Cairo University

Faculty of Engineering

Computer Engineering Department

Third year

PATTERN RECOGNITION & NEURAL NETWORKS

Handwritten Based Gender Classification

Team#10 Members

Name	Section	B.N.
Ayman Mohamed Reda	1	20
Mostafa Mohamed Elgendy	2	27
Ghieath Omar Saleh	2	7
Nour Aldin Mostafa	2	34

May 2022

Literature Review

Problem Definition:

The challenge of automatically classifying gender based on handwritten samples allows distinguishing between male and female writers' samples. According to several psychological research, we can differentiate between the two genders' writings due to various differences; in average, female handwritings are more uniform and regular, whereas male handwritings are spikier and slanted.

Implemented Techniques:

Preprocessing:

Mainly two kinds were used, (each feature used exactly one)

Method one:

1- Sharpen the image

2- Convert to gray

3- Blur using gaussian

4- Binary threshold using histogram analysis

5- Dilate then erode with a wider kernel to connect letters

Method two:

1- Sharpen the image

2- Double its height and width

3- Gaussian blur

4- Binary threshold using histogram analysis

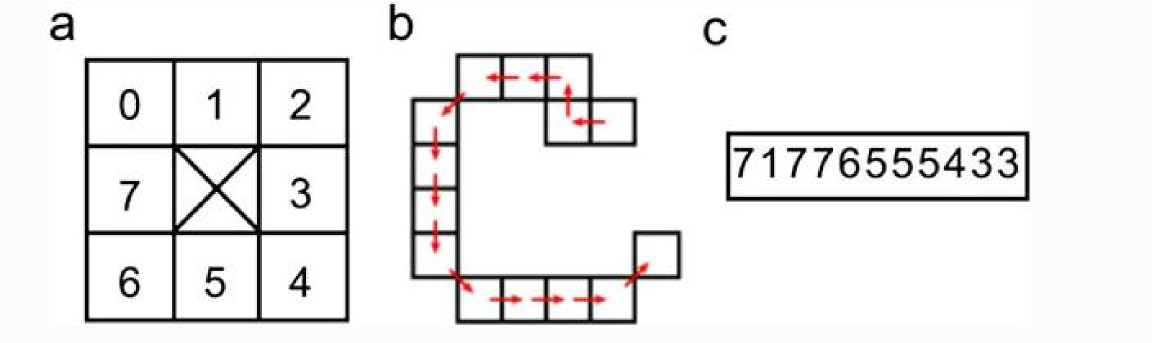
Hinge Feature: (uses preprocessing method two)

The contour-based feature was created to capture the curvature of the document images' ink trace, which is thought to be particularly discriminatory across handwritings. The Hinge, Quill-Hinge, and Delta-n Hinge features are the best contour-based features documented in the literature. The probability distribution of orientations of two legs of the produced "hinge" based on edges or contours collectively attached at a current pixel is the Hinge feature. It can also be defined as the probability distribution of orientations of two contour fragments attached at a common pixel. The Hinge feature has two parameters: the number of angle bins (p) and the leg length (r). In our implementation, we set p = 40, r = 25.

Chaincode Feature: (uses preprocessing method one)

Firstly, we find countours of the image, then we traverse through their pixels while saving the direction we go through, each direction is then mapped to a number, the resulting number is then called ‘chain code’ of this contour.

Chain codes are then stored in a list for each image.



This clearly captures detailed curvatures of the writing.

The features extracted and used are PDFs of patterns encountered, ex:
PDF[‘5’] -> probability of ‘5’ in all chin codes of image

PDF[‘53’] -> probability of ‘5’ followed by ‘3’ in all chain codes

Etc..

This is done to all possible patterns up to a length of 3.

Total number of features:

