TEXT CONTENT DEPENDENT WRITER IDENTIFICATION

Samuel A. Daramola[1], Morakinyo A. Adefunminiyi[2]

[1]*Associate Professor, Electrical and Information Engineering, Covenant University, Ota Ogun State, Nigeria.*
[2]*PG Student, Electrical and Information Engineering, Covenant University, Ota Ogun State, Nigeria.*

Abstract

Text content based personal Identification system is vital in resolving problem of identifying unknown document's writer using a set of handwritten samples from alleged known writers. Text written on paper document is usually captured as image by scanner or camera for computer processing. The most challenging problem encounter in text image processing is extraction of robust feature vector from a set of inconstant handwritten text images obtained from the same writer at different time. In this work new feature extraction method is engaged to produce active text features for developing an effective personal identification system. The feature formed feature vector which is fed as input data into classification algorithm based on Support Vector Machine (SVM). Experiment was conducted to identify writers of query handwritten texts. Result show satisfactory performance of the proposed system, it was able to identify writers of query handwritten texts.

Keywords: *Handwritten Text, Feature Vector, Identification and Support Vector Machine.*

----------***----------

1. INTRODUCTION

Biometric traits are classified as behavioral or physiological. Behavioral traits like signature and handwritten text written on paper are widely used for authentication of persons. Handwritten text identification involves authentication of a person using biometric features from handwritten text. Handwritten text recognition can be done manually or automatically. Automatic personal identification via handwritten texts is better than manual method because ordinarily people find it difficult to differentiate unforeseen handwritten texts. Effectiveness of automatic personal identification system is largely depends on it is ability to suppress high intra variation within handwritten texts of the same person, main causes of this variation include writing speed, pen handling and mental ability. Personal identification using handwritten text can be done in two ways based on the input sensors. They are called off-line and on-line [1][2]. Off-line identification method involves using pen to write text on paper while on-line identification is done by writing text expressions on digitizer. In case of on-line method in additional to text shape, dynamic features like speed and pressure are captured for processing [3][4].

Handwritten text recognition is relatively different from Optical Character Recognition (OCR), the objective of handwritten text recognition is to identify the author of a given text while that of OCR is to recognize characters of text written on paper and convert it to digital form without writer identification. Therefore handwritten text recognition uses behavioral characteristics to identify document's writer. Each person's handwriting is unique, characteristics of handwriting includes size, shape, slope of characters and regular or irregular spacing between words. A writer identification process makes use of one-to-many search method to decide authentic author of a written text among a set of known writers. One major application of writer identification system is in forensic science in which unknown writer of a contested document is revealed through analysis of written document such as police report, will, diaries and written statement by suspects, examination paper, election score sheet, money transfer form etc.

Many researchers have proposed several writer identification methods. These methods can be classified as text dependent or text independent writer identification. In text dependent approach identification of writers is based on specific targeted handwritten text [5][6]. While text independent approach identification of document's writer is based on any written text [7] [8][9][10].

In [11] text dependent writer identification system based on Support Vector Machine (SVM) was developed. Two edge based features were extracted from handwriting text. Also features were extracted at word and character level. Also in [12], writer identification system was proposed; nine geometric features were extracted from handwriting text using subset of "IAM" dataset. Each writer's model was generated based on individual Hidden Markov Model (HMM). In case of writer text dependent identification system based on neural network that was developed in [13], features extracted from handwritten text lines include width, slant, and three heights of the writing zones. Whereas in [14], codebook of connected component contours is generated to model character allographs using a bag of features model. On other hand, edge hinge feature from curvature of characters using relative angle of two line segments on a character's contour was produced in [15], other features like connected component contours, run lengths and slant features were included. And in [16], extraction of eight features from Malayalam text at character level was carried out. The features are classified as loop features, distance features and directional features. They are fused together to form feature vector for writers identification.

