Arabic Handwriting Recognition

PREPARED FOR

Dr. Patrick Saoud

PREPARED BY

Ahmed Aljmiai

Fatemah Alshaikh

Jennan Sowayan

Problem Statement

The Kingdom's vision 2030 promotes digital transformation. The Ministry of Justice especially has the challenge of having an enormous archive of official handwritten documents like property deeds, marriage certificates, birth certificates. These documents need to be digitized and stored in a relational database to solve deteriorated documents, fade-out ink, and ease of accessibility and prevent loss or forgery of documents.

Questions:

- Is image processing possible ?
- Can we victorize the images ?
- Is building a classifier that recognizes Arabic characters with all possible forms of the letter e.g. ‘ـذ ، ذ’ doable ?
- Can handwritten documents get digitized into a typed document?

Data Description

There are two datasets that serve the purpose of our project. The datasets ‘**Hijja2**’ and **‘Ahdd1’** are described as follows:

1. Ahdd1 [1]:

The MADBase is a modified Arabic handwritten digits database with 60,000 training and 10,000 test images. It was handwritten by 700 different people. Each digit (from 0 to 9) was written 10 times.

2. Hijja2 [2]:

Hijaa is a dataset for handwritten Arabic letters collected from Arabic-speaking school children. 591 participated in the making of the data with 37933 training and 9501 testing images. The dataset includes all possible positions of each letter, including the hamza.

Tools

1. **Jupyter Notebook**
2. **Python Libraries:** Pandas, NumPy, matplotlib, seaborn, sklearn, imageIO, Computer Vision library (openCV)

Project Workflow

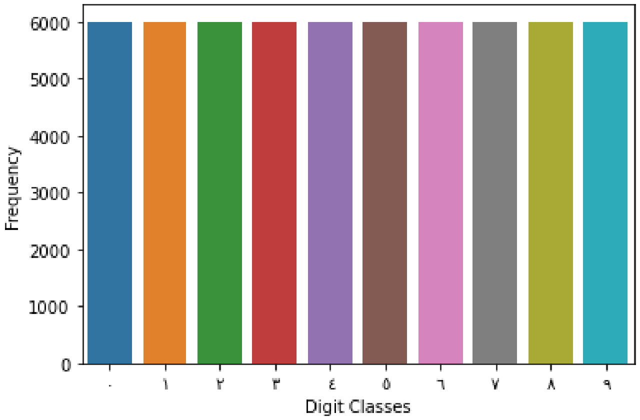
For the sake of experimenting with our classification project, we will start by training our model using the numbers dataset **‘Ahdd1’;** because it has fewer number of classes which will make it easier to evaluate the model performance and spot any shortcomings, and because it has balanced classes that will be considered equally when training the model.

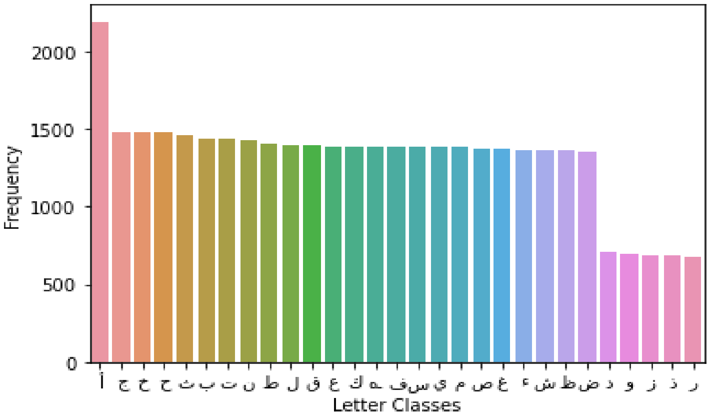
1. **Exploratory data analysis and metric selection:**
    a. Descriptive statistics
    b. Vectorize the image files into .csv files
    c. Normalize pixel values between -1 to 1 instead of 0 to 255
    d. Check for solutions to remove the white space from the letter images
    e. Select the relevant metrics for evaluating the models performance such as Accuracy score, Recall, Precision or F1-Score
2. **Baselining:**
    a. Fit a sample of the observations into KNN model, trying multiple K values
    b. Use the classification evaluation method from sklearn library, to evaluate the model performance and check if the chosen metric is relevant to the problem statement
3. **Establish a validation and testing scheme:**
    a. For each model trial, we will be performing the relevant validation method
    b. The test set will be held out to generate predictions against all models, in order to generalize it on unseen data
4. **Expand and refine model:**
    a. Try out several approaches to solve for our imbalanced classes such as image augmentation
    b. Building on top of the KNN model, we will be trying out other models such as random forest regressor, gradient boosted trees, logistic regression, and naive bayes
    c. Rescale data and/or tune hyperparameters to solve for overfitting or underfitting models to get the best fit

5. **Finalize, test, and interpret your model:**
    a. **Finalize and test:** test the models against each other and choose the best performing model according to the chosen evaluating metric
    b. **Interpret:** gather insights about the chosen model

MVP Goal

The two graphs below show each class's data distribution in the two datasets, and we can see that the letters classes are not balanced, which doesn’t work well in a classification project unless solved.





The MVP should recognize a sample of the Arabic digits e.g. the digits ‘ ٢‘ ,’ ١’.

Future Work

- Incorporate word segmentation in the preprocessing phase, so the model could recognize written documents
- Evaluate the performance of the chosen model on different datasets

4. References

Datasets

[1] Mohamed Loey (2017). Arabic Handwritten Digits Dataset, Ver.# 3. Retrieved Oct. 27th, 2021 from https://www.kaggle.com/mloey1/ahdd1.

[2] Najwa Altwaijry, Monera Al-Megren, Haya Al-Shumisi, Lamya Al-Arwan, and Isra Al-Turaiki (2019). *Hijaa2*: handwritten Arabic letters of Arabic-speaking school children between the ages of 7 and 12 in Riyadh, Saudi Arabia. Ver.# 2. Retrieved Oct. 27th, 2021 from https://github.com/israksu/Hijja2.git.