WRITER IDENTIFICATION IN HANDWRITTEN DOCUMENT USING OCR TECHNIQUE

Smt.Vidya D
Assistant Professor of Computer Science
Government First Grade College, Farahatabad

Abstract:

An image is a digital representation of real-world scene composed of discrete elements called pixels. Pixels are parameterized by position,intensity,time these parameters define still images,video, volume data and moving volumes. Digital image processing operations are electronic data processing operations on a 2-D array of numbers. The array is a numeric representation of an image.

Optical Character Recognition (OCR) deals with the automatic machine recognition of characters present in a document image. OCR converts the paper manuscripts to digital documents where entire contents can be read and processed by machines.

The utilities proved by OCR systems are manifold postal automation, banking automation, digital library creation and the design of reading aids for the blind in the form of text-to-speech converters

Image processing operations can be divided into image compression, image enhancement and restoration and measurement extraction. Image compression is a technique of minimizing the size of graphic file without degrading the quality of an image. Image enhancement can be defined as conversion of image quality to a better and more understandable level for feature extraction. Once the image is in good condition, the measurement extraction operations can be performed to obtain the useful information from the image.

OCR includes for the processing of images the stages like “Preprocessing” the raw data is collected and put into number of stages. “Segmentation” extraction of characters from words. “Feature Extraction”this method includes identification of vertical and horizontal and skew strokes in the characters. “Classification” by applying k-nearest neighbour methods or Euclidian distance method we classify the features finally by applying the “Post Processing technique” applying different algorithms it will find out the writer of a particular document.

Introduction:

As the the world moves closer to the concept of the paperless office, more and more communication and storage of documents is performed digitally. Documents and files that were once stored physically on paper are now being converted into electronic form in order to facilitate quicker additions, searches and modifications. Because of this there is a great demand for software, which automatically extracts analyzes and stores information from physical documents for later retrieval. All of these tasks fall under the general heading of document analysis, which has been a fast growing area of research in recent years.

A very important area in the field of document analysis is that of optical character recognition(OCR), which is broadly defined as the process of recognizing either printed or handwritten text from document images and converting it into electronic form . To date, many algorithms have been presented in the literature to perform the task for a specific language , and for multilingual environment , such OCR will not work . Therefore to make a successful multilingual OCR , script identification is very essential before running an individual OCR system. In this direction , most of the published work on automatic script identification in india deals with printed documents and very few articles are found for handwritten script identification problem for two major indian languages English and devanagari as an initial attempt.

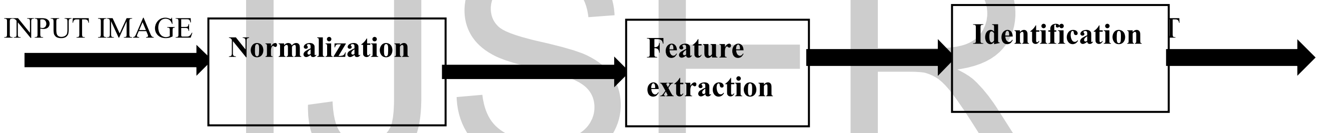
Optical Character Recognition (OCR) deals with the automatic machine recognition of characters present in a document image. OCR converts the paper manuscripts to digital documents where entire contents can be read and processed by machines.

The utilities proved by OCR systems are manifold, Postal banking automation, digital library creation and the design of reading aids for the blind in the form of text-to-speech converters.

The Process contains the stages :Data collection and preprocessing, feature extraction

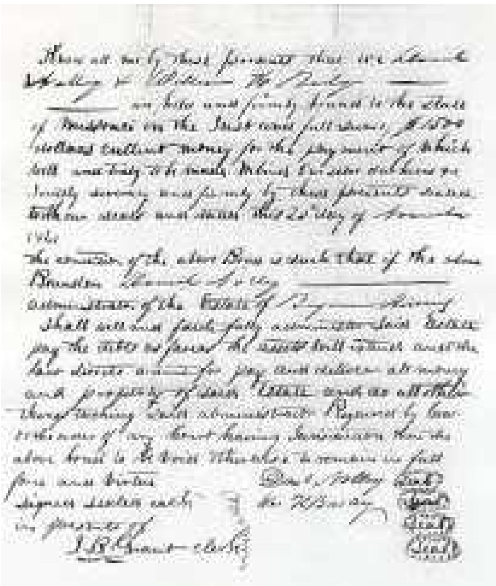
Algorithms, Experiemental results

Data collection and Preprocessing:



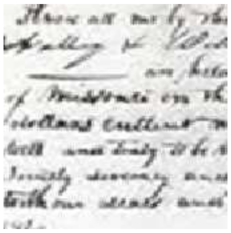
Data collection and Preprocessing:

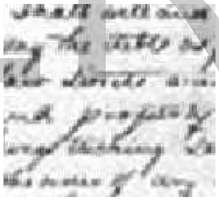
A sample of 250 hanwritten documents are collected from different writers. The purpose of data collection is not disclosed.The collected documents are scanned using HP scanner at 300 DPI, which usually yields a low noise and good quality document image. The digitized images are in gray tone and they have used Otsu's global thresholding approach to convert them into two-tone images. Threshold is a normalized intensity value that lies in the range [0, 1]. Otsu's method chooses the threshold to minimize the interclass variance of the threshold black and white pixels. The two tone images are the converted into 0-1 labels where the label 1 represents the object and o represent the background. The single or double quotation marks, hyphens and periods etc are removed using morphological operations.



Segmentation :

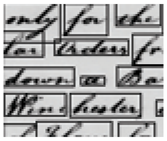
Partitioning of an image into several constituent components is called segmentation., we have broken our normalized image into 128 X 128 images, since the image given is bilevel in nature , we have to convert it into binary image and process is known as binarization.

