Feature Extraction from English Handwritten Text for Individuality Detection

Salankara Mukherjee and Ishita De Ghosh

October 18, 2019

Feature Extraction from English Handwritten Text for Individuality Detection

Salankara Mukherjee
Department of Computer Science
West Bengal State University, Barasat
Kolkata 700126, India
Email: getsalu@gmail.com

Ishita De Ghosh, Senior Member, IEEE, ACM
Department of Computer Science
Barrackpore Rastraguru Surendranath College
Kolkata 700120, India
Email: ishitade.ghosh@gmail.com

***Abstract*—Detection of individuality in handwriting is an important research topic. In this work various features are extracted from handwritten English text and then analyzed. Simple algorithms are used for the purpose. Experimental results are given for sample images from IAM data set. It is observed that for pre-printed text, skew-angle and baseline do not show much variation, whereas zone, slant and pen pressure show enough variation and hence individuality.**

***Index Terms*—Handwriting analysis, Feature extraction, Individuality detection.**

I. Introduction

In spite of extensive use of electronic documents and printed text, handwritten document image processing remains an active research area. It is essential in character recognition, writer identification, signature verification, individuality detection, personality analysis and many other applications. Automatic feature extraction is an integral part of handwritten document image processing and it is a non-trivial task because of inter-personal and intra-personal variations in handwriting. Inter-personal variations arises from the fact that each person writes in his own way, and different persons write differently. Even for the same individual, handwriting varies depending on age, physical or mental state, writing conditions and writing tools. This results in intra-personal variations. Among many applications that involve feature extraction from handwritten documents, some focus on the common characteristics of writings by many individuals, eg. character recognition [1], some focus on the common characteristics of writings by the same individual, eg. writer identification [2], while some others focus on common as well as distinct characteristics of writings by many individuals, eg. individuality detection [3], [4] and personality analysis [5].

Present work focuses on feature extraction for detecting individuality of handwriting. It explores some common but important features of handwritten English text viz. skew, baseline, zones, slant and pen-pressure for the purpose. The importance of each feature, the algorithm for extraction, and the experimental results on sample images from IAM databse [6] are provided. The paper is organized as follows. Section II gives a brief review on related work; section III gives discussion on the features mentioned above, and presents the algorithms to extract them; this section also includes illustrations of input and output images of the results obtained. Section IV presents experimental results obtained by applying the algorithms on samples from IAM Database in tabular form; finally concluding remarks are given in section V.

II. Previous Work

Individuality detection and personality analysis from handwritten documents are related research fields and they can be enriched from each other. A brief discussion on previous work in these two fields are given now. Importance of study on individuality of handwriting and development of computer-assisted procedures for comparing handwriting was discussed by Srihari and others in [3]. A subset of attributes used by expert document examiners, was used to quantitatively establish individuality by using machine learning approaches. Feature extraction and analysis is done in a local approach in [2]. Writing is divided into sub-images and morphologically similar sub images are grouped together for analysis. Link between handwriting and neurological aspects of human brain was reported by Plamondon in [7], where it was shown that formation of characters in the brain of a writer depends on his habits and each neurological brain pattern forms a distinctive neuromuscular movement which is similar for individuals with the same type of personality. Accordingly handwriting is a representation of human brain. A method of writer identification and behavior evaluation is discussed in [8]. For this the authors considered six features: size of letters, slant of letters and words, baseline and pen pressure. An SVM is trained by the features to guess behavioral traits. In [9] the authors propose a novel method for extracting a set of baseline-independent features, which are based on the combination of global and local information. They use a handwriting database in Arabic scripts for their work. An off-line, writer-independent handwriting analysis system to predict the personality traits of a writer automatically from features extracted from a scanned image of the writer's handwriting sample given as input ws poposed in [10]. The features include pen pressure, slant of letters and size of writing. In [11], the objective is to analyze the handwriting characteristics like baseline, slant, pen-Pressure, size, margin and zone to determine the emotion levels of a person. Machine learning approach like KNN with incremental learning is used in [12] to improve the efficiency of the tool, which analyze the handwriting features like margins, baseline and T-bars. In [4], the authors study handwriting of children to measure the development of individuality. An automated system is used to compare the writings at the word-level as well as paragraph-level. The results provide strong support that handwriting becomes more individualistic as the child grows. In [5], the authors study feature extraction from handwritten documents for personality analysis. They discuss from an algorithmic view-point on the following features, size, spacing, skew, slant and pressure. In [13], the authors propose the first non-invasive three-layer architecture based on neural networks to study the relationship between personality traits and offline handwriting. They also present the first database collected from 128 subjects containing personality classification and handwriting samples. They studied the features baseline, word slant, writing pressure, connecting strokes, space between lines, lowercase letter 't' and lowercase letter 'f'.

