวใจที่แข็งแกร่งในทุก 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

สกู๊ปหน้า 1

หนุนบันเทิงโกอินเตอร์ 'เพิ่มคุณภาพ' 'โมเดล'อะไรช่วยได้?

หนุนบันเทิงโกอินเตอร์

'เพิ่มคุณภาพ'

'โมเดล' อะไรช่วยได้?

สร้างศิลปินเกาหลี, ญี่ปุ่น

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.