



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

Introduction	<b>Partial Segmentation</b>	<b>Arabic Handwritten Word Recognition</b>			
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.	Thus, additional classes were considered for the new formulated PAW. The machine will learn these new classes instead of applying some heuristics and rules	<ul> <li>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.</li> <li>Preprocessing <ol> <li>Binarization (Otsu Algorithm)</li> <li>Size normalization 120 50</li> </ol> </li> </ul>			
We proposed a learning-based word spotting system	Difficulties Recognition Rate %	3. Smoothed grayscale Image			

we proposed a learning-based word sponning 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**

- 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.
- Partial segmentation of the words to their PAWs. Arabic text line segmentation.

Extracted **127 PAW** aites Dio Features Classes Disconnecting PAW **Gradient Features** 88.4 Gabor Features 82.5 Touching PAWs PAW Recognition Results Confusing PAW PAW class Ľ 35 بذ

بتا ريچ

- E

Segmenting a word

into its PAW

Class Number	PAW	Confusing class	PAW	Class	Printed	Confusion	Printed Sample	Number of				
3	5	15	1	ID	Sample	Class ID	of the confusion class	misclassified samples	Mis	classified sa	amples in the d	atabase
3	5	19	2	66	ت	10	ن	28	Ċ	υ	$\sim$	ف
3	<b>、</b>	23	j	3	J	19	د	20	~		د (	
10	Ú	45	5	19	د	26	و	19	>		2 2	
17	Ľ	44	Ľ	10				10				
33	تذ	34	بذ	10	ن	45	5	18	ථ	ٺ	S	ů
58	ئسيعه	73	سيعه	19	د	3	ر	17	)		、ノ	ر
Confi	using	PAW	in the	26	و	19	د	16	3	د	Э	و
database		25	ق	10	ن	15	6	ა	Ġ	ښ		
				22	صيل	7	ميـل	11	صل	of a	ريسو	Jus
'arti	al So	egme	ntatio	n <sup>82</sup>	ثه	42	ئــه	10	طت	مړ	ى2	0

**92 PAW** 

Classes

*89.3* 

82.5

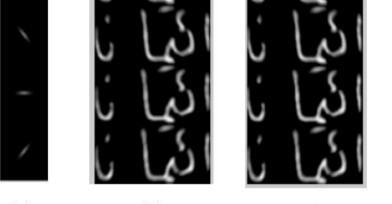
The most frequent confusion classes taken from the confusion matrix resulted from extracting the gradient features from the PAW images.

#### **Problems with Arabic Word Spotting**

1. Arabic script is cursive by nature

reature Extraction: **Gradient Features [3]** 

**Gabor Features [2]** 



**Fourier Features** [4] The Discrete Fourier Transform of the projection, upper and lower profiles of the words were extracted:

## **Arabic Text Line Segmentation**

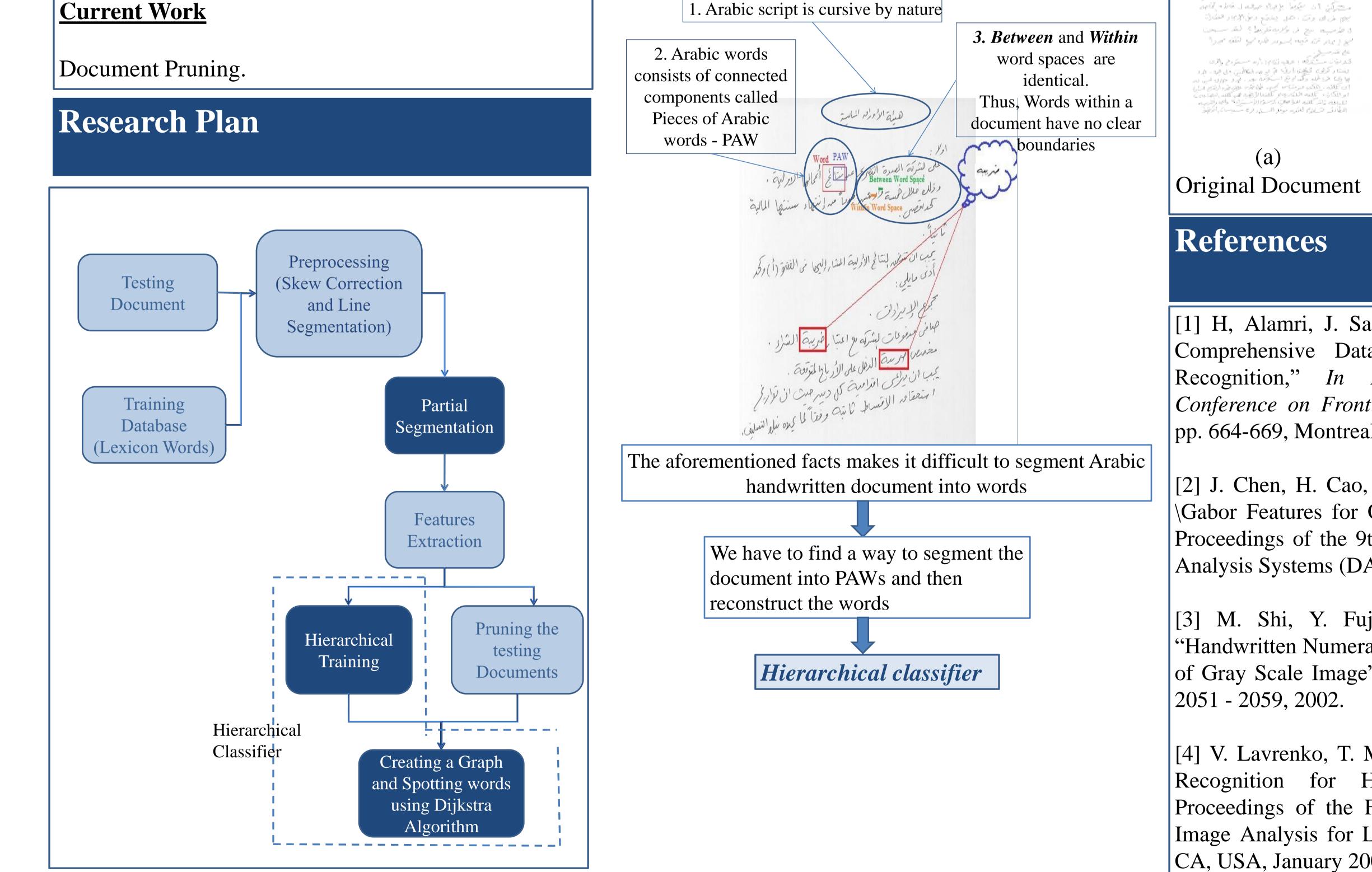
Arabic handwriting is cursive by nature and it includes small connected components called diacritics. These facts add constrains to the segmentation of the Arabic handwritten documents into lines.