III. Proposed Work

As discussed in previous sections, individuality of a handwriting is expressed by various attributes. We have worked with a number of attributes or features, but due to limited space present only a few of them is presented in this work, viz. skew, baseline, zones, slant and pen-pressure. For illustration purpose, handwriting samples collected from diverse background are used. A brief description of our work is given now. In general, the handwritten documents are scanned as RGB images and they require pre-processing before feature extraction. First step of pre-processing is converting the images from RGB to grayscale. Next step is median filtering for noise removal. Then grayscale to binary conversion is done by thresholding. After that, handwritten text portions are extracted from the binarized images and finally text lines are segmented. After the pre-processing steps, feature extraction procedures are applied on individual text lines. The features and their extraction procedure which are described here are; Skew and Baseline Calculation, Zone identification, Slant angle Calculation and Correction, Pressure calculation. Schema diagram in Fig. 1 will portray the overall design of our work.

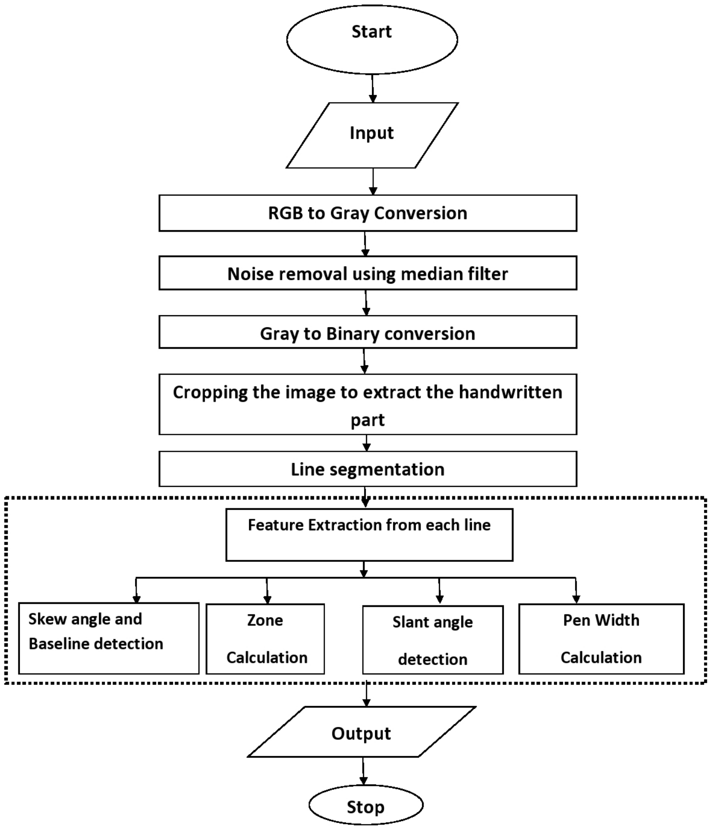
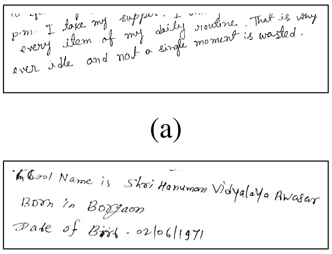


Fig. 1: Schema Diagram

A. *Skew and Baseline of writing*

Although skew and baseline of a writing are separate features, they are related features. Hence both of them are presented in this section. Skew is the angle in degrees between line of writing and the horizontal line, measured anti-clockwise. Skew-angle is zero for horizontal writing, positive for upward writing and negative for downward writing. It is seen in general that if the writer copies from a predefined text, the handwritten text lines tend to be horizontal. However if the writer writes random text freely on his chosen subject, text lines tend to be upward or downward and sometimes even erratic. Therefore random text expresses more individuality than predefined text in respect of skew-angle.

