Shape-based Analysis for Segmentation of Arabic Handwritten Text

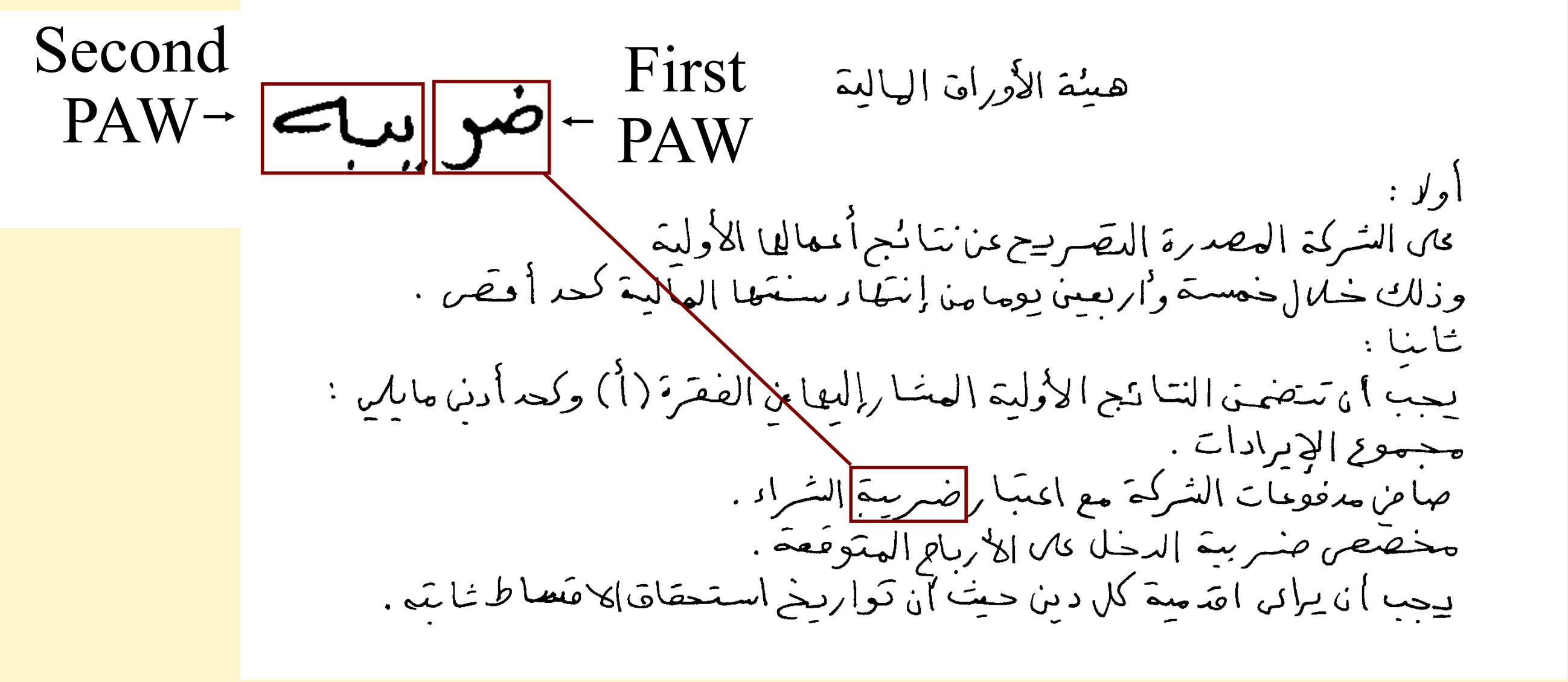
CENPARMI
Centre for Pattern Recognition and Machine Intelligence

Ph.D Candidate: Amani T. Jamal
Supervisor: Ching Y. Suen

UNIVERSITÉ Concordia UNIVERSITY

Introduction

Extracting main units from a handwritten document is an essential pre-processing step for two reasons [7]:
(1)Text recognition methods
letter-based and word-based
(2)Word-spotting or content-based image retrieval techniques
Most of the techniques in handwritten document retrieval and recognition will fail if the texts are wrongly segmented into words. However, sometimes the cause of failure in Arabic-related methods is the incorrectly segmented text into sub words or Parts of Arabic Word (PAWs).


