M. Tech (cs)

Writer recognition by analyzing handwritten documents and by using neural network classifier

A dissertation submitted in partial fulfillment of the requirements for the M. Tech(cs) degree on the Indian Statistical Institute

By

Sanjoy Dey

Under supervision of …..

I N D I A N   S T A T I S T I C A L   I N S T I T U T E

203, B. T. Road, Kol - 108

Key Words: LVQ, K-mlans, Bounding box, Word-extraction, Line-count, Writer identification

Chapter 1

Introduction

For many centuries handwriting has been used to identify an individual. The hypothesis about handwriting being a personal biometric relies on the fact the process of handwriting is an unconscious act learnt over time, and some pen movements are invariant and not easily changed when an attempt at forgery or disguise is made. When doubts of authenticity of handwriting arise, forensic document examiners are asked to conduct an analysis of the questioned document. They seek characteristics of handwriting (features) that are consistent in a person's normal writing by analysis of shapes and structure of the handwriting. While other branches of forensic science, such as DNA analysis, have been explained and proven by experimental analysis, the forensic techniques used in handwriting analysis have far less scientific support. The methods used by the examiners are intutively reasonable and have been derived from experience. It is the credibility of the document examiner that has been a key basis in a court of law rather than scientific basis of the techniques.

In recent cases the scientific acceptability of forensic analysis of handwriting has been successfully challenged. To provide a scientific support for the handwriting analysis pattern recognition techniques have been employed. It has been demonstrated that handwriting indeed can be used to identify an individual with high accuracy. However, it has not been shown that writers can be distinguised using the techniques forensic document examiner use in their analysis.

A writer can be characterized by his own handwriting. The problem of writer identification arises frequently in the court of justice where one must come to a conclusion about the authenticity of a document (e.g. a will). It also arises in banks for signature verification. In order to come to a conclusion about the identity of an unknown writer, two tasks must be considered:

1. The *writer identification*, where handwritten samples are retrieved from a database and the query sample is taken to find the writer of the query document.

2. The *writer verification*, where two samples of handwriting are taken and a conclusion is drawn whether they are written by the same writer or not.

Previous work: Most work in the writer recognition field has concentrated on signature verification, since signatures typically present more individuality. However, in many cases signatures are not available for analysis, only words or characters. Word based analysis work started with Steinke et al. [4]. Zois et al. [5] used features obtained from the morphological transformation of thinned word images to answer the different writer/same writer question using one single word.

Overview of Study: In case of handwriting two types of variation can occur: within writer variation and between writer variation. The first type of variation occurs within a writer's hanwritten document. The second type of variation occurs between handwritten documents of two different writers. The goal of our study was to establish the intuition that within writer variance is less than the between writer variance . The study have four phases: data collection, preprocessing, feature extraction, and recognition . The first three phases was done by $Tripathy$[11]. In the recognition phase he used KNN classifier, whereas I have used Neural Classifier. In the data collection phase handwriting samples of different writers were colllected. Preprocessing was done on the images so that the suitable features can be extracted. This phase have three steps: noise removal, getting line count and extracting words boundaries. In feature extraction phase required features of each word of a document were extracted from the preprocessed image for discriminating two different writing style. Recognition phase helps to prove that within writer variance is less than between writer variance.

Chapter 2

Preliminaries

A digital image is a representation of a two-dimensional ($Z \times Z$ where $Z = \{0, 1, 2, ...\}$) image as a finite set of nonnegative digital values (pixels) obtained by sampling the continuous domain along two orthogonal directions. The digital image contains a fixed number of rows and columns of pixels. Pixels are the smallest individual elements in an image, holding quantized values. we consider only binary pictures. The value of foreground point or object point in Z^2 is 1, and the value of background point in Z^2 is 0.

2.1 Mathematical morphology:

Sets in mathematical morphology represent objects in an image. In our work we have considered only binary morphological operators which deal with pixel value 1.The two primary operations are: dilation and erosion[1]. Other operations are formed by some combination of these two operations. Dilation and erosion are constructed from translation, set union and set intersection. Let A and B are subsets of Z^2 and t be a pixel of Z^2 then

Translation:

$A_t = \{c \in Z^2 | c = a + t \text{ for some a} \in A\}$

Dilation:

$A \oplus B = \{c \in Z^2 | c = a + b \text{ for some a} \in A \text{ and for some b} \in B\}$

Erosion:

$A\theta B = \{c \in Z^2 | c + b \in A \text{ and for every b} \in B\}$

Opening:

$A \circ B = (A\theta B) \oplus B$

Closing:

$A \bullet B = (A \oplus B)\theta B$

the set A is image object and set B is called structuring element(SE).

2.2 Thining:

Algorithm:

The method consists of successive passes of two basic steps applied to the contour points of the given regions, where, a contour point is any pixel with value 1 and having at least one 8-neighbour valued 0. With reference to the 8-neighbourhood notation shown below, step 1 flags a contour point p_1 for deletion if the following conditions are satisfied:

$a)$ $2 \leq N(p_1) \leq 6.$ and $T(p_1) = 1$,

$b)$ $p_2.\ p_4.\ p_6 = 0$,

$c)$ $p_4.\ p_6.\ p_8 = 0$

where $N(p_1)$ is the nunber of nonzero neighbors of p_1; that is,

$$N(p_1) = p_2 + p_3 + \ldots + p_8 + p_9.$$

p_9	p_2	p_3
p_8	p_1	p_4
p_7	p_6	p_5

and $T(p_1)$ is the number of 0-1 transitions in the ordered sequence $p_2, p_3, \ldots, p_8, p_9, p_2$.