8 for length 1

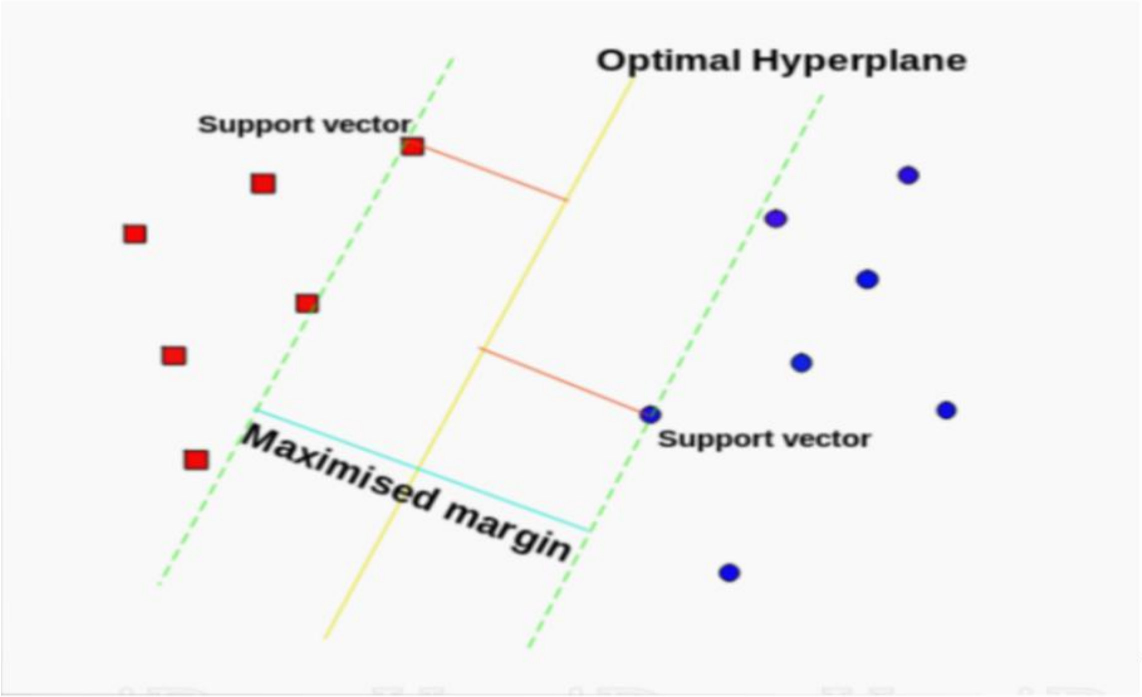
64 for length 2

512 for length 3

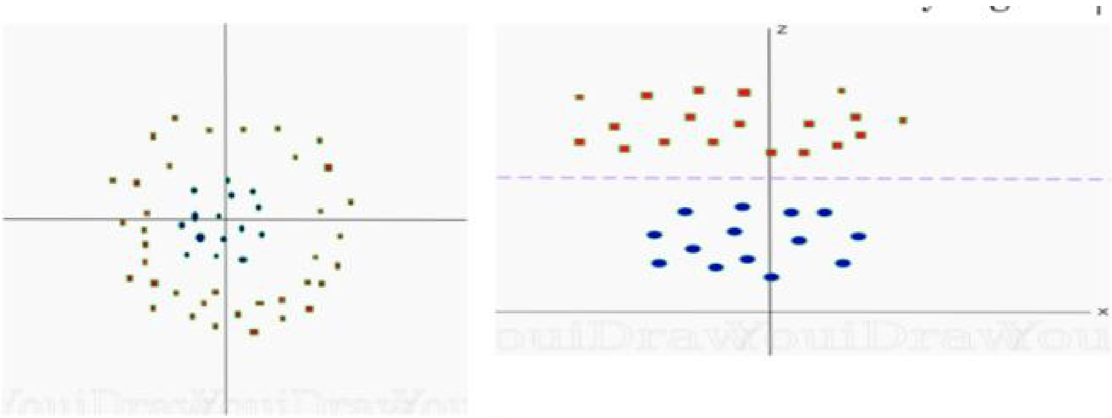
I.e.: a total of 584 numbers compose the chain code feature vector

Support Vector Machine Classifier:

SVM is a supervised technique tries to make a decision boundary in such a way that the separation between the two classes(that street) is as wide as possible.



Separation is not necessarily linear, a kernel could be non-linear depending on parameters given.



SVM parameters:

Kernel = rbf (default)

Radial Basis Function, is a kernel that allows curving in the separator, this was the best choice experimentally.