Second PAW
First PAW
ضر
يبه

هيئة الأوراق المالية

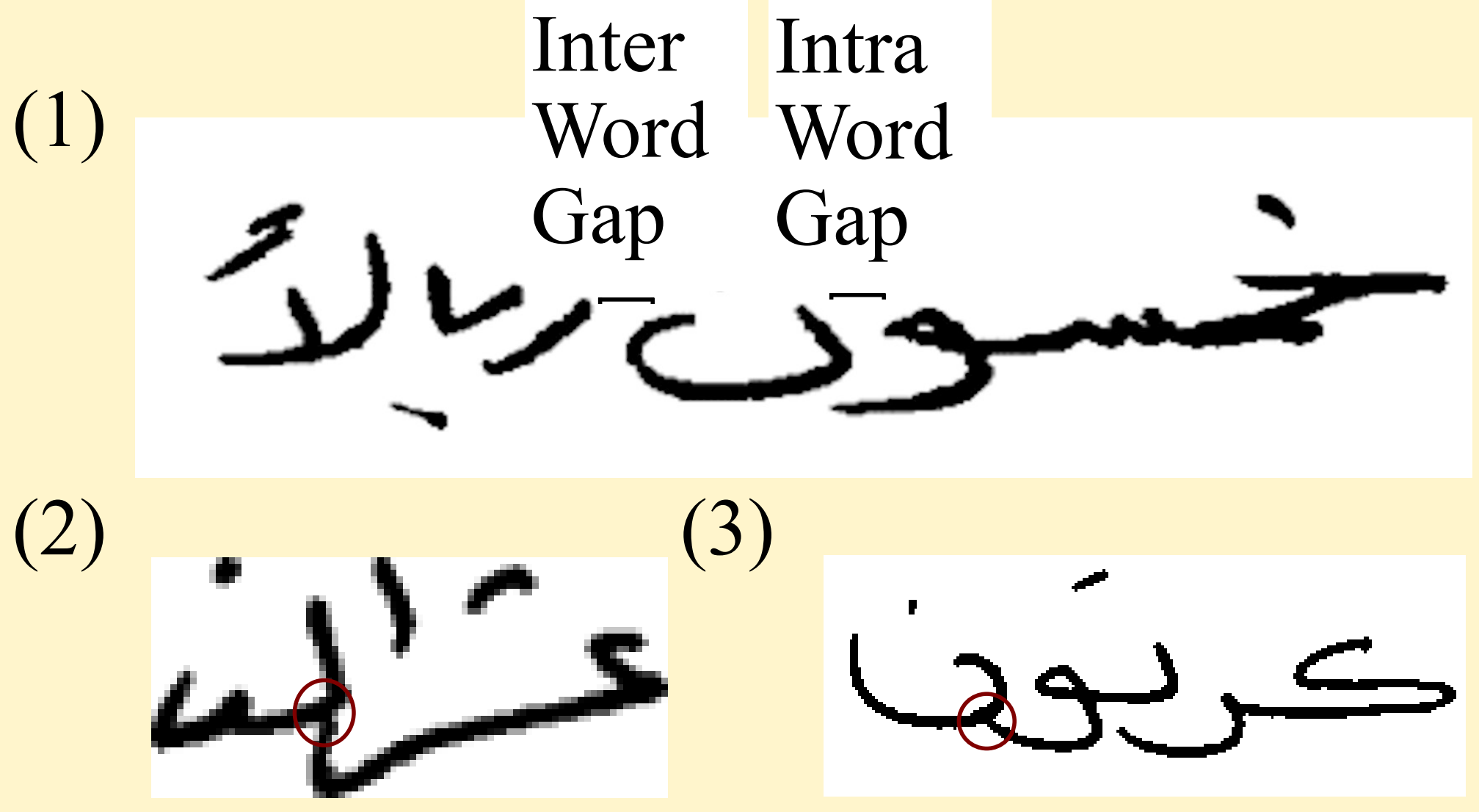
أولا :
على الشركة المصدرة التصريح عن نتائج أعمالها الأولية
وذلك خلال خمسة وأربعين يوما من إنتهاء سنتها المالية كحد أقصى .
ثانيا :
يجب أن تتضمن النتائج الأولية المشار إليها في الفقرة (أ) وكحد أدنى ما يلي :
مجموع الإيرادات .
صافي مدفوعات الشركة مع اعتبار ضريبة الشراء .
مخصص ضريبة الدخل على الأرباح المتوقعة .
يجب أن يراعى اقدمية كل دين حيث أن تواريخ استحقاق الاقساط ثابتة .