In step 2, the point p_1 is flagged for deletion if

$a)$ $2 \leq N(p_1) \leq 6.$ and $T(p_1) = 1$,

$b)$ $p_2.\ p_4.\ p_8 = 0$,

$c)$ $p_2.\ p_6.\ p_8 = 0$

After Step 1 has been applied to all border points in image, those that were flagged are deleted (changed to 0). Then step 2 is applied to the resulting image in the same way. The process is repeated until no more pixels are deleted from the image.

2.3 Corner detection and Facet model:

The facet model[6] describe corner detection. A corner then is said to occur when two edges meet at a certain angle or when an edge direction is changing very rapidly. The existence of a corner is associated, thus, with

1. there exists an edge.
2. there are significant changes in edge direction.

Solving the direction change of an edge in the direction of α, using facet model. We get
$\theta'_\alpha(0,0) = \frac{-2(K_2^2K_6 - K_2K_3K_5 + K_3^2K_4)}{(K_2^2+K_3^2)^{\frac{3}{2}}}$
The corner detection algorithm tests each pixel (0,0) of an image for

1. (0,0) is an edge point.
2. $\theta'_\alpha(0,0) > T$ for a given threshold T.

2.5 Connectivity number:

Connectivity numbers[1] are assigned to the pixels which belong to the domain of support. Purpose of this number is to show how a pixel of domain of support is connected to its like neighbors. Though we call it *connectivity number*, it is actually a label and has no arithmetic property. Connectivity number operator associates with each pixel belonging to domain of support one of the six different values: 5 values for boundary pixels and one value for the interior pixels. An example is given below.Table 2.1 shows meaning of each connectivity number.

```
  XXXXX   X             13221   0
   X                     2
   X                     2
 XXXXXXXX              12422222
   X    X                2    2
   X    X                2    1
  XXXX                  1211
  XXXX                  1551
  XXXX X                1551 0
  XXXX                  1111

Binary Object         Connectivity Number
    (a)                       (b)
```

Number	class	Meaning
0	Boundary pixel	Isolated
1	Boundary pixel	Edge
2	Boundary pixel	Connecting
3	Boundary pixel	Branching
4	Boundary pixel	Crossing
5	Interior pixel	interior

Table 2.1

8 connectivity case where b_s is source binary image, K_i's are masks.

$$C(r,c) = \max\{\sum_{i=2}^{9} (b_s \theta K_i)(r,c) - \sum_{i=10}^{17} (b_s \theta K_i)(r,c), 5(b_s \theta K_i)(r,c)\}$$

$$\begin{array}{|c|c|c|}\hline 1&1&1\\\hline 1&1&1\\\hline 1&1&1\\\hline\end{array}$$

$\boxed{K_1}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&1&1\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_2}$

$$\begin{array}{|c|c|c|}\hline 0&1&0\\\hline 0&1&0\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_3}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 1&1&0\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_4}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&1&0\\\hline 0&1&0\\\hline\end{array}$$

$\boxed{K_5}$

$$\begin{array}{|c|c|c|}\hline 0&0&1\\\hline 0&1&0\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_6}$

$$\begin{array}{|c|c|c|}\hline 1&0&0\\\hline 0&1&0\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_7}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&1&0\\\hline 1&0&0\\\hline\end{array}$$

$\boxed{K_8}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&1&0\\\hline 0&0&1\\\hline\end{array}$$

$\boxed{K_9}$

$$\begin{array}{|c|c|c|}\hline 0&1&1\\\hline 0&1&1\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_{10}}$

$$\begin{array}{|c|c|c|}\hline 1&1&0\\\hline 1&1&0\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_{11}}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 1&1&0\\\hline 1&1&0\\\hline\end{array}$$

$\boxed{K_{12}}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&1&1\\\hline 0&1&1\\\hline\end{array}$$

$\boxed{K_{13}}$

$$\begin{array}{|c|c|c|}\hline 0&1&1\\\hline 0&1&0\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_{14}}$

$$\begin{array}{|c|c|c|}\hline 1&0&0\\\hline 1&1&0\\\hline 0&0&0\\\hline\end{array}$$

$\boxed{K_{15}}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&1&0\\\hline 1&1&0\\\hline\end{array}$$

$\boxed{K_{16}}$

$$\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&1&1\\\hline 0&0&1\\\hline\end{array}$$

$\boxed{K_{17}}$

Fig 2.1 masks for connectivity number

Chapter 3

Data Acquisition

Our goal was to obtain meaningful handwritten samples so that it can capture the between writer and within writer variation. For this reason we have collected handwritten samples from multiple writers and multiple samples from each writer. The handwriting samples were sufficient in number to exibit normal writing habits and consistency in writing habits. For comparison purpose the handwriting samples were similar in texts. Handwriting style depends upon several factors, e.g., age, gender, ethnicity, handedness, writing instrument (pen and paper), time of writing etc. So the document content were such that it capture as many features as possible. Some of these features were considered in our work.

Source Document:

Two source documents of Bangla script were taken for our work (Fig.1 and Fig.2). One of the two source documents was copied by each writer. One of the documents contains 158 words and the other contains 120 words. All the characters were present in that documents. The documents were composed of most frequently occuring words in Bangla language. Fig 3 shows such words along with their frequencies. The source document also contained puctuation, distinctive letter and a general document structure that allowed extracting macro document attributes such as word and line spacing, line skew, etc. Each writer copied one of the two source documents four times using plain, unlined sheets and ball pen. We have collected four such documents from 30 writers, i.e., total 120 such documents.

Each handwritten document was scanned and converted into a digitized image using a desktop black and white scanner. The resolution of the scanner was 300 dpi, and the resulting images were stored as binary bitmap images as shown in Fig-4.