In this work, handwriting text images are not segmented to character or word level before feature extraction. Unlike most previous methods text segmentation preceded feature extraction. In this work errors as result of text segmentation are avoided by extracting robust feature vector from text image using moving window across the text. It involves scanning across a one pixel width text image with one pixel overlap away from subsequent window. In each block, geometric feature is extracted. Feature vectors generated from this window operation were used to establish writers' model. The output data from feature extraction stage is fed to classification algorithm.

2. DIGITAL IMAGE PROCESSING

The personal identification system proposed in this work firstly involves capturing of handwritten text image from users. The next stage is the image preprocessing, it involves preparation of image for next step which is called feature extraction method. The training stage is carried out using SVM to model handwritten text of each writer. The final stage is the classification algorithm where query handwritten texts are verified to discover the authentic writers.

2.1 Image Acquisition

Imaging of handwritten text acquired from twenty subjects was carried out using flatbed scanner. Each user was asked to write an expression tagged "Money transfer confirmed by me" on a white sheet of paper. Users of varies ages group are allowed to write on paper without any constraints. Thereafter the handwritten texts are scanned into computer system as gray-scale bitmap image at 300 dots per inch. Each dot or pixel is a shade of grey that has integer value range from 0 to 255, it means that each pixel can be represented by eight bits. An example of gray scale handwritten text images gotten from three different writers is shown in fig-1.

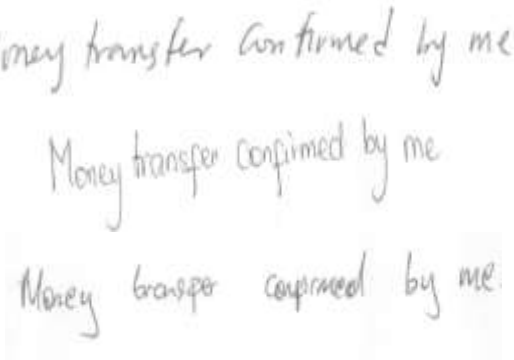

Fig-1, Gray-level handwritten text image.

2.2 Image Preprocessing

Three major preprocessing operations are performed on the captured gray-scale handwritten text images and thereafter they are passed to feature extraction stage. Preprocessing methods involve at this stage include filtering, thresholding and thinning.

Filtering operation is achieved using smoothing filters or sharpening filters. Smoothing filters are low pass filters that blurring image detail by removes noise and higher frequency components, examples of smoothing filter include Average, Gaussian and Median filter while sharpening filters perverse high frequency component such as fine details, line, points and edges. Sharpening filters can be achieved using gradient and laplacian masks. Given an image *f(x,y)* and a filter mask *h* of size mxn. Output smoothed image g(x,y) is obtained as result of convolution operation between image *f(x,y)* and filter mask *h* as the center moves to every pixel in the image as given in (1)

$$g(x,y) = \sum_{k=-a}^{a}\sum_{j=-b}^{b} h(j,k)\, f(x-j, y-k) \qquad (1)$$

Where a=m/2 and b=n/2

Level of smoothing depends on the size of the mask therefore different size of masks can be used to obtain desired degree of smoothing. In this work Gaussian filter is used for smoothing text images, for Gaussian filter, the kernel equation is as given in (2).

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} exp^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (2)$$

Standard deviation of the Gaussian σ control the size of filter mask and level of smoothing, for σ = 1.4, the Gaussian mask for size 7x7 is shown in fig-2, the normalization summation of all elements is equal to 1. Example of smoothed handwritten text images using Gaussian filter is shown in fig-3.

1	1	2	2	2	1	1
1	2	2	4	2	2	1
2	2	4	8	4	2	2
2	4	8	16	8	4	2
2	2	4	8	4	2	2
1	2	2	4	2	2	1
1	1	2	2	2	1	1

Fig-2, 7x7 Gaussian mask.

