Learning Based Word Spotting for Arabic Handwritten Documents Using Hierarchal Classifier

Muna Khayyat (PhD Student – 3rd Year – Concordia University)
Center for Pattern Recognition and Machine Learning (CENPARMI)
Supervisors: Dr. C. Y. Suen and Dr. L. Lam

Introduction

Word spotting has been widely implemented for Latin-based and Chinese documents. However, few word spotting systems have been implemented for Arabic handwritten documents. Yet, Arabic is spoken by a significant number of the world's population. Arabic script is cursive by nature; besides, in Arabic writing words have no clear boundaries; these facts make the implementation of word spotting for Arabic handwritten documents a significant challenge.

We proposed a learning-based word spotting system that uses Support Vector Machine (SVM) to recognize sub-words rather than complete words. We will use this partial segmentation concept to resolve the word boundary problem in Arabic handwriting. In addition, a shortest distance algorithm is used to spot Arabic handwritten words.

Work Progress

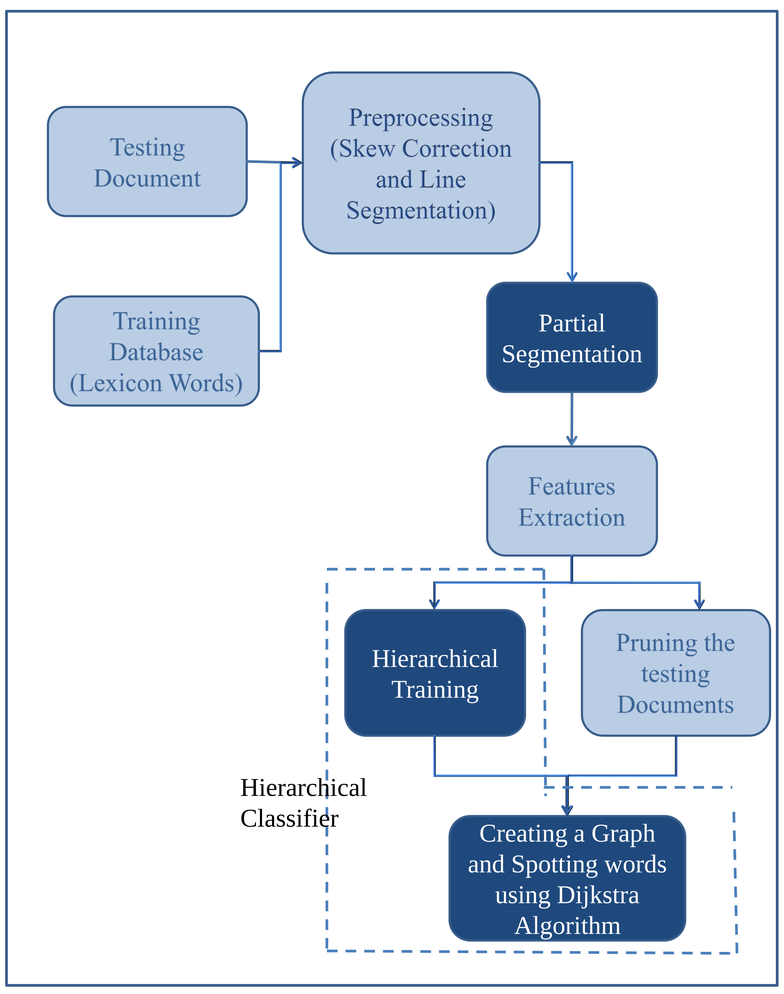
Progress to date

1. An Arabic word recognition system using three different feature sets has been constructed. The features were tested on CENPARMI and other databases. The highest classification accuracy on CENPARMI Arabic words database is 95.46 % using gradient features. Another feature sets will be extracted in the future. Those features are supposed to bring higher classification accuracy.
2. Partial segmentation of the words to their PAWs.
3. Arabic text line segmentation.

Current Work

Document Pruning.

