Faculty of
Engineering
Credit Hours System

Cairo University

Pattern Recognition

“Final”

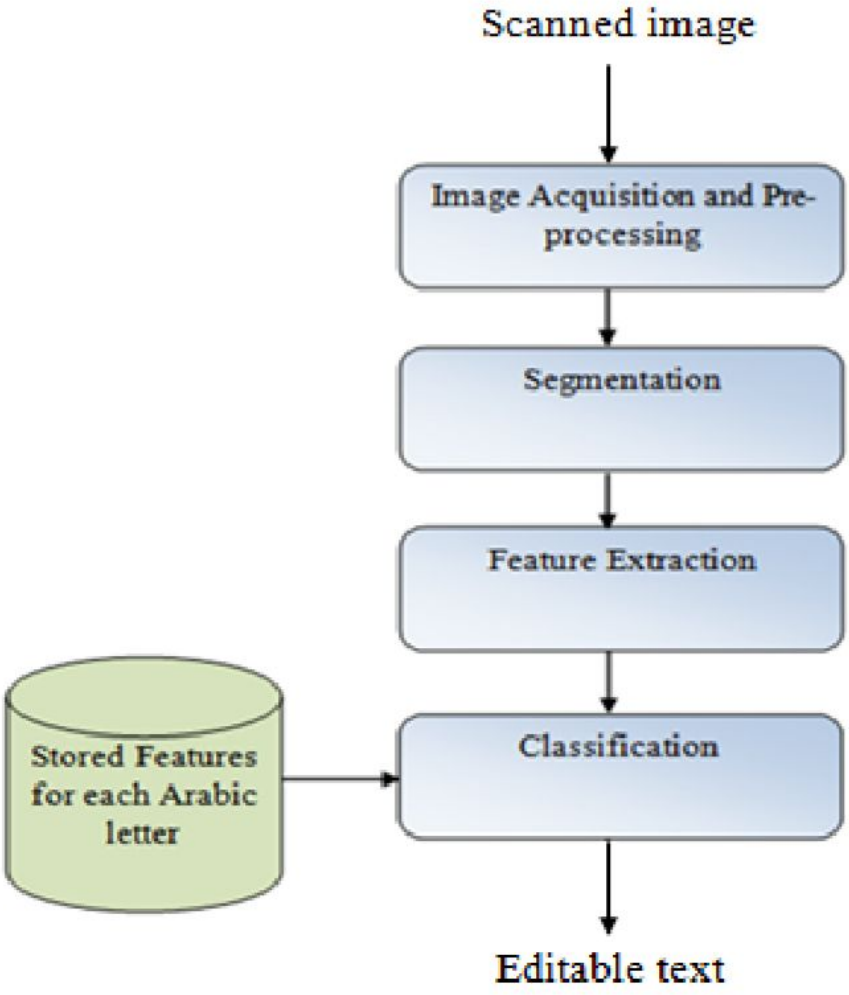
Team No. 14:

Name	ID
Mohamed Bassel Mohamed	1152253
Ahmed Mohamed Khalifa	11517313
Marwan Medhat Gamal	1152030
Mohamed Haitham	1152056

Table of Contents

i.Project Pipeline

Optical character recognition or optical character reader (OCR) is the recognition process of text obtained from media in the form of typed, handwritten or printed text into machine-encoded text form. The text in question may be presented in the form of a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

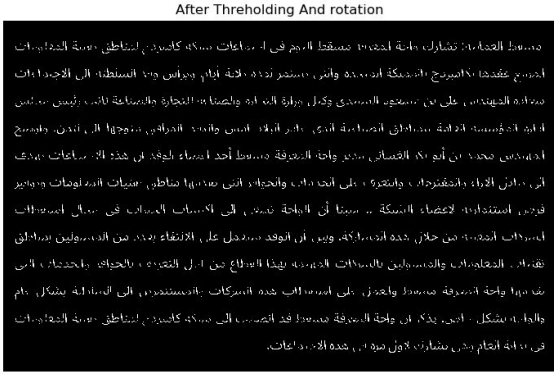


ii.Preprocessing Module:

In preprocessing we calculate the skew angle if the scanned image is rotated and fix it and apply Otsu's thresholding method to convert the scanned image to binary image