Baseline is a significant feature for handwriting analysis. It is the pre-printed or imaginary line on which the letters in English script reside. When a person starts writing a line of text in a blank paper, he estimates an imaginary straight line parallel to the edge of the paper and try to put his words on this line. But words or letters in words go up and down from this imaginary straight line forming upward baseline or downward baseline. Even convex, concave or wave-like baselines are found in handwritings. So every person constructs his own baseline based on his subconscious mind. Examples of varying skew-angles and baselines are shown in Fig. 2. Each single text line of a document

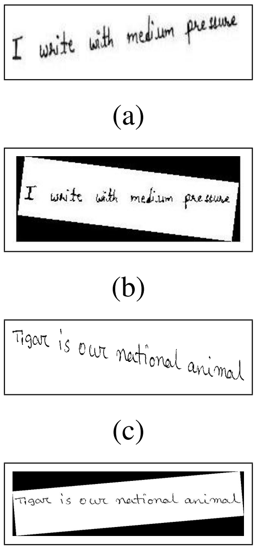


(a)

(b)

Fig. 2: Examples of texts with varying skew-angles and baselines. (a) Ascending skew-angles and baselines, (b) Descending skew-angles and baselines

is considered for skew angle detection and baseline calculation. Horizontal projection profile method is used for skew angle detection [14], [15]. At first a limiting $angle$ is decided and the segmented text-line is rotated from $-angle$ to $+angle$. For each rotation, number of black pixels in each row is counted and the row with the highest horizontal frequency is identified. Corresponding pixel-count and the angle-value is stored. At the end a total number of $2 * angle + 1$ pairs are obtained. Among them the highest horizontal frequency value is identified. Corresponding angle-value gives the required skew angle. For skew correction, the image is rotated with the negative of the skew-angle.



(a)

(b)

(c)

(d)

Fig. 3: Detection and correction of Skew and calculation of Baseline. (a) Skew angle=+7, Ascending Baseline (b) After skew correction of (a), (c) Skew angle=-5, Descending Baseline (d) After skew correction of (c)

As described in [13], baseline is classified as one of the following: Leveled or Normally Straight Baseline, Ascending Baseline and Descending Baseline depending on the range in which the skew-angle falls. If the angle is within the range -5 degree to $+5$ degree the text-line under consideration has a **leveled baseline**. If it is within the range $+5$ degree to $+30$ degree the text-line has a **ascending baseline**. If it is within the range -30 degree to -5 degree the text-line has a **descending baseline**. In Fig. 6 two different skewed images having respectively ascending and descending baselines are shown. Skew corrected images are also shown in the figure.
Algorithm for skew angle detection and baseline classification is given now.

1) Start
2) Convert the RGB image to gray scale image.
3) Apply median filtering process to the gray scale image for noise removal.
4) Convert the gray scale image into binary image.
5) for $i = -angle$ to $+angle$
   i. Rotate image in $angle_i$
   ii. Calculate the boundary points of the rotated image and get the bounding box of the image.
   iii. Calculate horizontal frequency of each image. Store the maximum frequency value of each

image in the array max_hor_freq[ ].
Store also corresponding angle in the array $angle$[ ].
$[End\ \ of\ \ outer\ \ for\ \ loop]$

6) Choose the maximum value from the array max_hor_freq[ ].With this,identify the corresponding angle max_angl from the array $angle$[ ].
7) Rotate the source image in that particular maximum angle max_angl and the skew correction is done.
8) if(max_angl>-5 and max_angl$<+5$)
The image has a Leveled Baseline
elseif(max_angl$>=-30$ and max_angl<-5)
The image has a Ascending Baseline
elseif(max_angl$>+5$ and max_angl$<+30$)
The image has a Descending Baseline
$[End\ \ of\ \ if]$