Objectives

. Holistic Segmentation (Word)
. Semi-Holistic Segmentation (PAW)

Challenges

(1) Lack of well defined boundaries
(2) Touching words
(3) Touching PAWs



Previous Works

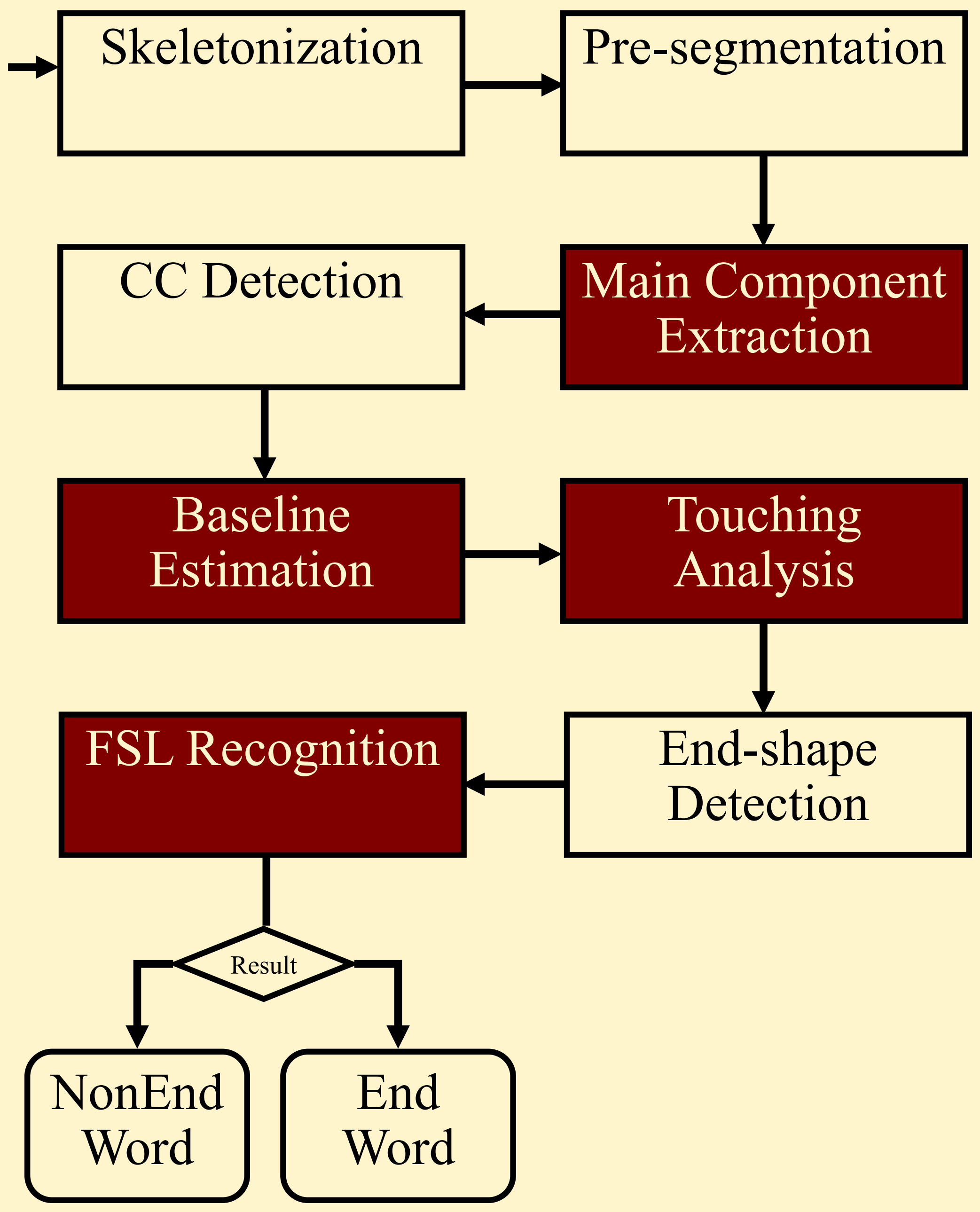
System Type	No. Images	Image Type	Result
Threshold	106-200	IFN/ENIT	66-91
Classification	100	Document	60
Scaling	5	Document	71.5-97.5