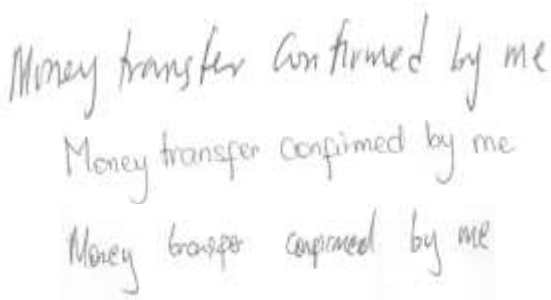

Fig-3, Smoothed handwritten text images

Thresholding is a method of creating binary image from gray-level image. The technique involves reduction of 8bits information (0 - 256) gray value in each image pixel to 1bit, that is value of 0 or 1. There are three major methods of thresholding a gray-level image. They are called global, Otsu and adaptive thresholding. Global thresholding is the appropriate method to use if histogram of the gray-level image is bi-modal, that is a partition exists between the object pixels and background pixels on the histogram plot. Then single threshold value can be selected within the interval of the valleys for good segmentation.

Otsu's method also relies on image bi-modal histogram information to perform image thresholding using class clustering method. Threshold value that separate two classes, that is image foreground and background is determined based on inter and intra class variation. Optimum thresholding value is achieved when intra-class variance and inter-class variance attain minimal and maximal value respectively.

Adaptive thresholding is one of the methods of converting gray-scale image to binary image. If the image histogram is not bimodal the best option is to use adaptive thresholding. The image is divided into small blocks; thresholding operation is performed on each block using single or optimum threshold value. In this work global threshold method is adopted. All the pixels that have gray value above selected threshold value are set to 1 and remaining pixels are set to zero. Examples of binary handwritten text image obtained using single threshold value are shown in fig-4.

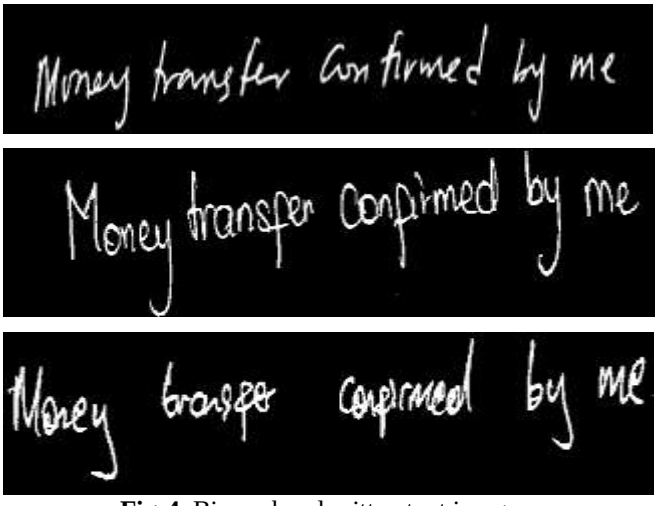

Fig-4, Binary handwritten text images.

Thinning is a morphological operation that acts on binary image; it prunes pixel width of each point in the image to one pixel width. The process reduces the number of foreground pixels and reveals the skeleton of the image that shows the true representation of the text image written style. Also the process helps to speeds up subsequent operations. Fig-5 shows example of handwritten text images that have been thinned.

3. FEATURE EXTRACTION STAGE

The feature extraction method is developed to generate feature vector that can effectively describe each user text written style and shape. The thinned binary handwritten text images are the input to the feature extraction algorithm. The image is divided into blocks by sliding a window of specific size over it from left to right with predefined overlapping.

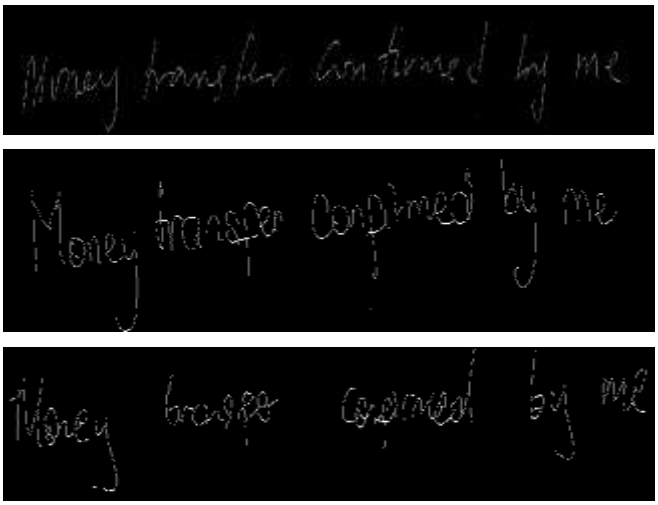

Fig-5: Thinned binary handwritten text image.