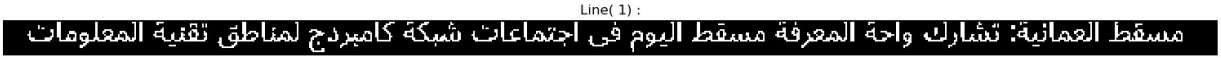
Orignal Image

After Threholding And rotation

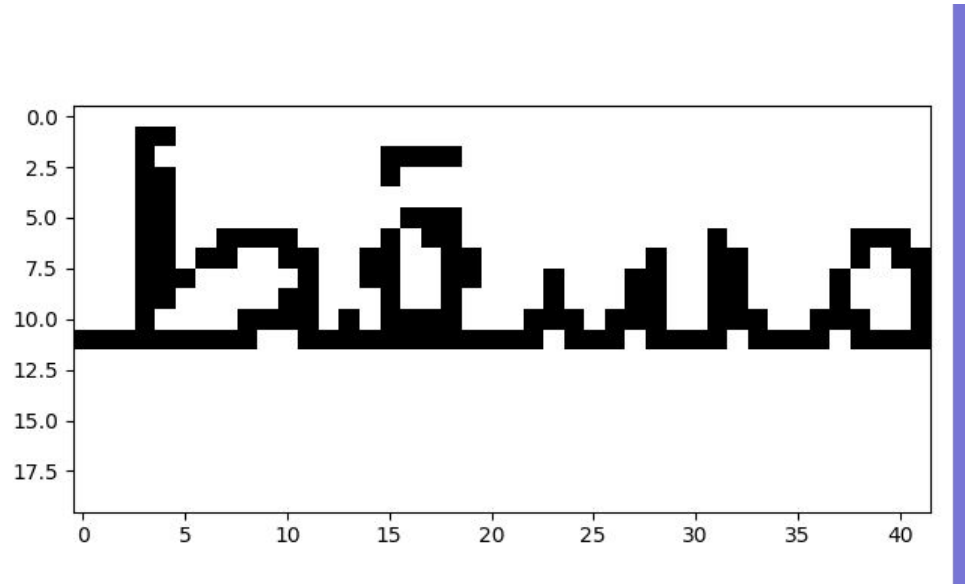


Line Segmentation: we make a horizontal projection and detect the gaps and form the gaps we segment the lines

Line(1) :



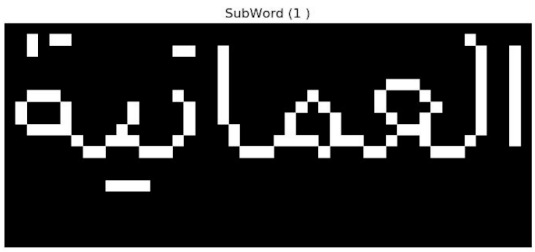
Word Segmentation: after we segment line into separate words by using Vertical projection and deciding on the values between characters

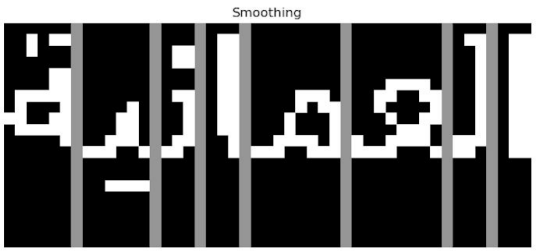


Character Segmentation: First we detect both base line (making horizontal projections and sum each row and pick the most important one) and we detect the Maximum transition line(we count in each row the transition from black to white and pick the row index with maximum transition)

We use the Maximum transition index and iterate on each col and if the transition is detected

We calculate the start and end index as from black to white and from white to black is my end index and then we choose a cut between them





iii. Feature Extraction/Selection Module

We choose those features:

1. Number of connected components
2. Number of holes
3. Height to Width ratio
4. Max transition columns
5. Max transition Row
6. White pixels to black pixels ratio
7. Segment the image to 4 regions and get the white to the black ratio in each region (ratio 1, ratio 2, ratio 3, ratio 4)
8. white pixels in ratio 1 / white pixels in ratio 2
9. white pixels in ratio 3 / white pixels in ratio 4
10. white pixels in ratio 1 / white pixels in ratio 3
11. white pixels in ratio 2 / white pixels in ratio 4
12. white pixels in ratio 1 / white pixels in ratio 4
13. white pixels in ratio 2 / white pixels in ratio 3

iv. Model Selection and Training Modules

After we train our model and making a dataset we choose a Gaussian SVM classifier then Neural Networks achieved much better accuracy , As we try to use different SVM classifiers like linear and Non-Linear Polynomial and Gaussian and Amongst the Gaussian kernel and polynomial kernel, we can see that Gaussian kernel achieved a better prediction rate while polynomial kernel misclassified sometimes. Therefore the Gaussian kernel performed slightly better. However, there is no hard and fast rule as to which kernel performs best in every scenario. but due to some runtime issues.

vi. Performance Analysis Module.

