

Cairo University Computer Engineering Department

Faculty of Engineering Third year

**pattern recognition & Neural networks**

**Handwritten Based Gender Classification**

**Team#10 Members**

|  |  |  |
| --- | --- | --- |
| **Name** | **Section** | **B.N.** |
| Ayman Mohamed Reda | 1 |  |
| Mostafa Mohamed Elgendy | 2 | 27 |
| Ghieath | 2 |  |
| Nour Aldin Mostafa | 2 |  |

**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:**

**Hinge Feature:**

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:**

**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.

Diagram

Description automatically generated**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% |

**References**

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

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