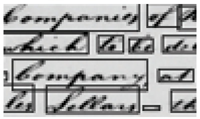


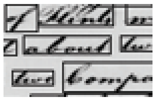


Morphological operations:

The two tone images are then converted into 0-1 labels where the label 1 represents the object and 0 represents the background. The small objects like single or double quotation marks , hyphens and periods etc are removed using morphological opening.







Feature Extraction:

Each sample or pattern that we attempt to classify is a block of text (128 x128 pixels). The important property of English script is the existence of the vertical strokes in its characters and has less number of horizontal strokes. The right and left diagonal strokes plays an important role.

To extract the characters or components containing strokes in vertical, horizontal, right and left diagonal directions they have performed the erosion operation on the input binary image with the line-structuring element. The length of the structuring element is thresholded to 70% of average height of all the connected components of an image. The resulting image is used for morphological opening in four directions to obtain the strokes present in the connected components of the image as shown in fig.

Average Stroke length: It is defined as the average of average length of individual strokes found in the connected component of an image. Let N as number of connected components

The values of seven features extracted here :

1. Average Vertical Stroke Length(AVSL):

$$\text{AVSL(pattern)}=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{N}\sum_{k=1}^{N}(length(strokei)\right)$$

2. Average Horizontal Stroke Length(AHSL):

$$\text{AHSL(pattern)}=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{N}\sum_{k=1}^{N}(length(strokei)\right)$$

3. Average Right Diagonal Stroke Length(ARDSL):

$$\text{ARDSL(pattern)}=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{N}\sum_{k=1}^{N}(length(strokei)\right)$$

4. .Average Left Diagonal Stroke Length(ALDSL):

$$\text{ALDSL(pattern)}=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{N}\sum_{k=1}^{N}(length(strokei)\right)$$

Where length(stroke) is the number of pixels in a stroke .

Stroke Density(SD)- this is the number of strokes per unit length(x-axis) of the connected component. It is defined for the strokes present in vertical, horizontal, right and left diagonal directions of a connected component.

$$\text{VSD=} \quad \sum_{k=1}^{N}\left(\frac{ni}{width(componenti)}\right)$$

$$\text{HSD=} \quad \sum_{k=1}^{N}\left(\frac{ni}{width(componenti)}\right)$$

$$\text{RDSD=} \quad \sum_{k=1}^{N}\left(\frac{ni}{width(componenti)}\right)$$

$$\text{LDSD=} \quad \sum_{k=1}^{N}\left(\frac{ni}{width(componenti)}\right)$$

5. **Average stroke density(ASD):** is defined as

$$\text{ASD(pattern)=}\frac{VSD+HSD+RDSD+LDSD}{N}$$

Where VSD, HSD, RDSD AND LDSD are vectors of size 1 x N

Proposed Algorithm:

Input: scanned image

Output: Binarized image

Method: WED classifier used to distinguish the writer

The steps involved in algorithm are:

1. Input initial image
2. Pre-process the input image i,e, binarization using Otsu's method and remove speckles using morphological opening
3. Extract strokes of each connected component of an image in vertical, horizontal, right and left diagonal directions using morphological erosion and opening operations.
4. Computer the average stroke lengths of images in all directions.
5. Evaluate the average stroke density of an image.
6. Calculate average aspect ratio of an image.
7. use nearest neighbor to classify the new image

Experimental Results:

The results are obtained by using NN-classifier shown in below table:

Writers	Set A	Set B	Set C	Set D	Set E
Set A	---	80%	70%	70%	90%
Set B	80%	---	70%	70%	80%
Set C	80%	70%	---	90%	70%
Set D	70%	70%	90%	---	90%
Set E	90%	80%	70%	90%	-----

Conclusion:

We have proposed a robust method for handwritten script identification of pair of writers. The aim is to facilitate the multiple handwritten OCR and script based retrieval of offline handwritten documents. It can also be noted that the proposed algorithm performs very well even for text line wise script identification

A number of experiments have been conducted. The experiments use 10 differentwriter lasses. Features were extracted from handwriting images using the multi-channelGabor filtering and the grey scale co-occurrence matrix (GSCM) technique. Identificationwas performed using two different classifiers (the weighted Euclidean distance (WED) andthe K-nearest neighbour (K-NN) classifiers). The results achieved were very promising,and an identification accuracy as high as 96.0% was obtained. The two classifiers haveshown good performance, but at some stages the KNN classifier's performance is relativelypoor when compared to the WED classifier.We are currently investigating ways of reducing the impact of such factors on theperformance of the proposed global approach. We will also consider local approacheswhich seek writer specific features to improve the recognition accuracy. In the future,both global and local approaches will be integrated as one system for a betteridentification accuracy

References:

[1] Plamond and G. Lorette, “Automatic Signature Verification and Writer Identification-The State of Art,Pattern Recognition, 1989, Vol.22, No.2, pp107-131.

[2] W. Kuckuck, “Writer Recognition by Spectra Analysis “, Proc. Int. Conf. In Security ThroughScience Engineering, 1980, West Berlin, Germany, pp.1-3.

[3] T. N. Tan, “Written Language Recognition Based On Texture Analysis”,Proc. IEEE ICIP’96,Lausanne, Switzerland, Sept. 1996, Vol.2, pp.185-188.

[4] T. N. Tan, “Texture Edge Detection by Modelling Visual Cortical Channels”, Pattern Recognition, 1995,Vol.28, No.9, pp.1283-1298.

[5] A. K. Jain, and S. Bhattacharjee, “Text Segmentation Using Gabor Filters for Automatic Document Processing”, Machine and Vision Applications, 1992, pp.169-184.

[6] G. S. Peake and T. N. Tan, “Script and Language Identification from Document I