Feature is extracted at area occupy by the window, therefore feature element is obtained from every area occupy by the window as it move from left to right. The feature extraction steps are stated as follows:

- Step 1: Resize and obtain bounded thinned binary handwritten text image. Example is as shown in figure6
- Step2: Determine the size of window
- Step3: Set number of overlapping pixel.
- Step4: Slide window over the thinned binary image to divide the image to 36 smaller blocks. Figure 7 shows for example ten out of the thirty six blocks.
- Step5: Extract feature called mean centroid from each block.
- Step6: Form feature vector of 36 elements, Table 1 show example of feature vectors obtained from three different text image samples.

Fig- 6: Bounded thinned binary text image

Figure-7: Parts of block text image.

Table 1. Feature vectors for three different images.

No	Data 1	Data 2	Data 3
1	4.53488224	12.38031525	9.33817476
2	3.45168660	6.86022538	3.91014770
3	2.92667433	3.25610200	2.83660513
4	8.88962993	6.21064804	2.50587736
5	8.06341917	8.18699370	2.08696116
6	3.72129567	5.34424170	2.37960997
7	4.26013989	7.37065097	1.84135074
8	7.21623005	6.52470017	4.08325380
9	7.89360382	11.00691679	3.19968781
10	4.33285064	9.21598548	4.57277024
11	4.26295108	6.86913956	2.98169322
12	8.12868180	9.09138834	4.95866631
13	6.69972154	10.92457946	6.76451155
14	9.78906810	8.74400284	4.99804262
15	17.92384759	7.52221820	10.40938746
16	5.07721463	4.01696092	6.85775013
17	4.38308015	3.70579609	1.82574185
18	3.26426010	6.50076624	4.74704732
19	7.03765709	11.81375619	9.84151925
20	7.04935645	11.29388042	5.14524852
21	3.18791348	6.44414825	5.85924645
22	5.05271813	4.99686548	4.27633496
23	4.79949178	4.29052236	3.82861823
24	4.79155607	4.48467434	4.83982411
25	5.59597457	3.58149313	8.71839516
26	3.66391304	5.80491320	3.79316748
27	4.66648379	11.8331122	2.75638869
28	5.34633522	12.0999499	2.66570876
29	4.66961411	4.54999429	1.49646881
30	2.02494479	3.14113533	3.21111889
31	2.43479013	6.32723827	3.48928287
32	6.47949391	14.18773091	5.15260922
33	11.80428729	14.18773091	5.15260922
34	4.81542451	7.56849121	3.63902746
35	5.83899639	6.51802772	3.24395016
36	4.08391309	6.38929940	2.56322433

4. SUPPORT VECTOR MACHINE FOR WRITER IDENTIFICATION.

Support Vector Machine (SVM) is a learning algorithm that has the capacity to classify feature vectors into different regions separated by optimal hyper plane using structural risk minimization method. Decision surface that is the optimal hyper plane represents the largest link between negative and positive examples. SVM deals with non-linearly separable data by project original input feature space into a higher dimension feature space using nonlinear kernel.

Application of SVM in resolving personal identification problem using handwritten text images is achieved using training and testing algorithm. The system is trained using positive and negative text images to establish an optimal hyper plane that increases margin between the support vectors. The testing algorithm is used to determine classes of query handwritten text images based on the distance of the query feature vector from the hyper-plane.