Chapter 4

Preprocessing

Preprocessing comprises of the following subtasks.

1. Noise removal
2. Estimation of line count
3. Extraction of word's bounding box.

Noise Removal: Image is opened with 3×3 mask having all the values 1.

Estimation of line count: Let IMORI be the original image. Standard 2 pass component labeling algorithm is used for labeling connected components of black pixels of each image.Then for each component compute the distribution of length of horizontal and vertical runs of zeros terminated at both end by ones. We take the median of these distributions as the estimated intra-word loop-width and loop-height, respectively for that image[3]. Then suitable rectangular Structuring Element(SE) is applied in closing of the image.The long side of SE is along the default axis. For Our experiment length of SE is twice the estimated loop-width and height is half of the estimated loop-width (Fig 4.1). Now the image is scanned . The direction of scanning is perpendicular to that of the default axis. During scanning the number of background to objects transitions is counted for each column. Naturally, any duplicate transition between pixels bearing the same component label and background is ignored. The maximum transition count over all columns is taken to be the number of lines of handwritten scripts in the document. Note that, however all columns need not have same number of lines. We take this image IMCOMP.

Extraxtion of word's bounding box

The image IMCOMP is taken and the connected components of this image are found. Now as a result of closing (fig-4.2) there are the chances of merging two or more words. So to find this, each component from IMCOMP is scanned vertically in the original image (not in IMCOMP), if found they are divided into separate component. Thus we find all the words of the document. Now, for line merging each component is scanned horizontally. We have already got the line count, and hence average line height. Any component having height greater than the average line height, is then divided into two parts by taking horizontal projection, and choosing minimum count in-between the component height. So we have all the words extracted and hence their bounding boxes.

Chapter 5

Feature Extraction

Features are properties or attributes of objects that can represent an object. Thus the success of any object recognition system relies on extraction of suitable features of the object(s). There are two type of features: conventional features and computational features. Conventional features are those used by forensic examiners to establish the authorship of questioned documents. These features are obtained from the handwriting by visual and microscopic examination. Many of those features like writing quality are defined ambiguously and thus are subjective. Computational features are those that can be strictly defined in terms of computational algorithms used to extract them. Such features are unambiguous and hence remove subjectivity from feature measurement. To determine whether handwriting can be used to identify a person all computational features are suitable. Computational features are features that have known software/ hardware techniques for their extraction. The two types of features have some correspondence.

Computational features can be divided into macro- and micro-features. Macro-features can be extracted at the document level (entire handwritten manuscript) or at the paragraph, line, word, and character levels. In our work we have used computational features.

So far we have got the bounding box for each word. Now we will divide each word to m horizontal and n vertical boxes called cells. In our experiment value chosen for m and n were 4 and 3 respectedly (fig 5.1). Now we do thinning of the original image and store it in IMTHIN matrix (say), and also we compute Medial Axis Transform (MAT) and store it in IMMAT matrix (say).

In this phase two types of features were considered.

1. **Word level features:** Pen pressure (thin version), Pen pressure (MAT version), Aspect ratio, Depth.

2. **Cell level features:** Isolated points, Corner, Crossing, Branch, End point, Pixel density.

Word level features

Let (i_1, j_1) and (i_2, j_2) be two diagonal points of a bounding box such that $i_1 < i_2$ and $j_1 < j_2$ then

Pen Pressure: Two types of pen pressure were computed for a word (bounding box).

1. *Thinned version*: Number of black pixels in original image for a word, divided by number of black pixels in thinned image for the same word.

$$\text{Pen Pressure (thin)} = (\sum \text{IMORI}(i,j))/(\sum \text{IMTHIN}(i,j))$$
$$\text{where } i_1 \leq i \leq i_2 \text{ and } j_1 \leq j \leq j_2$$

2. *MAT version* : Sum of all the values in IMMAT for a bounding box, divided by number of black pixels in original image for the same bounding box.

$$\text{Pen Pressure (MAT)} = (\sum \text{IMMAT}(i,j))/(\sum \text{IMORI}(i,j))$$
$$\text{where } i_1 \leq i \leq i_2 \text{ and } j_1 \leq j \leq j_2$$

Pixel density: Pixel density is defined as total number of pixels having value one in IMTHIN for a bounding box divided by total number of pixels in bounding box.

$$\text{pixel density} = (\sum \text{IMTHIN}(i,j))/((i_2 - i_1)(j_2 - j_1))$$

Aspect Ratio: Width of bounding box divided by height of bounding box.

$$\text{Aspect ratio}=|(j_1\text{-}j_2)/(i_1\text{-}i_2)|$$

Depth: Highest value in IMMAT for a bounding box.

$$\text{Depth} = \max(\text{IMMAT}(i,j))$$
$$\text{where } i_1 \leq i \leq i_2 \text{ and } j_1 \leq j \leq j_2$$

Cell level features:

End point: A pixel which having exactly one black pixel in it's neighbourhood is regarded as end point. A point (i_c, j_c) is an end point if

$$\sum(\text{IMTHIN}(i_c+i, j_c+j)) = 2$$
$$\text{where } -1 \leq i \leq 1 \text{ and } -1 \leq j \leq 1$$

Corner points: Facet model (see sec. 2) was used for finding the corner points in IMTHIN.

Isolated points, crossing, branch were computed through finding the connectivity number (see sec. 2) in the IMTHIN.

For each Cell six features were calculated and four for each word (bounding box), i.e. Total of 6mn + 4 features for each word.

Chapter 6

Recognition

Training Set: Set of document images for which writers are known. It contains all the words of all the writers (except which are to be tested). Writer information is attached with every word in training set.