9) Stop

B. Zone division of writing

The feature zone division of handwriting are explored now. English texts with upper and lowercase letters have three zones, viz. upper, middle and lower. Although in print-documents zone division is very clear, it is not so for the handwritten ones. Distribution of a handwriting in the zones shows how the writer makes specific use of their mind, emotions and physical elements in their environment [5]. In general, children start to learn their English handwriting in four-line pages, so that the letters have proper zone divisions. But as they grow older, their individuality in handwriting becomes more prominent and zone division may become less clear. Often the size of handwriting is decided by a benchmark of 3mm in middle zone in a full height of 9mm as normal writing exceptions of which are classified as large or small writings [8]. Above concept is not useful for zone calculation since size of writings vary.

The basic principle of our zone calculation is that since all letters have their presence in the middle zone, horizontal projection profile of a text-line have high frequency values for middle zone. After pre-processing, horizontal projection profile of a binary text-line image is calculated. The maximum horizontal frequency is identified from this profile. Now the concept is that horizontal frequency of each row in the middle zone should be greater than one-third of the maximum value. This concept is used to identify the middle zone. Once middle zone is identified, upper and lower zones are identified using the bounding box of the text-line and bounding-box of the the middle-zone. An illustration of zone division using our algorithm is depicted in Fig. 4. The algorithm for detection of the zone division and calculation of zone-width is as follows.

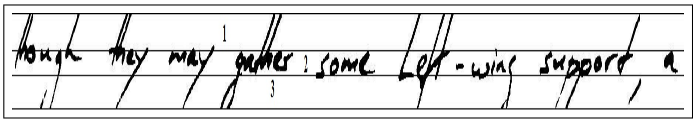


Fig. 4: Detection of zones and calculation of zone-width in number of pixels. Upper-zone width=30, Middle-zone width=20, Lower-zone width=27

1) Start
2) Read the pre-processed binary image as input.
3) Calculate the number of black pixels in each row.
4) Find out the maximum horizontal frequency.
5) Calculate $1/3$ of maximum horizontal frequency. This is called T.
6) Traverse from bottom-row to top-row and check where the black-pixel count of the row exceeds T. This particular row number is r_1.
7) Traverse from top-row to bottom-row and check where the black-pixel count of the row exceeds T. This particular row number is r_2.
8) Calculate the width of the upper zone by $r_2 - top - row$.
9) Calculate the width of the middle zone by $r_1 - r_2$.
10) Calculate the width of the lower zone by $bottom - row - r_1$.
11) Stop

C. Slant of writing

Slant of a handwriting is the inclination of the letters towards the right or the left. In general, handwritings with disconnected letters have little or no slant. English cursive writing has a natural right slant. It is noted that the individuals who are trained to write in cursive, tend to have a right slant even when they write disconnected letters. Handwritings with left-slant are uncommon. Slant angle is measured in degrees clockwise from the vertical line to the axes of letters. If the axes of letters make zero degree angle with the vertical line then the writing has zero or vertical slant. If the axes of letters make positive angle with the vertical line then the writing has positive or right slant. If the axes of letters make negative angle with the vertical line then the writing has negative or left slant. Variation in slants is illustrated

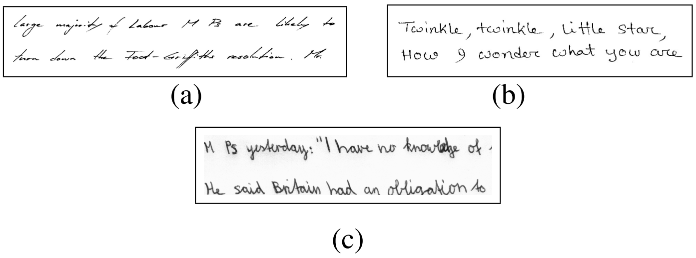


Fig. 5: Example of texts with various slants. (a) Right slant, (b) Near-zero slant, (c) Left Slant

