Arabic Handwriting forgery detection.

Sylvia Hani, Diana Abdel Naser, Hetem EL Herzawy, Caroline Kamal.
Supervised by: Dr. Sherin Moussa, Eng. Youssef Mohamed.

4 October 2017

Abstract:
Developing a new technology to help detecting forgery in an Arabic handwritten documents, based on Arabic OCR to detect Arabic language, then using HMM to extract features from every language and build a huge data-set to train characters recognition based of characters features for every individual handwriting style. then based on the results of each individual a handwriting; another handwriting features are being matched with it to detect forgery based on accuracy detection, done by matching string matrices.

1 Introduction

1.1 Background

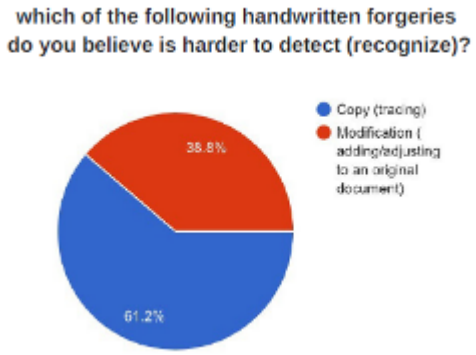
A lot of different organizations have plotted the difficulty for detecting forged documents for their contracts signatures and trusted handwritten documents. detecting forgery could be challenging due to the style of the handwriting of different people and what differ a single handwriting to another. In this project we will be focusing on applying Arabic OCR to read Arabic Handwriting, then extraction of features of different hand-writings from different people. then running a matching algorithms to detect the accuracy between handwriting's features that were extracted to detect forgery.

1.2 Motivation

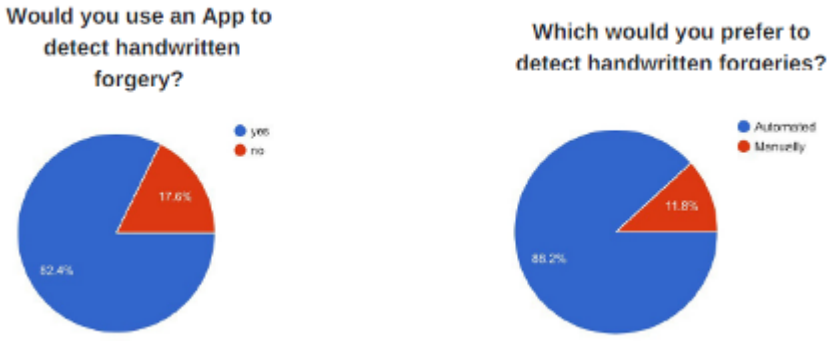
Forgery is a series crime that could lead to prison up to 25 years according to law 206, 207 in the Egyptian book of law, that's why detecting forgery is essential.Many forgery detection systems have been created over the past few years, but not detecting Arabic handwriting forgery, according to a survey that we created and 175 persons took it, 65 percent of them stated that they actually encountered forgery.

1.2.2 Academic Motivation

1- Arabic OCR.
2- Feature Extraction from Arabic handwriting Algorithms.
4- Accuracy of detection.
4- Creating data sets.
5- Matching Algorithms.
6- performance optimization.
References:
• Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling.
• Recognition of Off-line Arabic Handwriting words Using HMM Toolkit.



Caption 1: The tracing of the handwriting between 2 copies of same word is more complicated than the modification detection.



Caption 2: Usage of an app for detecting forgery is more required, which will be automated to make it easier and faster than the manual way.
By this result we need the features extraction to the handwritten to apply matching algorithms by implementing code of those algorithms in the application.

1.3 Problem Definitions

Several challenges have been faced by detection of forged handwriting systems because of the similarities found in them, Arabic handwriting is even challenging because it is a cursive language. We aim to recognize Arabic handwriting style and detect forgery occurrence. and optimizing the accuracy by using different feature extraction method.

2 Project Description

Detecting Arabic Handwritten Forged Documents using Arabic Character recognition, Arabic handwriting feature extraction methods to detect a personal style for each handwriting and then matching algorithms to detect forgery based on the features extracted form the handwriting and their matching accuracy.