Test Set: Set of images for which writer is to be recognized using training set.

Euclidean distance:

The **Euclidean distance** between two points $P = (p_1, p_2, ..., p_n)$ and $Q = (q_1, q_2, ..., q_n)$, in Euclidean n-space,is defined as:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

K-Means algorithm: In pattern recognition, the k-means is a method for clustering where objects are organized into clusters. The objects in a cluster are similar in some way and are dissimilar to the objects belonging to other clusters.In our work, each feature is assigned a dimension to form a multidimensional feature space. Features of all the objects (words) with a priori known classes alongwith corresponding document numbers (reqired for classification) are extracted and assembled into feature vectors and those are stored along with corresponding class labels and document numbers. The K-means algorithm decreases intra-cluster error and increases inter-cluster error. The main idea behind this algorithm is to define k centroids (means) one for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest centroid (In our experiment the nearest centroid is calculated using a Euclidean distance). When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function.

The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{p} ||x_i^{(j)} - c_j||^2 ,$$

where $||x_i^{(j)} - c_j||^2$ is a chosen distance measure (Euclidean distance, in our case) between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the p data points from their respective cluster centres.

Each word is represented as a point in 6mn +4 dimensional feature space.

Algorithm:

The number of clusters to be formed (i.e,the value of K) is chosen.

For each writer, a data set is prepared by extracting all the word features of the documents of the corresponding writer.

For each writer,

step 1: initial means (centroids) of K clusters are selected randomly from the data set.

```
step 2: For each word in the data set
step 3:   compute euclidean distance with all the K means of the K clusters.
step 4:   assign the word to the cluster having minimum euclidean distance
          with that word.
        End for
step 5:   recalculate the positions of the K means.
step 6:   Repeat Steps 2 and 3 until the centroids no longer move.
```

End for

Output: K means for each writer, i.e, K cluster is formed for each writer.

End

Learning Vector Quantization(LVQ) algorithm : In Neural Network the LVQ is a method for classifying phenomena based upon observable features. The essence of the method is in deriving a set of reference vectors by iterative matching and adaptation to a training set. Following the convention, we represent such reference vectors as weight vectors each associated with a node in the output layer with no lateral connections. Geometrically, the reference vectors determine a Voronoi tessellation that partitions the feature space into different classes. Each node defines a decision region that includes all the points in the feature space that are closer to it than to any other node. The decision of the classifier is identical for all the points inside this region. In other words, the node may be considered as a representative having all the attributes of the points surrounding it. A node therefore represents a subset of the feature space with constant class membership. The output layer is used for classification when an input sample is matched to the weight vectors associated with the nodes of the output layer and the class of the input vector is matched with the class of the closest vector associated with a node in the output layer. In our experiment, q ($\leq$K) cluster- means each of 6mn+4 dimensional, formed at the end of k-means, have been chosen for each class from K-means method. The associated class numbers have been stored with these means. Those q means have been asssigned as the q initial weight vectors(one weight vector for each output node) for each class, i.e, total of 30q weight vectors for 30q nodes in the output layer. Each node in the output layer is 6mn+4 dimensional. The nodes in the input layer represent an input word. The nodes in the input layer are not laterally connected, but they are fully connected to the nodes of output layer. The learning equation used is ,

η(t)= $\eta_0(1-\frac{t\times\frac{I}{100}}{I})$ where I = training data size, t = no.of iteration, η(t) = learning rate at iteration t, η_0 = initial learning rate.

Algorithm :

Assign the output nodes by the means (output of K-means algorithm) associated with corresponding classes.

For each iteration (each iteration corresponds to the whole training data set) $t > 0$

For each data point x in the training set

compute the euclidean distance between the data point and each weight vector w_j of the output node.

Find the output node $v_j^{'}$ corresponding to the minimum distance.

if the class of x is the same as the class of v_j, $w_j^{'}$ is updated by

$w_j^{'}$(t+1)= $w_j^{'}$(t)+η(t)($x-w_j^{'}$(t)), otherwise,$w_j^{'}$ is updated by

$w_j^{'}$(t+1)= $w_j^{'}$(t)-η(t)($x-w_j^{'}$(t)),

End For

End For

Output: Converged weight vectors.

End

Training accuracy of recognition: Algorithm

Take an array for each document, where each index denodes an output node of LVQ and initialize all the values to zero. Assign misclassification to zero.

For each document of the training set

For each word of the document

compute euclidean distances with the weight vectors of the output nodes of LVQ.

find the minimum distance.

increment the value of the array of the document by 1 for that output node corresponding to minimum distance.

End For

Find the class of the output node corresponding to the maximum value of the array of the document.

If this class is not same as the class of the document then increment misclassification by 1.

End For

Output: accuracy = 1-(misclassification / total no. of documents in training set).

End

Testing accuracy of recognition is calculated by the same procedure as in training accuracy calculation. The only difference is that here, instead of training data set testing data set is considered.

Chapter 7

Experimental Results and Discussion

Training Data set was prepared by choosing three documents randomly from four documents of each writer. Testing data set was prepared by choosing the remaining document of each writer after choosing training documents. Naturally, we got four training data set and corresponding four testing data set. For every training document, image words were extracted and for each word 4 features were extracted, i.e., pen pressure (thin version),pen pressure (MAT version), depth, and aspect ratio. Each word is then divided into 4×3 cells and from each cell six features i.e., Isolated points, corner points, crossings, branch, end points and pixel density were extracted.