Research Plan

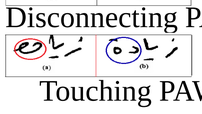


Partial Segmentation

We reconstructed CENPARMI database [1] by segmenting it into PAWs. Some PAWs may be touched or disconnected . However, people seems to connected or disconnect almost in a similar manner. Thus, additional classes were considered for the new formulated PAW. The machine will learn these new classes instead of applying some heuristics and rules trying to connect the disconnected PAW and vise versa.

Difficulties

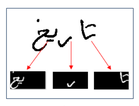

Disconnecting PAW


Touching PAWs

Class Number	PAW	Confusing class	PAW
3	ر	15	ا
3	ر	19	د
3	ر	23	ز
10	ن	45	ة
17	تا	44	يا
33	تَـ	34	بَـ
58	تسعه	73	سبعه

Confusing PAW in the database

Partial Segmentation


Segmenting a word into its PAW

Results

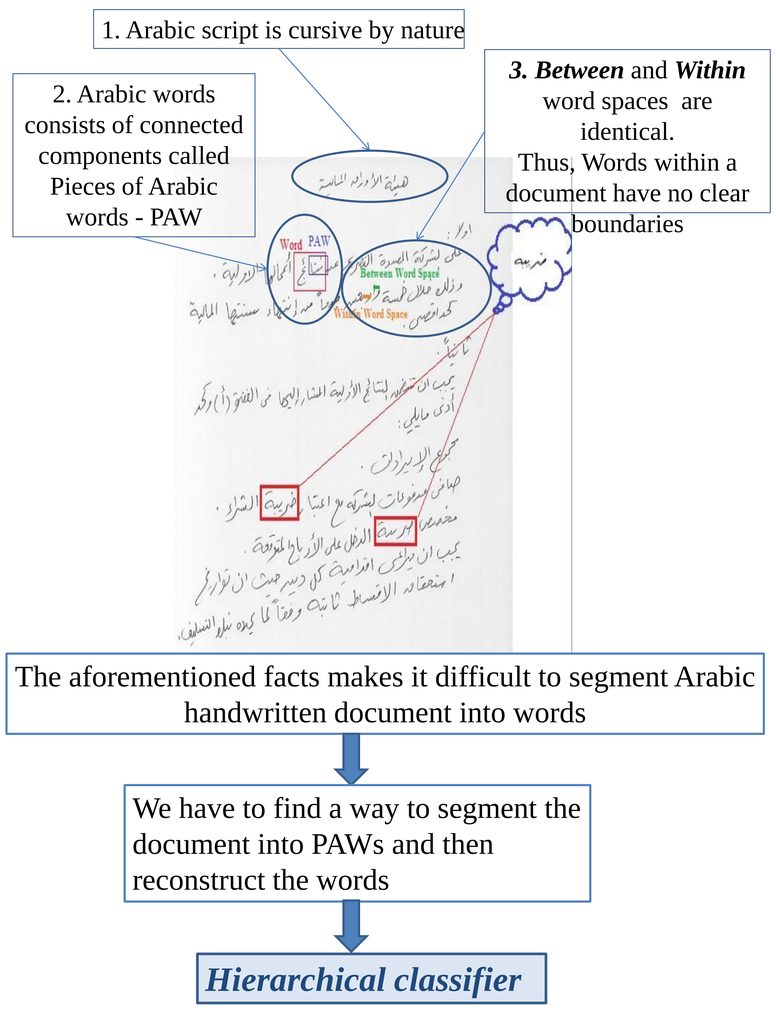
Extracted Features	Recognition Rate %	
	127 PAW Classes	92 PAW Classes
Gradient Features	***88.4***	***89.3***
Gabor Features	***82.5***	***82.5***

PAW Recognition Results

Class ID	Printed Sample	Confusion Class ID	Printed Sample of the confusion class	Number of misclassified samples	Misclassified samples in the database
66	ت	10	ن	28	
3	ر	19	د	20	
19	د	26	و	19	
10	ن	45	ة	18	
19	د	3	ر	17	
26	و	19	د	16	
25	ق	10	ن	15	
22	صديل	7	ميل	11	
82	ثمه	42	مه	10	

The most frequent confusion classes taken from the confusion matrix resulted from extracting the gradient features from the PAW images.