Approach

Utilize the knowledge of Arabic Writing

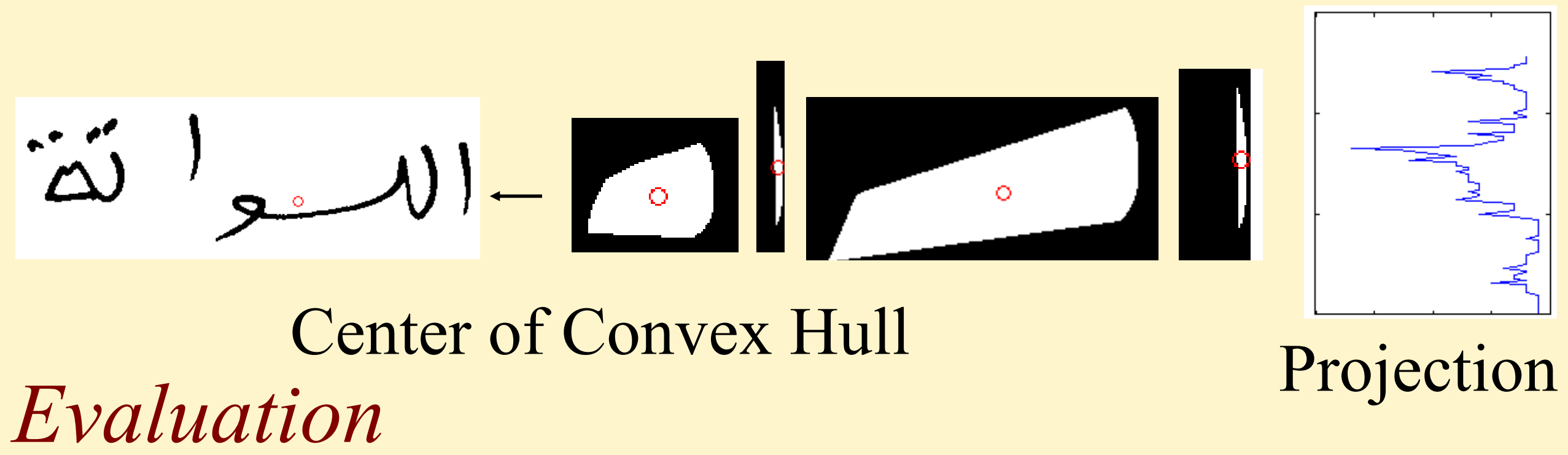
مجموع
أربعة
عنصر

Methodology [1]



Baseline Estimation [3]