5. EXPERIMENTAL RESULT

Experiments are conducted to establish the capability of the proposed personal identification method to detect originator of contested handwritten text. Training of the proposed system is carried out using text images store in the system. Five handwritten text images are collected from each of the twenty people. Three images per person are passed to training algorithm based on Support Vector Machine (SVM) to establish each user model. Query text images are sent to the system for confirmation, 95% of the query images are detected correctly for the authentic handwritten text writers.

6. CONCLUSION

Personal identification system based on handwritten text images and Support Vector Machine (SVM) has been developed for detecting originator of queried handwritten text. This was achieved by extracting discriminating feature from a set of suspected text images written by known writers. Text image is divided into smaller overlapping blocks to obtain robust local feature using a moving window. The experimental result obtained from the proposed text content dependent identification system using non-segmented character image is very encouraging.

REFERENCES

[1]. R. Plamondon and G.lorette. 1989. Automatic Signature verification and writer Identification- The state of the art. Pattern Recognition vol.22, no.2, pp.107-131.

[2].Daramola Samuel and Ibiyemi Samuel. 2010. Novel Feature Extraction Technique for Off-line Signature Verification System. International Journal of Engineering Science and Technology vol. 2(7), pp. 3137-3143.

[3].A. Chaabouni, H. Boubaker, M. Kherallah, A.M. Alimi and H.E. Abed. 2011. Combining of Off-line and On-line Feature Extraction Approaches for Writer Identification. In Proc. of the 11th International Conference on Document Analysis and Recognition, pp.1299-1303.

[4]. S.A Daramola and T.S Ibiyemi. 2010. Efficient On-line Signature Verification System. International Journal of Engineering & Technology (IJET-IJENS), vol. 10, No: 04, pp.42-46.

[5] Constantine Anton Cosmin Ştirbu Romeo, Vasile Badea. 2010. Identify Handwriting Individually Using Feed Forward Neural Networks. International Journal of Intelligent Computing Research (IJICR), vol. 1, Issue 4, pp. 183-188.

[6]. Somaya Al-Maadeed. 2012. Text-Dependent Writer Identification for Arabic Handwriting Journal of Electrical and Computer Engineering; pp1-8.

[7]. B. Helli, M.E. Moghaddam, 2008. A text-independent Persian writer identification system using LCS based classifier, in: IEEE International Symposium on Signal Processing and Information Technology, (ISSPIT). pp. 203–206.

[8] B. Helli, M.E. Moghaddam. 2010. A text-independent Persian writer identification based on Feature Relation Graph (FRG), Pattern Recognition. Vol.43, pp.2199–2209.
[9].I. Siddiqi, N. Vincent. 2010. Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. In Pattern Recognition. vol.43, pp.3853 –3865.
[10] Golnaz Ghiasi, Reza Safabakhsh. 2013. Offline text-independent writer identification using codebook and efficient code extraction methods, Image Vision Computer, vol.31, no.5, pp.379-391.
[11].Saranya K and Vijaya M S . 2013. Text Dependent Writer Identification using Support Vector Machine. International Journal of Computer Applications. vol.65– No.2, pp.6-11.
[12].A. Schlapbach and H. Bunke. 2007. A writer identification and verification system using HMM based recognizers, Pattern analysis and applications. pp. 33-43.
[13].U. Marti, R. Messerli, H. Bunke. 2001. Writer identification using text line based features. Proc. of 6th International Conference on Document Analysis and Recognition pp. 765–768.
[14].M Bulacu, L Schomaker. 2007. Text-independent writer identification and verification using textural alloygraph feature. Pattern Analysis and Machine Intelligence, IEEE Transactions vol.29, Issue:4, pp.701 - 717.
[15]. L Schomaker, M Bulacu. 2004. Automatic writer identification using connected-component contours and edge-based features of uppercase western script . Pattern Analysis and Machine Intelligence, IEEE Transactions vol.26, Issue: 6, pp.787 – 798.
[16].Sreeraj M, Idicula S. 2011. Identifying Decisive Features for Distinctive Analysis of Writings in Malayalam. International Magazine on Advances in Computer Science and Telecommunications. vol. 2, no. 1. pp.13-20.