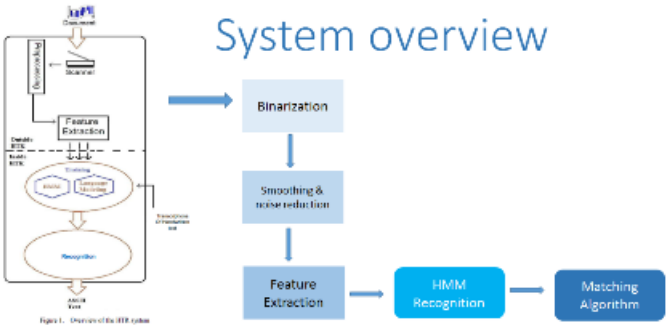
2.1 Objective

Our system is designed to become an effective, accurate tool for detecting Arabic Handwritten Forged documents. Arabic handwritten forgery are often detected by experts MANUALLY, it takes time, accuracy is not always achieved, lack of experts; are all factors why our system is needed in the market. Forgery is considered a crime and it could lead to huge consequences in legal documents, bank checks, doctors prescriptions, tests, signatures. Our system will compare the original document and the proposed forged document to detect similarity in their features with accuracy percentage in an optimum time. by using Algorithms like HMM, HTM,Viterbi algorithm, BBN,rotation operation, horizontal projection and vertical projection profile.

2.2 Scope

1. System can detect Arabic handwriting by using Arabic OCR.
2. System can extract different features from handwritten documents.
3. data set stored and tested for matching the features to detect forgery.
4. The user will have higher accuracy detection in less time consumed.

2.3 Project Overview



3 Similar System Information

Ramy El-Hajj, Laurence Likforman-Sulem, Chafic Mokbel [1] proposed Arabic Handwriting Recognition by hidden Markov Modelling (HMM),The paper motivation is to propose a structural features in the HMM framework which exclusively results in a 9 percent relative improvement in performance. By handwriting feature extraction, in this paper this focus on the baseline of handwriting by sliding window technique for baseline detection, detecting 24 different features vectors per frame some based on baseline, others by position. HMM process divided into 2 modules (training, recognition)modules, segmental Expectation Maximization (EM) algorithm is used for the training module. which is not performed on a character per character basis; rather whole words are used for this training while the character models are shared between the words, IFN/ENIT Database to contain large number of data sets. Resut: Overall, we demonstrated a 17 percent relative reduction in word error rate over the baseline system

Shirin Saleem, Huaigu Cao, Krishna Subramanian, Matin Kamali, Rohit Prasad and Prem [2] Natarajan proposed in Improvements in BBN's HMM-based Offline Arabic Handwriting Recognition System, a handwriting characters recognition by BBN's Hidden Markov Model (HMM), focusing on baseline detection, GSC(Gradient, Structure and Concavity), different strokes detection and extraction, ruled-lines detection and removal. all done by window sliding technique, the local maxima from the horizontal projection profile, 2 modules of HMM phases are introduced (trained, not trained) to calculate the Word Error Rate(WER) of recognition by HMM improvement after every feature extraction technique.

Yuchen Luo, Rui Xia and M. Abdulghafour [3] proposed in Offline Chinese Handwriting Character Recognition through Feature Extraction;The paper mo-

tivation: In this paper, a structural feature-extraction algorithm was developed to obtain feature information of characters. An efficient matching program resulting a high accuracy of recognition. Recognition is done by position matrix and adjacency matrix are all combined by one matrix, chinese characters feature extraction recognition by first applying thresholding by using Ostu algorithm smoothing and noise reduction filters then thinning filters to be able to extract end, start, branch points of the handwriting by using open and closed loops path.Result:Experimental results are accomplished by the use of 1000 random characters to test the effectiveness. The accuracy of the recognition system is significant, The results of experiments show that most characters are successfully recognized.

Hicham El Moubtahij, Khalid Satori and Akram Halli [4] proposed in Recognition of Off-line Arabic Handwriting words Using HMM Toolkit (HTK), HMM standard Viterbi algorithm is used for recognition of Arabic hand-writings. Paper motivation: Hidden Markov Models Toolkit (HTK). This is a robust and simple approach which main advantage is that no prior segmentation of words is needed. This paper focuses on recognizing different skew angels and slant angels by sliding window technique and works on their correction by applying rotation operation, horizontal projection and vertical projection profile.HTK models (HMM) the feature vector with a mixture of Gaussians.In the recognition stage, the Viterbi algorithm is used which searches for the most likely sequence of a character given the input feature vector. For the training data we use Baum–Welch algorithm, a variant of the expectation maximization (EM) algorithm for optimization of the HMM.Result: After many experiments on the Arabic database we achieved a recognition rate of 80 percent.

Dinesh Dileep [5] proposed in A FEATURE EXTRACTION TECHNIQUE BASED ON CHARACTER GEOMETRY FOR CHARACTER RECOGNITION; Paper motivation: this paper has been inspired from the work in the literature explains many high accuracy recognition systems for separated handwritten numerals and characters. Feature extraction done after applying Binarization, Background Noise removal, Skeletonization filters to detect starters, intersections, minor starters by Zoning method which is zoning The image into 9 equal sized windows, feature extraction was applied to individual zones rather than the whole image, after zoning character traversal is applied, each zone extract:

1) Number of horizontal lines.
2) Number of vertical lines.
3) Number of Right diagonal lines.
4) Number of Left diagonal lines.
5) Normalized Length of all horizontal lines.
6) Normalized Length of all vertical lines.