. Learning-based Approach
. DB generation (based on 5 pixel) [8]
. Pre-processing : Horizontal Normalization
. Baseline –relevant feature extraction


Center of Convex Hull
Projection

Evaluation

Class	error in pixels	Percentage
One word	<=5	47.21 %
	<=10	90.39 %
Two words	<=5	18.15 %
	<=10	30.70 %
Three words	<=5	15.22 %
	<=10	47.28 %

Future Work

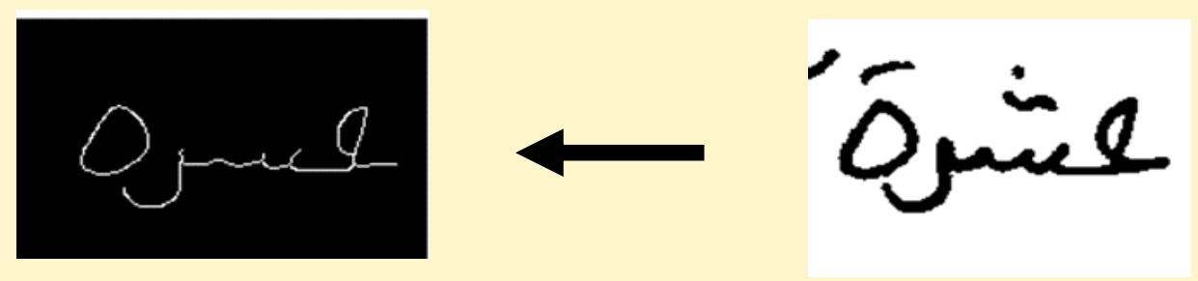
. Hough transform technique to extract the horizontal line segments
. Locating the holes
. Rotating the image to find the peak
. PAW

Used Databases

. IFN/ENIT [4]
. CENPARMI
. Documents [2]
. Words [5]
. Cheques [6]

Main Component Extraction

. Middle line locating
. Morphological Reconstruction



Evaluation

# images	Performance
20	87.5%

Future Work

Use some heuristic rules

Touching Analysis

. Database generated from CENPARMI word DB, IFN/ENIT, CENPARMI cheque DB
. 3 Classes: ascender, descender and baseline touching