Probability Density Function (PDF) of the document

رملص محبه مداشداة ظنان تو وتنسيو فلال مرماي كم كانت معصامي مداشية فلات ومسيوفين مواكم كان كبه فجماد المغنى مسبعاتهم رجبيه كالراحية تورقد إعارى لا المدخل مستعكم في مصبر المراجع له فكراج كميه ظمان الدفن مستقائم رجب أعراجه لويخر إيعن و السعيف ومجهد فلياهرض معاشوهاته المجاشر بالمتهان مده الأتي الدميت دجهير للإيوم متغولت يتملك اشقان مده المتسب كما مستويه ملغوله. إذا كنت مستعمل تقدى منو اللها ي مستقريه مالغوانه . الذا كلنه مستلحل تقديه العس اغلل فيها ليخص خاخصف الحستثل دالال قداعكم رجسه الحال مأينا البطحان خالف فر المستبن والغرى تقه ميكن رقبتيري أن رقة مداف ركل من الملكيم تد تجهر لمستأجريه العرف وانطقه. بش رقع سرامیت؟ کمکر من اللکیہ تد ککہ المستاً جریہ بهم خراف وقت على ينتقر والح الايحار فعقال مستتنكي أن الميكيل بإجماد مهاندن فالمرم إلتهم נ מתיידים נוש זה הצקים ועציון אידי אידי المجمع من أن وحمد ، عان المنظمة وعن الإيمار المتخدرة اليهم ( جابر المثنة الحييه السويد الحلي الي الملكة المحرر ل الأميسة بسيخ عن مركزية للوبط؟ المقد سسحت الميح الرجابر المقد شيه المسومر الحان الميح المقت محبد(أ فتداخات مستشكرته وعرقيه لتكامرا بأره مستقردهم والمرى ینی فتر سیستی . از هر نظرت می شدند از عباریه از یا مرا در ۵ مستقود هم و آقرال وزیر از تر می شدید اور ۲ م این چره افغافینی دی چره چره این از مرا مکان و ترک موضح است کرده مجرور و خیری از عباری اسی م این محلکان و ترک بند الحکام برمتر تحلیما از این ها مارت هماری م این بده از از اللانه احمام جری کردها از این می محله و تصریح این بده از از اللانه احمام جری کردسکار از می مسیوستی افکانی این ماده از از اللانه احمام جری کردسکار از می مسیوستی این می دشتار گرفته فیضته ارتی از نوم. اعلمتن بالی فرد. عرا وختار کردی کنید، ارد بخ بوجه المقاطبین ای عرق الما المنا المنا المله وقد اج المسكل المرام المرد جاران الم ان كلفه، بالكندية فسرحقياً ما أنجميه، الجاة القرب عليمان المرافقين في واللكان، إكلتم التكاريمار كلتما للغرام في كلتم إنجاع للباعب لماتذ كلتمه المطحاني كرسخا الأستيل كالمواقد وأشرب this level white there and all الطألف الشيطاح للخلدية موقع المسيقاد فرته ستسيساته إقرابهم (a) (b)(c)Line Segmentation Preprocessing [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 O<sup>2</sup>ine 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.

[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.



# **2011 ICDAR** Doctoral Consortium