7) Normalized Length of all right diagonal lines.
8) Normalized Length of all left diagonal lines.
Result:This paper proposed a feature extraction technique that may be applied to classification of cursive characters for handwritten word recognition. The method proposed was tested after training a Neural Network with a database of images.

3.1 Similar System Description

Over the past years many system tried to detect Character recognition(OCR) in many languages,and many succeeded and delivered accurate results, and many systems tried to detect forgery of handwritten documents, by pixel matching algorithms.Our systems differs because it delivers something that is new to the market, a system that combines both OCR, Feature extraction based on each individual style, and then apply matching algorithm to detect accuracy based of Arabic handwriting features. In [1], [2] it focused on one of the distinctive feature that the Arabic language has and that's Baseline, done by Sliding window technique; to detect strokes, dots placement, etc. in [3] focuses on starting, ending points of handwriting and detecting it. [4] focuses on extracting gradient, structure, concavity) features. plus other features as intensity, length, slant angels, skew angels) and data set is collected to be trained and tested for accuracy.

3.2 Comparison with Proposed Project

/

Points of comparison	Algorithm used	Accuracy Achived	training sample
Arabic Handwriting Recognition Using Baseline Dependent Features and Hidden Markov Modeling	HMM. Sliding window. Veterbi Algorithm	74.90%	21,500
Recognition of Off-line Arabic Handwriting words Using HMM Toolkit (HTK)	HMM(HTK) toolkit. Sliding window Algorithm Gaussians models SFSA modelling	80.33%	1,819
Offline Chinese Handwriting Character Recognition through Feature Extraction	Ostu Algorithm Smoothing, Thinning filters Open, close loops Algorithm Position,Adjecency Matrix	Not mentioned	3755
Our Proposed System	HMM. Sliding window. Veterbi Algorithm Ostu Algorithm Smoothing, Thinning filters Open, close loops Algorithm + Matching string Matrixes /Knuth-Morris-Pratt	-	>1000

3.3 Screen Shots from previous systems (if needed)

Similar System 1

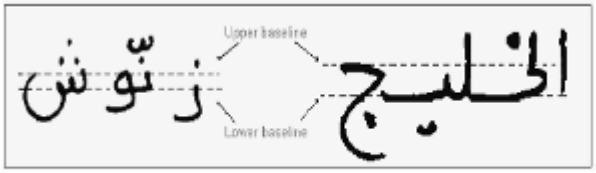
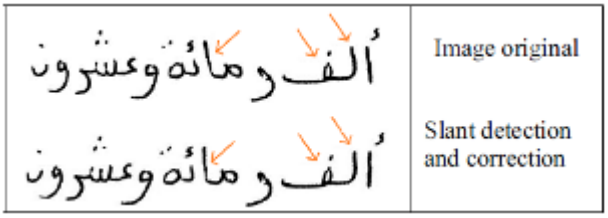


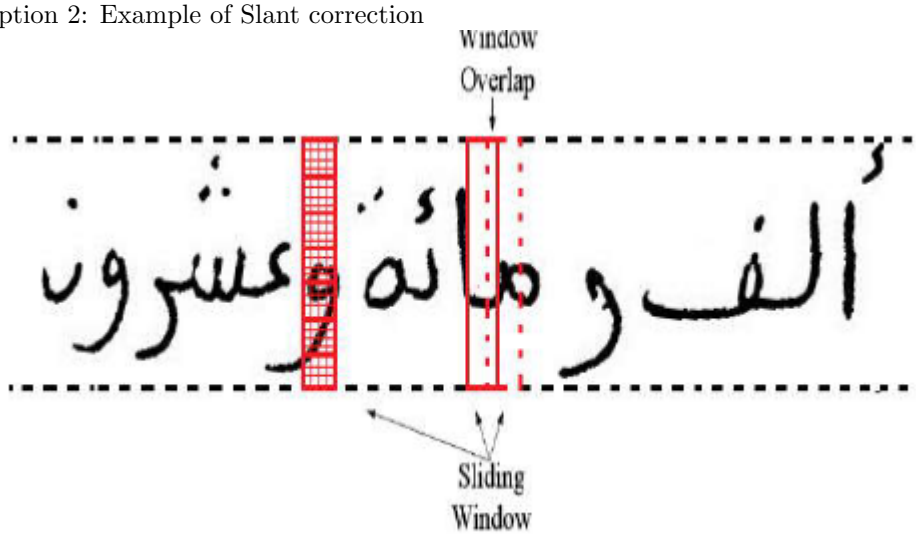
Figure 2. Upper and lower baselines on sample data

Caption 1: Upper and lower baselines on sample Data.

