Writer Identification

Mentor:
Dr. Gaurav Harit

Made By:
Aakash Hasija(Ug201310001) - CSE
Amit Jain(Ug201310005)- CSE

Project Overview:

In this project we will be building machine learning models for writer identification using two approaches:

1. Text Independent Approach (**Present Work**)
   We did this via the 3 steps mentioned below:
   a. Pre-processing(Segmentation of image)
   b. Feature Extraction(via clustering)
   c. Identification(via correlation similarity measure)
2. Text Dependent Approach (**Future work**)
   In this we will be using character recognition for feature extraction and then identify writers based on similarity of each character that the writer has written.

Description of Important modules:

1. **Preprocessing:** For preprocessing of the data ,we have carried out a component by component division of the text.For each connected component in the text,we fix the vertical origin and move each window from left to right to find the first text (black) pixel.The figure shows division of the word 'advertisement' into subimages using the above mentioned method.We tried 2 window sizes of

   i. 30x30 ii. 40x40
2. **Feature extraction:** We obtained an average of **3000** images per writer from step 1 and there were **78 writers** in our database. Now our next task was to extract features for every writer from these sub images. We did this by clustering similar images into 1 class and creating a fixed number of classes for every writer. We used K-means clustering for this purpose and every pixel was used as feature for the clustering.
3. **Identification:** For Identification we defined a similarity measure called correlation similarity measure between the handwritten document D and an unspecified handwritten document T by the following relation:

$$SIM(D,T)=\frac{1}{card(D)}\sum_{i=1}^{card(D)}\underset{y_j \in T}{Max}(sim(x_i,y_j))$$

Where sim(x,y) is : $sim(x,y) = \frac{n11 \times n00 - n10 \times n01}{\sqrt{(n11+n10)(n01+n00)(n11+n01)(n10+n00)}}$

n11: Number of corresponding pixels that are 1 in both binarized images
n00: Number of corresponding pixels that are 0 in both binarized images
n01: Number of corresponding pixels that are 0 in first and 1 in second binarized image
n10: Number of corresponding pixels that are 1 in first and 0 in second binarized image

Summary of Result:

We used **IAM dataset** to train and test our machine learning model.After Noise removal, our database had images of handwritten text of 78 writers.We preprocessed the data through segmentation of image through which we obtained approximately 3000 images per writer. On these sub-images we did K-means clustering with number of clusters as **20**, **30** and **50** and the **final accuracy** in the three cases using correlation similarity measure and euclidean distance for classification is as shown in fig.1.

We have also generated 3-D plots(as shown in fig.2) showing value of correlation similarity measure for every pair of author (x,y) such that *x* ∈ *Training Set* and *y* ∈ *Test Set*

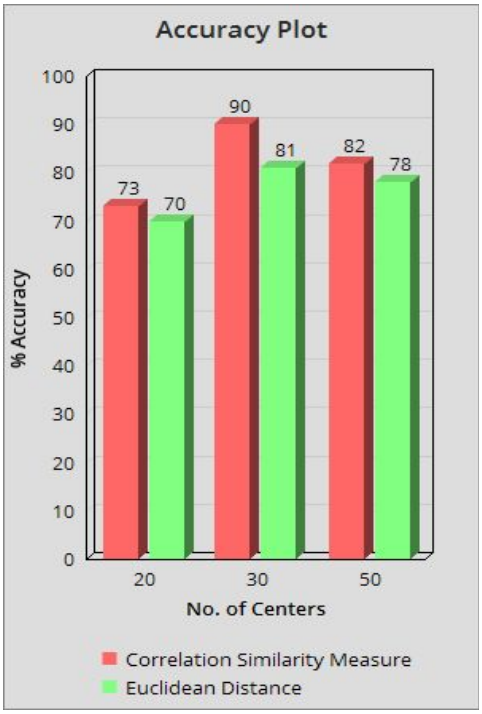

Fig 1

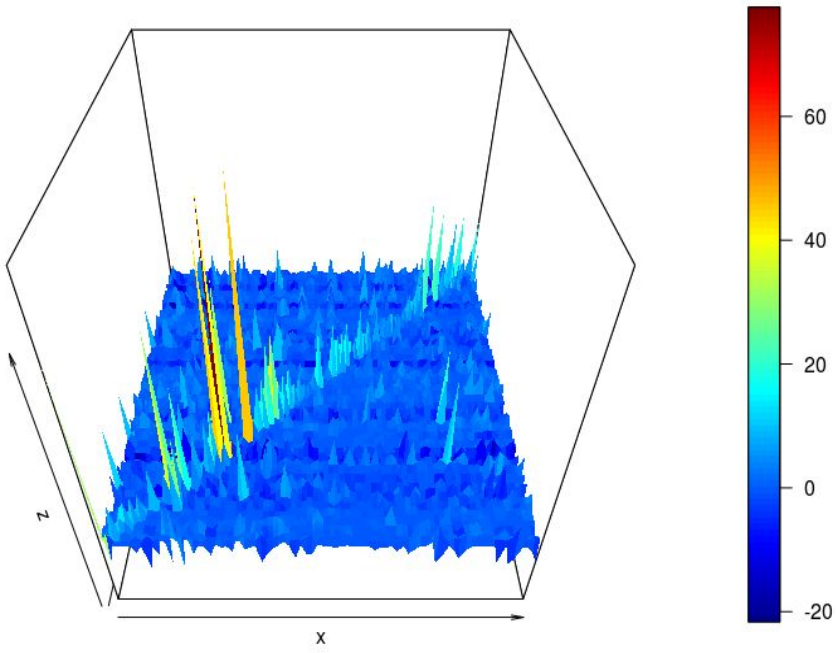

Fig 2.

The **peaks at the diagonals** of the plot in fig.2 represent the **high value of correlation similarity measure** between the training data and test data of **same author**, thus proving our similarity measure to be effective.