Threshold for corner detection (facet model) was chosen to be 0.9. System was trained using 30 writers 3 documents each, and one document from each writer was used as test set. For finding accuracy of recognition of writers we used LVQ. As the source document contains more than 100 different words, and we have 30 writers, we need, for each writer, more than 100 nodes in output layer of LVQ. Natuarally, space complexity and time complexity of the program will increase. That is why K-means algorithm is used. In K-means each output mean corresponding to a particular cluster of words represents some words which are closer to that mean than any other means because K-means algorithm decreases intra-cluster error and increases inter-cluster error. So one output node of LVQ can represent many words. K-means algorithm is applied for clustering of the whole data set (training and tesing both) of each writer. The value of K was initially chosen 50. At the end of K-means, considering minimum intra-cluster variation, 10 cluster means were selected from the 50 clusters. LVQ was trained by using these 10 cluster means of each writer. These selected means for each writer were initialised as initial weight vectors of the output nodes of LVQ. Initial learning rate was chosen as 0.01 and then the learning rate was gradually decreased with increase in iteration. Then after training phase,training accuracy and testing accuracy are calculated separately. The same procedure is applied when we are selecting 20 cluster means, 30 cluster means, 40 cluster means, 50 cluster means at the end of K-means, taking value of K = 50. We are also doing the same work when we are taking the value of K = 40, K = 60, K = 70 separately in K-means algorithm. The only difference is that when K = 40 is chosen we are selecting 10 clusters means, 20 cluster means, 30 cluster means, 40 cluster means at the end of K-means method and when K = 60 is chosen we are selecting 10 clusters means, 20 cluster means, 30 cluster means, 40 cluster means, 50 clusters means, 60 cluster means at the end of K-means method and when K = 70 is chosen we are selecting 10 clusters means, 20 cluster means, 30 cluster means, 40 cluster means, 50 cluster means, 60 cluster means, 70 cluster means at the end of K-means method. The K-means and consecutively LVQ were applied separately to all the four training data sets and the corresponding testing data sets.

Table 7.1, 7.2, 7.3, 7.4 show average training accuracy, average testing accuracy, best result for traning accuracy, best result for testing accuracy. Rows are representing the clusters selected. Columns are representing values of K. From the table it is clear that for a particular value of K, as the number of clusters selected increases, accuracy in both training and testing increases. Fig 7.1, 7.2 , 7.3, 7.4 show the graphs (accuracy vs. iteration for Training accuracy (series-1) , Testing accuracy (series-2)) when 40 clusters were selected out of 40 in K-mean, 50 clusters were selected out of 50 in K-mean, 60 clusters were selected out of 60 in K-mean, 70 clusters were selected out of 70 in K-mean respectively. From the graph it is clear that training accuracy increases with increase in iteration, but testing accuracy decreases after a particular iteration with increase in iteration.

clusters selected	out of 40	out of 50	out of 60	out of 70
10	71.9444	60.2778	56.1111	43.8889
20	91.1111	85.5555	85.5555	74.9999
30	93.6111	92.4999	92.2222	86.1111
40	94.1667	94.1667	93.8889	93.8889
50	$\times$	94.4444	93.6111	93.6111
60	$\times$	$\times$	95.5556	95.5556
70	$\times$	$\times$	$\times$	97.7778

Table 7.1: Average Training Accuracy (%)

clusters selected	out of 40	out of 50	out of 60	out of 70
10	63.3334	55.0000	43.3333	31.6667
20	74.9999	73.3333	54.9999	53.3333
30	81.6666	80.8333	71.6667	70.0000
40	87.4999	84.1666	76.6666	71.6667
50	$\times$	93.3333	92.4999	78.3334
60	$\times$	$\times$	92.4999	91.6667
70	$\times$	$\times$	$\times$	92.4999

Table 7.2: Average Testing Accuracy (%)

clusters selected	out of 40	out of 50	out of 60	out of 70
10	72.2222	53.3333	43.3333	42.2222
20	93.3333	83.3333	80.0000	74.4444
30	94.4444	92.2222	92.2222	85.5555
40	94.4444	95.5556	95.5556	93.3333
50	$\times$	96.6667	96.6667	95.5556
60	$\times$	$\times$	94.4444	94.4444
70	$\times$	$\times$	$\times$	97.7778

Table 7.3: Best Training Accuracy (%)

clusters selected	out of 40	out of 50	out of 60	out of 70
10	63.3333	53.3333	36.6666	30.0000
20	73.3333	63.3333	53.3333	53.3333
30	83.3333	80.0000	70.0000	63.3333
40	86.6667	83.3333	73.3333	70.0000
50	$\times$	96.6667	73.3333	70.0000
60	$\times$	$\times$	93.3333	90.0000
70	$\times$	$\times$	$\times$	90.0000

Table 7.4: Best Testing Accuracy (%)

Discussion

The error in the recognition system is due the following reasons:

1.**Improper word segmentation:** Fig 7.5 shows two same words for a single writer. In fig 7.5a whole word is included in one bounding box. In fig 7.5b same word is broken down into two different bounding boxes, causing the variations in feature vectors.

2.**Within writer variance:** When a writer starts writing he makes his handwriting to look good. But after sometime, he / she gets bore and wants to finish anyhow. Fig 7.6a is taken from first copy and fig 7.6b is taken from fourth copy of the same writer. The first is good looking and the second is not good looking. So the feature vectors vary for the same words of the same writer.

Fig 7.7 shows both documents where these words were taken from.

3.**Variation in Output in K-means:** The output produced in K-means depend on the initial values of the means and it frequently happens that suboptimal partitions are found.

4. **Overlearning in LVQ:** In case of overlearning the final locations of the weight vectors of the output nodes are very close to the training samples. So there are less chances that an unknown sample will match with the same class of weightvector of the output node.