Future Work

. Develop a classifier to identify the type of touching.
. Detect segment point

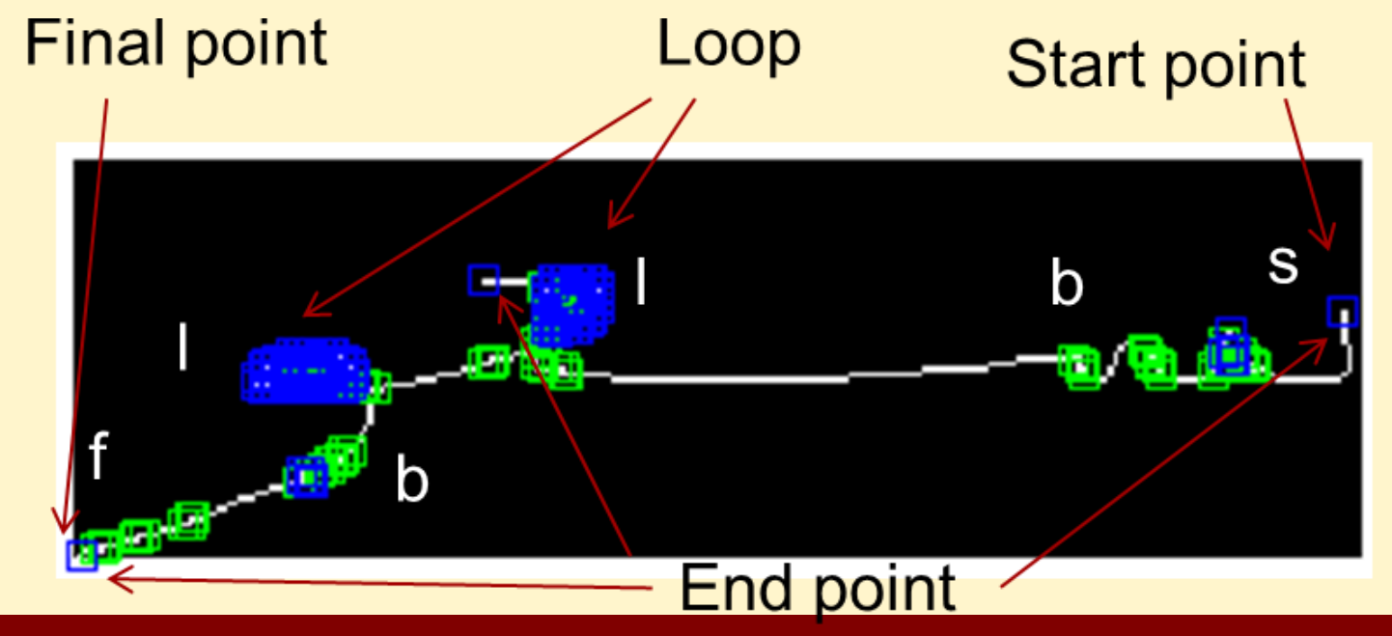
Final Shape Letter Recognizer

ـه ل س ى ص ع ح م ك ن ه ط ف ق ث

. 15 Classes
. Support Vector Machine (SVM)

Future Work

Final Shape Letter extraction



Evaluation

. Detection Rate (DR)

$$DR = o2o/N$$

N is the count of ground-truth elements
o2o is the number of one-to-one matches

. Recognition Accuracy (RA)

$$RA = o2o/M$$

M is the count of result elements

. Performance Metric (FM)

$$FM = \frac{2\ DR\ RA}{DR + RA}$$

References

[1] A. T. Jamal and C. Y. Suen “Shape-based Analysis for Automatic Segmentation of Arabic Handwritten Text,” In Advances in Artificial Intelligence, pp. 334-339. Springer Berlin Heidelberg, 2013.

[2] M. Khayyat, L. Lam C. Y. Suen, F. Yin and C. Liu “Arabic Handwritten Text Line Extraction by Applying an Adaptive Mask to Morphological Dilation,” In 10th IAPR international. Workshop on Document Analysis Systems (DAS), pp. 100-104, 2012.

[3] A.T. Jamal, N. Nobile, and C. Y. Suen “Learning-based Baseline Estimation,” In 11th International Conference "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-11-2013) (September 23-28, 2013, Samara, The Russian Federation). Accepted

[4] M. Pechwitz, S. S. Maddouri, V. M¨argner, N. Ellouze and H. Amiri, “IFN/ENIT- Database of Handwritten Arabic Words,” In Colloque Inter. Francophone sur l’Ecrit et le Document (CIFED), Vol. 2, pp. 127–136, 2002.

[5] H. Alamri, J. Sadri, C. Y. Suen and N. Nobile, ”A Novel Comprehensive Database for Arabic Off-line Handwriting Recognition,” In Proceding of 11th International Conference on Frontiers in Handwriting Recognition (ICFHR.08), pp. 664-669, 2008.

[6] Y. Al-Ohali, M. Cheriet and C. Y. Suen, “Databases for Recognition of Handwritten Arabic Cheques,” Pattern Recognition, Vol. 36, No. 1, pp. 111-121, 2003.

[7] C. Huang, and S. Srihari,”Word Segmentation of Off-line Handwritten Documents,” In Proceedings of the Document Recognition and Retrieval (DRR) XV, IST/SPIE Annual Symposium, Vol. 6815, 2008.

[8] M. Pechwitz, H. Abed, and V. Märgner,”Handwritten Arabic Word Recognition Using the IFN/ENIT-database,” Guide to OCR for Arabic Scripts, pp. 169-213, 2012.