The segmentation performance is almost 80% to 90% through different files

```
Image Word# 571 Text Word#: 571
571 571
Appending to training set with accuracy of char segmentation => 82.13660245183888  % Running Time: 6.3249852657318115 for File apr68.png
Image Word# 614 Text Word#: 614
614 614
Appending to training set with accuracy of char segmentation => 80.29315960912052  % Running Time: 5.140742301940918 for File apr69.png
Image Word# 675 Text Word#: 675
675 675
Appending to training set with accuracy of char segmentation => 80.5925925925926  % Running Time: 6.288101673126221 for File capr7.png
Image Word# 1867 Text Word#: 1867
1867 1867
Appending to training set with accuracy of char segmentation => 82.11033743974289  % Running Time: 18.20384430885315 for File apr70.png
Image Word# 861 Text Word#: 861
861 861
Appending to training set with accuracy of char segmentation => 84.3205574912892  % Running Time: 8.090716123580933 for File apr71.png
Image Word# 1007 Text Word#: 1007
1007 1007
Appending to training set with accuracy of char segmentation => 81.23138033763655  % Running Time: 10.966155767440796 for File apr72.png
Image Word# 350 Text Word#: 350
```

```
Appending to training set with accuracy of char segmentation => 87.47433264887063  % Running Time: 7.9128193855285645 for File capr3.png
Image Word# 88 Text Word#: 88
88 88
Appending to training set with accuracy of char segmentation => 85.22727272727273  % Running Time: 1.219120740890503 for File apr34.png
Image Word# 356 Text Word#: 356
356 356
Appending to training set with accuracy of char segmentation => 85.1123595505618  % Running Time: 4.341842412948608 for File apr35.png
Image Word# 613 Text Word#: 613
613 613
Appending to training set with accuracy of char segmentation => 83.68678629690048  % Running Time: 8.05824089050293 for File apr36.png
Image Word# 154 Text Word#: 154
154 154
Appending to training set with accuracy of char segmentation => 79.87012987012987  % Running Time: 1.5755507946014404 for File apr37.png
```

Neural Network

We achieved an accuracy of 95% on our testset.

```
C:\Users\student\Desktop\Team_14_Cr\OCR-for-Arabic-Scripts\03-Source Code>python edit.py output Expected
capr3.txt: 107
Total distance = 107
Average Accuracy = 95.11%
```

SVM Gaussian

We achieved an accuracy of 91% on our testset.

```
D:\_Marwan\OCR-for-Arabic-Scripts\OCR-for-Arabic-Scripts\03-Source Code>python edit.py output expected
capr3.txt: 183
Total distance = 183
Average Accuracy = 91.64%
```

vii. Other developed modules.

Classification

We used PyTorch with different architectures to try which one worked best, we used a 5M datapoint dataset, with 2 hidden layers, using the cross entropy loss function. We got an accuracy of 97% on our testset. We trained the project using Microsoft Azure instance, from features extracted from the images. And used Google Colabs to use our training set directly from Drive.

viii. Enhancements and Future work.

In our Project, the major obstacle is the segmentation process because of certain characteristics of the Arabian language. This is the stage where the maximum error occurs and therefore as future work we would want to concentrate our work on this stage to attain maximum accuracy in the segmentation phase which deals with the problems of overlapping and characters.

viv. Team work distribution.

Mohamed Bassel : Character segmentation and integration

Marwan Medhat : Feature Extraction

Mohamed Haitham : Preprocessing, Line Segmentation and Classification(SVM)

Ahmed Khalifa : Word Segmentation and Classification(Neural networks)

References

1. Optical Character Recognition of Arabic Printed Text Safwa Taha, Yusra Babiker, and Mohamed Abbas
2. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
3. Printed Arabic Optical Character Recognition using Support vector machine
4. https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/