Chapter 8

Conclusion and Scope for Future Work

We have just proved that each individual has consistent handwriting that is distinct from the handwriting of another individual. We suggested a methodology that validates that everyone writes differently. We have extracted four different types of word features of a word and six different types of cell features of the word. More features can be extracted.

References

[1] B.Chanda and D.Dutta Majumdar, *Digital Image Processing and Analysis*, Prentice-Hall of India, New Delhi, 2000.

[2] S.N. Srihari, Sung-Hyuk Cha, Hina Arora, and S. Lee "Individuality of Handwriting" J Forensic Sci, July 2002, Vol. 47.No.4.

[3] A. K. Das, A. Gupta and B. Chanda "A Fast Algoritm For Text Line And Word Extraction From Handwritten Documents" Image processing and Communications, vol.3, no.1-2, 1997, pp.49-61.

[4] K. Steinke Recognition of writer by handwriting images. Pattern Recognition, 14(1-6): 357-364, 1981.

[5] E.N. Zois and V. Anastassapoulous. Morphological waveform coding for writer identification. Pattern Recognition, 33(3): 101-105, 2000.

[6] http://www.cs.technion.ac.il/Labs/IsI/Projects_done/VisionClasses/Vision/

[7] D.A. Valkaniotis, J.Sirigos, N.Antoniades and N.fakotakis : Off-line Writer Recognition Using Neural Network,Wire Communications Laboratory, University of Patras,26500 Patras,Greece

[8] Gonzalez and Woods, *Digital Image Processing*, Pearson Education, Delhi, 2002.

[9] *Pattern Classification and Scene Analysis* - Duda and Hart, Willey Interscience Publication, USA, 1973

[10] *Neural Network* - Simon Haykin, Pearson Education, Delhi, 1999

[11] P.Tripathy, B. Chanda and B.B. Chaudhuri, *Writer recognition by analyzing word level features of handwritten documents*, International Conf. on Advances in Pattern Recognition(ICAPR'07), Kolkata, India, pp-73-78,2007

বাবু আর এই রুটে চা ফেরি করে না। হয় অন্য কোন রুটে করছে বা এই কাজ ছেড়ে দিয়েছে। কিন্তু তেমন কিছু হয়েছে বলে তো মনে হয় না। তবে হয়ে থাকলে আমি একটা খবর পেতাম। তার সঙ্গে আমার যে সম্পর্ক ছিল তা থেকে এটুকু আশা করা যায়। একবার অবশ্য একটি ছোট ব্যাপার নিয়ে আমাদের মধ্যে খুব কথা কাটাকাটি হয়েছিল। তবে সে কি আর মনে আছে? এ সব আলাপের প্রথম দিকে হয়েছে।

যে দিন ও প্রথম এখানে আসে আমি এবং হরিবাবু সেই সময় এই পথে রোজ যাতায়াত করি। তিনি আবার তাঁর মনের মত চা না হলে পছন্দ করেন না। তাই আজ এর কাছ থেকে, কাল তার কাছে আমরা চা খাই। এর মধ্যে দেখা হল ওর সঙ্গে। সে থাকে লাইনের ওই পারে। যদি তেমন একটা বেচাকেনা না হয়, তখন চলে আসে এই দিকে। কারণ অন্য কাজ করার খুব বেশী উপায় তার জানা নেই এবং যখন এমন দিন পর পর আসতে থাকে, তখন টাকা রোজগারের জন্য এ রকম কিছু একটা করতে হয়।

Fig-1 source document 1.

বাবু থাকে লাইনের ওই পারে। ট্রেনে আমি এবং হরিবাবু রোজ অফিসে যাই। ও চা ফিরি করে। হরিবাবু তাঁর মনের মত চা না হলে খেতে চান না। বাবুর চা-ই আমরা খাই। ক'দিন ধরে ছেলেটাকে দেখি না। ও কি আর এই রুটে চা ফিরি করে না? কিন্তু তেমন কিছু হয়ে থাকলে আমি নিশ্চয় একটা খবর পেতাম। ওর সাথে আমার যে সম্পর্ক তা থেকে এইটুকু আশা করা যায়। তবে একবার একটি ব্যাপার নিয়ে আমাদের মধ্যে বেশ কথা কাটাকাটি হয়েছিল। সে সব অনেক দিন আগের কথা।

একবার ও বলেছিল, ব্যাবসা ভালো চলছে না। যদি এমন চলতে থাকে তবে চা বিক্রি ছেড়ে দেবে এবং অন্য কোন কাজ দেখবে। লাইনের ওই পারে নতুন গুমটি হয়েছে, হয়তো তার মধ্যে একটা আমাদের বাবুর। আমি ওর উন্নতি আশা করি।

Fig-2 source document 2.

না	32877
করে	28201
এই	28127
ও	27681
হয়	22765
এবং	20265
যে	18997
থেকে	15985
তার	14501
আর	14222
করা	12650
কিন্তু	12462
বা	12388
যায়	11758
হয়ে	11098
সঙ্গে	11008
এক	10651
কোন	10575
জন্য	9942
একটি	9751
মধ্যে	9421
সেই	9223
করতে	9192
হবে	9073
ছিল	8673
তিনি	8593
কি	8256
আছে	8200
সে	8038
একটা	7756
কথা	7746
বলে	7674
আমার	7478
নিয়ে	7376
নয়	7284
তা	7270
আমি	7226
দিয়ে	7045

