**contact information**

* Alexander boyev 314393158

alexabo4@ac.sce.ac.il

* Moshe Faerman 204469449

moshesh81@ac.sce.ac.il

**Description**

a python script that accepts user's dataset path, in this project the dataset “hhd\_dataset” [1], represents 5073 pictures of the Hebrew alphabet letters. The Hebrew alphabet contains 27 letters, so each letter represent a class and contains about ~188 image of handwritten letters.

The program will execute the following steps:

* The program will read the data into a dict variable structure, each “key” will represent the letter and the “value” will contain a “list” of images of that specific letter.
* Pre-processing step will “grayscale” each image, we will pad each image with white borders and make it squared, after that we will resize the image to 32x32
* We will split the data randomly but equally distributed to three categories: Train, Validation and Test. With the sizes of 80% : 10% : 10% accordingly.
* We will use sklearn library for the KNN model and use it with “Euclidean distance” metric. The model will train on the “train” set, afterwards we will try to evaluate the model with the “validation” set and find the best K ( in range 1 to 15 with steps of 2) that provides the highest accuracy.
* We will save the best model, accuracy and the processed data into a “pickle” file to shorten the progress in case we would like to run the program again.
* The best saved KNN model could be loaded, now we can fit the “test” set and report the results.
* We will generate a “result.txt” file that will contain:
  + Best K value
  + The accuracy of each letter.
* We will also provide a confusion matrix, that will be saved on scv file.

**Environment**

The script developed and ran on Windows 10 pro edition, PyCharm 2019.3.3 (Community Edition) with Python 3.7 interpreter.

**Requirements**

To run the program, you will need to install the following libraries:

**import** cv2  
**import** os  
**import** math  
**import** random  
**import** numpy **as** np

**import** pandas **as** pd  
**from** sklearn.neighbors **import** KNeighborsClassifier  
**from** sklearn.metrics **import** accuracy\_score,confusion\_matrix  
**import** pickle

**import** sys**import** matplotlib.pyplot **as** plt  
**import** itertools

**Data assumptions**

The data contains 5073 images of 27 Hebrew handwritten letters, even though 5073 images sounds allot, for 27 classes its not. Even if we split equally and train the model on 80% of the data its roughly 150 images per class. For such classification problem with that high number of classes we will need much bigger dataset, even tough we achieved 63.5% accuracy for K=9. Also, the accuracy per letter is overall decent.

Chart

Description automatically generated with medium confidence

We can see that each letter had roughly 18 samples to test, and only few of them ( like letter 20 and 22 ) got very low accuracy.

**How to Run Your Program**

**Preconditions:**

1)install all the required models.

2)copy the dataset [1] into the project folder.

The assignment delivered with the pickle file, so you could just load the model from the pickle and evaluate the “test” set and get the result output. Otherwise just cancel the comments and make functions available again, this way the script will make all the pipeline from scratch. Load -> preprocess -> Split -> Evaluate -> Test -> Results.

Graphical user interface, text, application

Description automatically generated

There are two options to run the program:

1. use the terminal and type: “knn\_classifier.py .\hdd\_dataset” while knn\_classifier.py is script name and .\hdd\_dataset is the path.
2. Press “Run” in the PyCharm environment or any other python script IDE.

**References**

[1] I. Rabaev, B. Kurar Barakat, A. Churkin and J. El-Sana. The HHD Dataset. The 17th International Conference on Frontiers in Handwriting Recognition, pp. 228-233, 2020.