C = 10

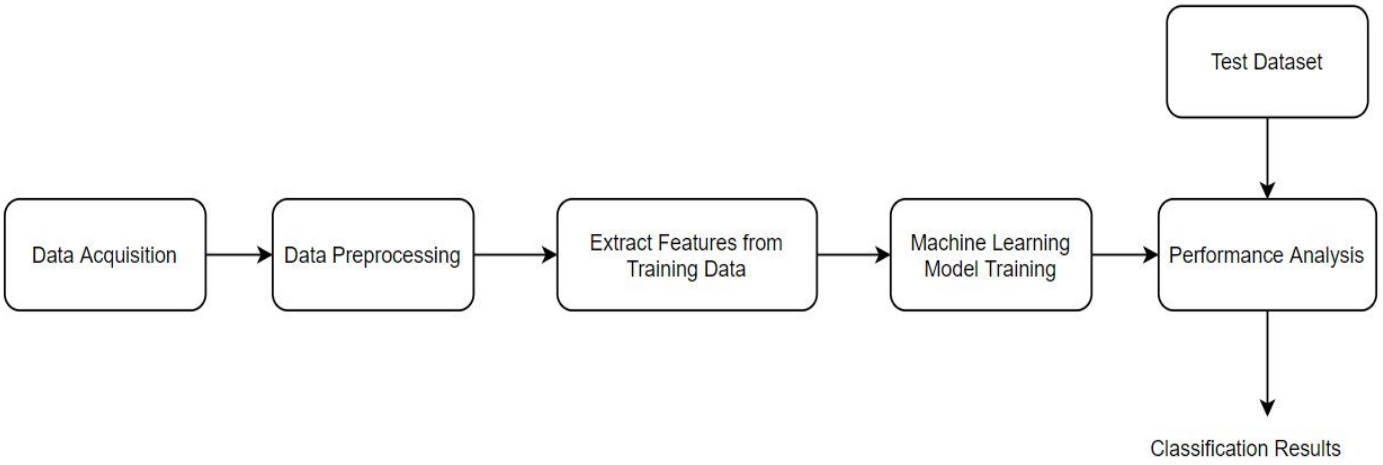
This parameter controls the penalty for a miss-classification, higher C will lead to a higher variance, at the classifier tries to minimize total number of penalties.

Another Techniques:

The Histogram Oriented Gradient (HOG) and the Local Binary Pattern (LBP) and grid features were utilized to solve the problem of gender classification based on handwriting where local features were favored over global ones. Another approach utilized the same set of features in addition to a feature obtained from a segmentation-based fractal texture analysis (SFTA) and features extracted from grey level co-occurrence matrices (GLCM).

Discrete Wavelet Transform (DWT) and Symbol Dynamic Filtering (SDF) are two global characteristics that can be used to tackle the problem. Using discrete wavelet transformations, the handwritten document is first split into sub-bands. A maximum entropy partitioning is then performed to these sub-bands, resulting in data sequences.

Project Pipeline



Data Acquisition:

We used our own built dataset (CMP_23 dataset) for training and testing. The pipeline begins with reading all the male and female images (note that actually we are only saving the images' paths in the memory not the images because the dataset size is huge, and we couldn't store the whole dataset in the memory).

Data Preprocessing:

The input to the data preprocessing step is the path of the image to be preprocessed then the image is read and preprocessed and we pass it to the feature extractor (so that we don't have to store all the preprocessed images).

Feature Extraction:

The features (hinge & chaincode features) are extracted from the preprocessed image and stored in numpy arrays. We save the numpy arrays into external .npy files in the “Features” directory (.npy files are used to store numpy arrays in binary format efficiently) so that we can load them directly into numpy arrays without waiting for feature extraction each time we run the program. After loading the feature vector, we use the function train_test_split to split this feature vector into training dataset and test dataset with a ratio given as a parameter to the function (we split the dataset for developing purposes only however in the delivered program the training data is the whole CMP_23 dataset).

Machine Learning Model Training:

After splitting the feature vector into training and testing vectors, the training vector is passed to the model to train it.

Performance Analysis:

The performance of the classifier was analyzed by using the dataset generated from the train_test_split function (which randomizes the selection of the training and test datasets) in which the whole dataset was divided by the percentages (50%, 50%) across the training dataset and the testset. Then, the classification accuracy was averaged along 100 runs of the classifier. The results and used techniques are shown in the following table.

Used Techniques	Features Extracted	Average Classification Accuracy	Percentage of Test Dataset
SVM classifier	Hinge Chaincode	81.7%	50%
SVM classifier PCA	Hinge Chaincode	60%	50%
SVM classifier	Hinge Chaincode COLD	67.77%	50%
Random Forest Classifier	Hinge Chaincode	76%	50%
SVM + Random forest	Hinge Chaincode	79.84%	50%
SVM + KNN + Random forest	Hinge Chaincode	79.25%	50%
SVM + 40 males, 70 females from ICDAR	Hinge Chaincode	74%	50%

References

[1] https://link.springer.com/chapter/10.1007/978-3-030-51935-3_25

[2] https://ieeexplore.ieee.org/document/6977065

[3] Automatic prediction of age, gender, and nationality in offline handwriting, BY Somaia and Abdelaalie

[4] SVC Parameters When Using RBF Kernel