এ	6808
পারে	6758
থাকে	6732
তাঁর	6704
মনে	6465
হয়েছে	6385
সব	6150
হল	5873
তো	5709
নেই	5698
কিছু	5673
এর	5644
সময়	5510
আমাদের	5367
করেন	5133
তখন	5120
দেখা	4991
আমরা	4828
অনেক	4723
যদি	4600
তাই	4556
কাছে	4412
তাদের	4297
পর	4167
তবে	4133
আবার	4076
করার	4031
বলা	3910
দিকে	3897
প্রথম	3774
কারণ	3752
গেল	3650
কাজ	3591
এমন	3557
যখন	3545
এখন	3526
হতে	3507
উপর	3478

বড়	3462
বিভিন্ন	3449
বিশেষ	3430
ফলে	3340
ঐ	3306
যেমন	3276
যেন	3255
দেওয়া	3200
মত	3152
আগে	3144
খুব	3084
পড়ে	3080
হলে	3072
হইতে	3034
যা	3028
এসে	2985
নাম	2985
পাওয়া	2967
শেষ	2938
কী	2914
দিতে	2877
পরে	2843
করিয়া	2829
হচ্ছে	2819
তাকে	2818
করেছেন	2815
চলে	2808
মতো	2780
কেন	2733
তারা	2733
বললেন	2730
ঠিক	2727
প্রায়	2687
দিন	2666
শুধু	2655
ছিলেন	2643
ক্ষেত্রে	2641
গিয়ে	2637

বছর	2613
অন্য	2611
শুরু	2599
গেছে	2597
টাকা	2574
নতুন	2565
মানুষের	2554
একটু	2509
কেউ	2481
আজ	2467
হয়েছিল	2465
নিজের	2452
যেতে	2445
ধরে	2395
অর্থাৎ	2385
বেশ	2382
মানুষ	2377
কোনো	2364
দু	2364
আরও	2349
বেশি	2276
তারপর	2252
দেখে	2235
যাবে	2205
দ্বারা	2187
প্রতি	2178
কে	2173
এখানে	2163
করিতে	2141
সম্ভব	2122
একজন	2100
পর্যন্ত	2087
করলেন	2081
হাতে	2069
সŽর্কে	2066
দেশের	2062
ভাল	2038
বলতে	2028

ব্যবহার	2013
দেখতে	2005
সৃষ্টি	1983
দেয়	1979
তুমি	1973
আপনার	1951
দুটি	1941
বলেন	1936
প্রয়োজন	1935
তাহলে	1931
ব্যবস্থা	1931
দুই	1928

ওই	1914
ছাড়া	1903
বসে	1874
তোমার	1863
সমস্ত	1862
সেখানে	1856
জল	1855
টি	1853
হলো	1852
ইত্যাদি	1849
তাঁদের	1827
ওপর	1813

বলল	1813
করেছে	1810
চেষ্টা	1798
সকল	1789
নানা	1765
হাত	1762
ওর	1752
হইয়া	1746
ছবি	1740
কম	1735
ভাবে	1728
হইবে	1725

Fig. 3. Words and their frequencies.

বাবু আর এরূপে চাকরি করে না। হয় অন্য কোন রূপে করছে বা
এই কাজ ছেড়ে দিয়েছে, কিন্তু তেমন কিছু হয়েছে বলে তো মনে হয়
না। তবে হয়ে থাকলে আমি একটা খবর পেতাম, তার সঙ্গে আমার
যা সম্পর্ক ছিল তাতে এটুকু আশা করা যায়, একবার অবশ্য একটা
ছোট ব্যাপার নিয়ে আমাদের মধ্যে খুব কথা কাটাকাটি হয়েছিল,
তবে সে কি আর মনে আছে ? এসব আলাপের প্রমাদিকে
হয়েছে।

যেদিন ও প্রথম এখানে আসে আমি এবং হরিবাবু সেই
সময় এই পথে রোজ যাতায়াত করি, তিনি আবার তাঁর পছন্দমত
চা না হলে পছন্দ করতেন না, ওই আড্ডের কাছ থেকে, কাল ওর
কাছ থেকে আমরা চা খাই, এর মধ্যে দেখা হলে ওর সঙ্গে, ও
লাইনের ওপারে থাকে, যদি তেমন একটা বেচা কেনা না হয়
তখন চলে আসে এই দিকে, কারন অন্য কোন কাজ করার উপায়
তাঁর বেশী জানা নেই, এবং যখন এমন দিন পর পর চলতে
থাকে, তখন টাকা রোজগারের জন্য এরকম কিছু একটা করে
হয়।