SimilerSystem 2

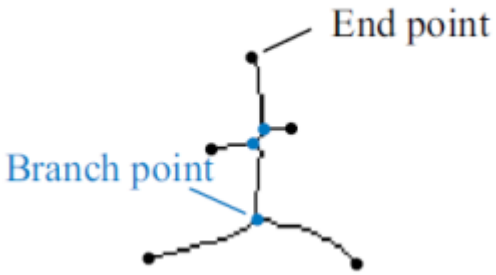


Caption 2: Example of Slant correction



Caption 3: Dividing the line text into windows and cells.

Silmlar System 3



Caption 4: Vertexes of an image, end points are the ones connecting with one branch, branch point are the ones connecting with 2 branches.

4 Project Management and Deliverables

4.1 Tasks and Time Plan

4.2 Budget and Resource Costs

1- HP Scanjet 200 Flatbed Photo Scanner - 1,420 EG

4.3 Supportive Documents

1- The market motivation has confirmed that detecting a forged handwriting *is difficult, takes times, not accurate nor precise, needs skilled professionals*.

2- *No* systems exists in *Egypt* to detect Arabic handwritten forgery

-Head of teller in HSBC
El Hussain Abdoun

Phase 1:
-Collecting information Research:till end of September. -Proposal presentation: 4th of October. -Implementation GUI Desgin:end of November. -Writing SRS: till 15th of November. -SRS Presentstion:15th of November. -Writing SDD:till end of December. -SDD Presentation: 2nd week of January.

5 References

1- Ramy El-Hajj, Laurence Likforman-Sulem and Chafic Mokbel, "Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling" in Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05). PoBox 100 Tripoli, LEBANON, 51687 Reims, FRANCE, 75013 Paris, FRANCE. 2005

2- Shirin Saleem, Huaigu Cao, Krishna Subramanian, Matin Kamali, Rohit Prasad and Prem Natarajan. "Improvements in BBN's HMM-based Offline Arabic Handwriting Recognition System" in Proceedings of 10th International Conference on Document Analysis and Recognition, Cambridge MA, USA. 2009

3- Yuchen Luo, Rui Xia and M. Abdulghafour, "Offline Chinese Handwriting Character Recognition through Feature Extraction", in Proceedings of 13th International Conference Computer Graphics, Imaging and Visualization 2016.

Nanjing, China. 2016

4- Hicham El Moubtahij, Khalid Satori and Akram Halli, "Recognition of Off-line Arabic Handwriting words Using HMM Toolkit (HTK)", in Proceedings of 13th International Conference Computer Graphics, Imaging and Visualization, 2016.

5- Dinesh Dileep, "A FEATURE EXTRACTION TECHNIQUE BASED ON CHARACTER GEOMETRY FOR CHARACTER RECOGNITION",Department of Electronics and Communication Engineering, Amrita School of Engineering, Kollam, Kerala, INDIA.

6- Volker Märgner, Haikal El Abed, Guide to OCR for Arabic Scripts , Springer London Heidelberg New York Dordrecht.

7- Jianying Hu, Michael K. Brown, Senior, and William Turin, HMM Based On-LineHandwriting Recognition, TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 18, NO. 10, OCTOBER 1996, IEEE.

8- Jianying Hu, William Turin and Michael K. Brown, Language modeling using stochastic automata with variable length contexts, Lucent Technologies, Bell Laboratories, 700 Mountain Avenue, Murray Hill, NJ 07974, USA and ATT Research Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974, USA.

9- Alaa M. Gouda' and M. A. Rashwan, Segmentation of Connected Arabic Characters Using Hidden Markov Models, CIMSA 2004 L E E Elm emvtional Conference on Computalional Intelligence for Measurement Systems and Applications Baston, YD, USA, 14-16 July 2004.

10- Shutao Li a,, Qinghua Shen a, Jun Sun, Skew detection using wavelet decomposition and projection profile analysis, College of Electrical and Information Engineering, Hunan University, Changsha 410082, China b Fujitsu RD Center Co., Ltd., Eagle Run Plaza B1003, Xiaoyun Road No. 26, Chaoyang District, Beijing 100084, China.

11- Homayoon S.M. Beigi, Krishna Nathan, Gregory J. Clary and Jayashree Subrahmonia, Challenges of handwriting in arabic, farsi with the same handwriting style, J.T Watson research center, IBM.

12- M. Pechwitz, V. Margner, Baseline Estimation For Arabic Handwritten Words, Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02).