Recognition of Thai Characters and Text from Document Template

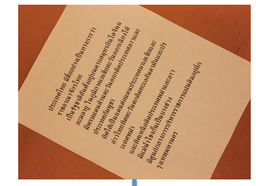
Nattapoom Asavareongchai, Evan Giarta
Department of Electrical Engineering, Stanford University
EE368 Digital Image Processing, Autumn 2016

Motivation

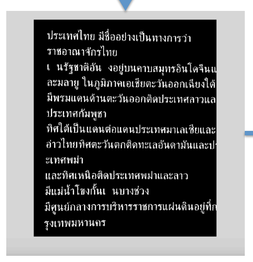
Text documents, such as reports and journals, in different languages usually require people fluent in those languages to read and translate them manually. This could take a long time to do. Why not have it automated by a program?

Solution: Build an image processing system to process text document images taken with a phone/camera and translate them text.

The structure of different written languages are different and Thai is one language that involves more subtle complex written patterns than languages like English. It becomes a challenging problem to tackle.



Preprocess



Recognition

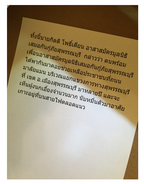
ประเทศไทย มีชื่ออย่างเป็นทางการว่า ราชอาณาจักรไทย เป็นรัฐชาติอันตั้งอยู่บนคาบสมุทรอินโดจีน ...

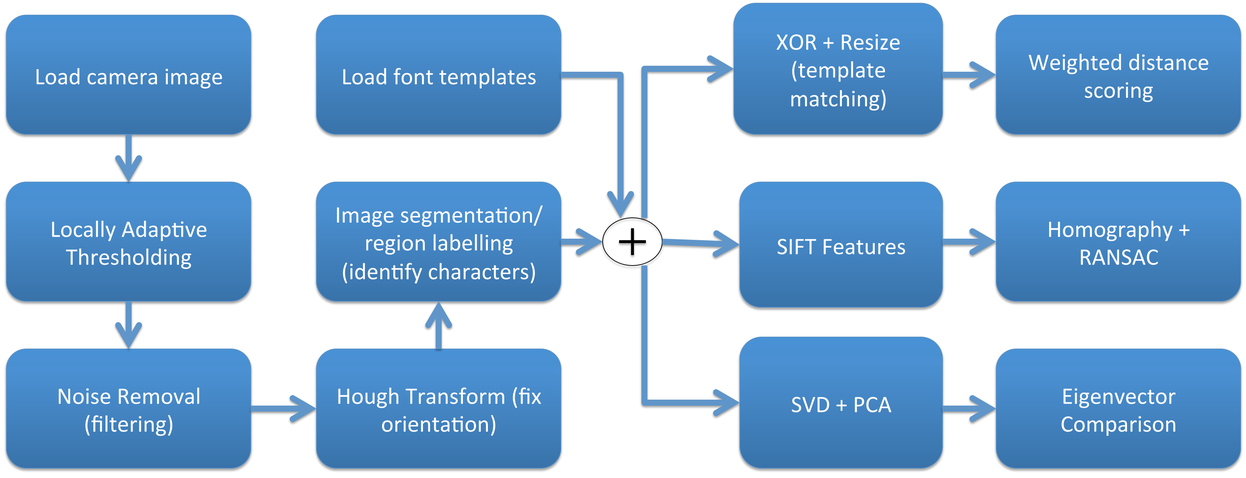
Translate

Thailand has an official name. Kingdom of Thailand The state is located on the Indochina Peninsula....

Processing Pipeline and Methods

Camera Image





Results

Correctly classify:

Misclassify:

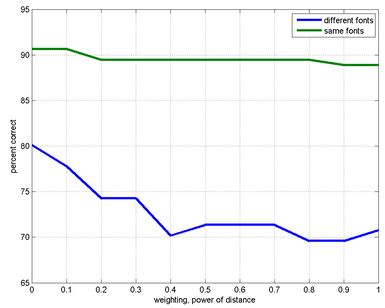
Experimental Results

Documents \ Detection Method	XOR + imresize + weighted distance score	SVD + PCA	SIFT + RANSAC
Clean Test Document (same font)	91%	73%	64%
Clean Test Document (different font)	80%	36%	36%
Test Camera Document (same font)	95%	69%	59%
Test Camera Document (different font)	83%	36%	18%

NOTE: Table shows the percentage of correct character detection

- The testing documents we used contains actual sentences from Thai newspaper and articles. This is so that it will represent the frequency of each characters being used
- For XOR method, same characters are either consistently correctly classified or consistently misclassify. This is not the case for SVD and SIFT

Graph of XOR results vs. weighted penalty:



Limitations & Future Improvements:

1. Does not work with photos taken with warped page, (sizes of characters varies in the photo).
2. Possible new technique, gather data from many fonts, then create eigenimages/fisherimages of each character and find best classfication.

Related Work/ Reference

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.