Problems with Arabic Word Spotting



The aforementioned facts makes it difficult to segment Arabic handwritten document into words

We have to find a way to segment the document into PAWs and then reconstruct the words

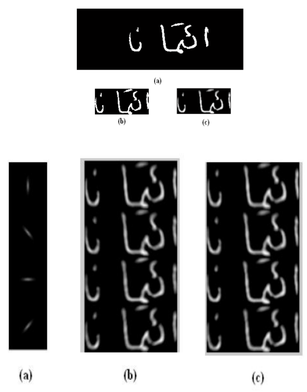
Hierarchical classifier

Arabic Handwritten Word Recognition

Three sets of features were extracted to recognize Arabic handwritten words, with each set of feature passed to one classifier. (Support Vector Machine - SVM) The confidence levels and classification results of the classifiers were used for the final classification. The CENPARMI database is used for validation.

Preprocessing

1. Binarization (Otsu Algorithm)
2. Size normalization *120 50*
3. Smoothed grayscale Image



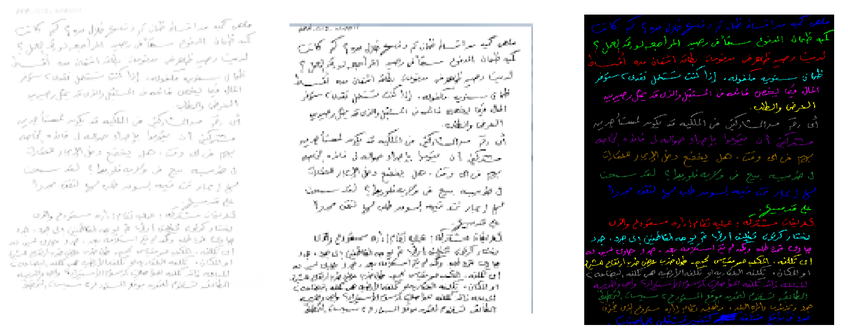
Feature Extraction:

1. **Gradient Features [3]**
2. **Gabor Features [2]**
3. **Fourier Features [4]** The Discrete Fourier Transform of the projection, upper and lower profiles of the words were extracted:

Arabic Text Line Segmentation

Arabic handwriting is cursive by nature and it includes small connected components called diacritics. These facts add constrains to the segmentation of the Arabic handwritten documents into lines.


Probability Density Function (PDF) of the document


(a) Original Document (b) Preprocessing (c) Line Segmentation

References

[1] H, Alamri, J. Sadri, C. Y. Suen, and N. Nobile "A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition," *In Proceedings of Eleventh International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 664-669, Montreal, Canada, August 2008.

[2] J. Chen, H. Cao, R. Prasad, A. Bhardwaj, and P. Natarajan, \Gabor Features for O²ine Arabic Handwriting Recognition", In Proceedings of the 9th International WORKSHOP on Document Analysis Systems (DAS '10), pp. 53 - 58, 2010.

[3] M. Shi, Y. Fujisawa, T. Wakabayashi, and F. Kimura, "Handwritten Numeral Recognition Using Gradient and Curvature of Gray Scale Image", *Pattern Recognition*, vol. 35, No. 10, pp. 2051 - 2059, 2002.

[4] V. Lavrenko, T. M. Rath, and R. Manmatha, "Holistic Word Recognition for Handwritten Historical Documents", In Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL), pp. 278 - 287, Palo Alto, CA, USA, January 2004.

TC10 Technical Committee on Graphics Recognition

2011 ICDAR Doctoral Consortium

IAPR TC-11 Reading Systems