Fig-4 sample document

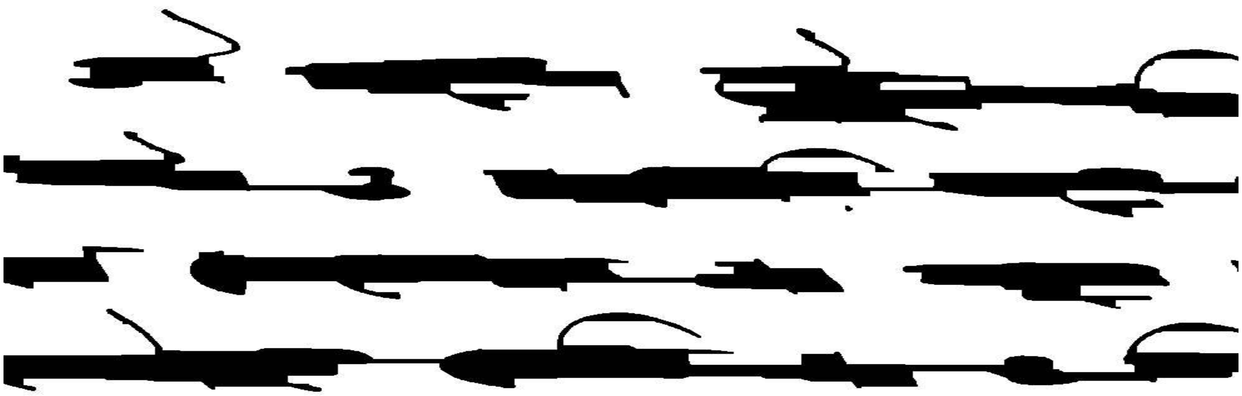


Fig 4.1 Result after closing

Fig 4.2 word extraction

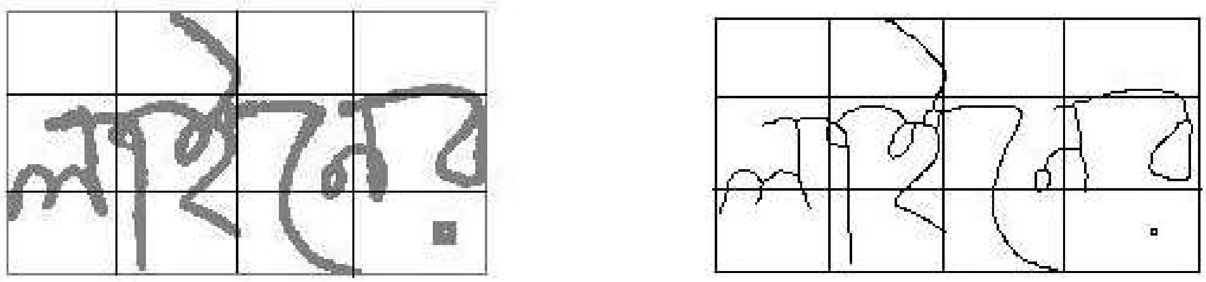
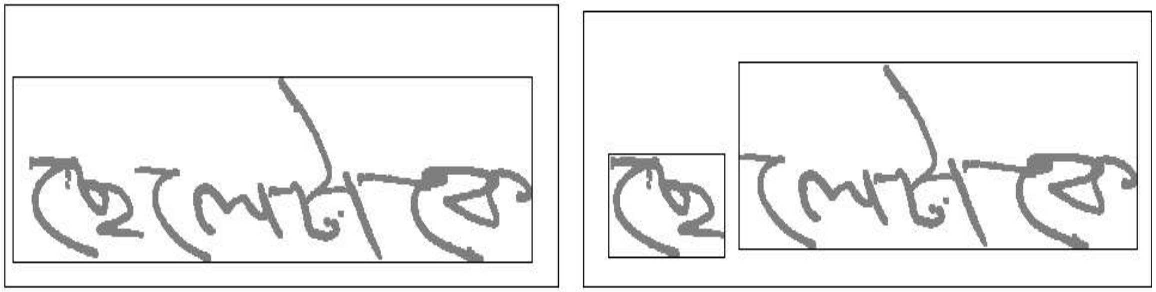


Fig. 5.1 Division of a word into Cells.



(a) (b)

fig 7.5: Two same words

(a) (b)

fig 7.6 : same word, same writer but in two different documents

বাবু থাকে লাইনের ওই পারে। ট্রেনে আমি এবং হরিবাবু রোজ অফিসে যাই। ও চা ফিরি করে। হরিবাবু তাঁর মনের মত চা না হলে খেতে চান না। বাবুর চা-ই আমরা খাই। ক'দিন ধরে ছেলেটাকে দেখি না। ও কি আর এই রুটে চা ফিরি করে না? কিন্তু তেমন কিছু হয়ে থাকলে আমি নিশ্চয় একটা খবর পেতাম। ওর সাথে আমার যে সম্পর্ক তা থেকে এইটুকু আশা করা যায়। তবে একবার একটি ব্যাপার নিয়ে আমাদের মধ্যে বেশ কথা কাটাকাটি হয়েছিল। সে সব অনেক দিন আগের কথা।

একবার ও বলেছিল, ব্যাবসা ভালো চলছে না। যদি এমন চলতে থাকে তবে চা বিক্রি ছেড়ে দেবে এবং অন্য কোন কাজ দেখবে। লাইনের ওই পারে নতুন ওমটি হয়েছে, হয়তো তার মধ্যে একটা আমাদের বাবুর। আমি ওর উন্নতি আশা করি।

বাবু থাকে লাইনের ওই পারে। ট্রেনে আমি এবং হরিবাবু রোজ অফিসে যাই। ও চা ফিরি করে। হরিবাবু তাঁর মনের মত চা না হলে খেতে চান না। বাবুর ~~চাই~~ চা-ই আমরা খাই। ক'দিন ধরে ছেলেটাকে দেখি না। ও কি আর এই রুটে চা ফিরি করে না? কিন্তু তেমন কিছু হয়ে থাকলে আমি নিশ্চয় একটা খবর পেতাম। ওর সাথে আমার যে সম্পর্ক তা থেকে এইটুকু আশা করা যায়। তবে একবার একটি ব্যাপার নিয়ে আমাদের মধ্যে বেশ কথা কাটাকাটি হয়েছিল। সে সব অনেক দিন আগের কথা।

একবার ও বলেছিল, ব্যাবসা ভালো চলছে না। যদি এমন চলতে থাকে তবে চা বিক্রি ছেড়ে দেবে এবং অন্য কোন কাজ দেখবে। লাইনের ওই পারে নতুন ওমটি হয়েছে, হয়তো তার মধ্যে একটা আমাদের বাবুর। আমি ওর উন্নতি আশা করি।

Fig 7.7 : Two different handwritings of a writer