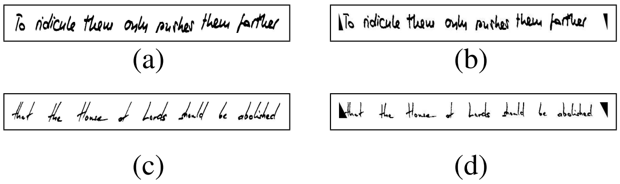


Fig. 6: Detection and correction of slant. (a) Handwriting sample with moderate right-slant (b) Slant corrected image, Detected slant-angle=26.53 (c) Handwriting Sample with severe right-slant (d) Slant corrected image, Detected slant-angle=44.97

in Fig. 5. In [16] a slant angle calculation procedure is described. Depending upon this concept we try to build our algorithm. In this algorithm we choose a shear factor range, in between which we make affine transformation of the image. Each time we calculate the vertical frequency count of every transformed image. The shear factor for which the vertical frequency count is maximum is identified. This shear factor is now used for slant angle calculation. The algorithm for detection and correction of slant-angle is as follows.

1) Start
2) Convert the RGB image to gray scale image.
3) Apply median filtering process to the gray scale image for noise removal.
4) Convert the gray scale image into binary image .
5) for $i = -5$ to $+5$
   i. Make an affine transformation using the matrix $[100; i10; 001]$ where i is the shear factor.
   ii. Calculate the boundary points of the transformed image and get the bounding box of the image.
   iii. Calculate vertical frequency of each image. Store the maximum frequency value of each image in the array max_ver_freq[ ]. Store also corresponding shear factor in the array $shearf$[ ].

   $[End\ of\ outer\ for\ loop]$
6) Choose the maximum value from the array max_ver_freq[ ].With this,identify the corresponding shear factor max_shear from the array $shearf$[ ].
7) Make affine transformation on the source image in that particular maximum shear factor max_shear and the slant correction is done.
8) Slant angle is calculated by the formula $tan^{-1}(1/max\_shear)$
9) End

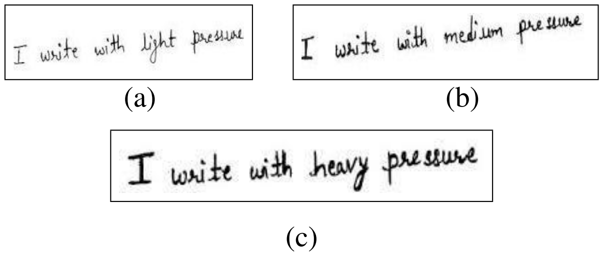


Fig. 7: Same text with different pressure: (a) Light Pressure, (b) Medium Pressure, (c) Heavy Pressure

D. Pressure of writing

Pressure of writing is not same for all individuals. It indicates the force exerted by the writer during writing. Type of the pen as well as the gripping style of it are also revealed by pressure. Pen pressure may be heavy, light or medium. Variation in pressure is illustrated in Fig. 7. Now we discuss about two methods used in our work for calculating the pressure of writing. For the binarized text image it is noted that the total number of black pixels is large when the pressure is high. In the first method total number of black pixels in the text image is counted at first. Then a thinning operation using the ZS algorithm is done. This procedure makes the letters in the writing one pixel thin. Finally the difference between the total number of black pixels before and after thinning is calculated. Large difference implies high pressure, whereas small difference means low pressure. An example of thinned image is provided in Fig. 8. Sample outputs of the method is depicted in Fig. 9. Here same text is written by the same writer but with different pressure. When applied on the first sample, the above algorithm generates pixel-counts of 4643 and 1941 before and after thinning respectively. So the difference in count is 2702. When applied on the second writing, the above algorithm generates pixel-counts of 6798 and 2986 before and after thinning respectively. So the difference in count is 3812.

So increment in pressure results in increment in black pixel count. The output is in accordance with the manual inspection which tells that the first one is written with low pressure whereas the second one is written with high pressure. The second method is based on the idea

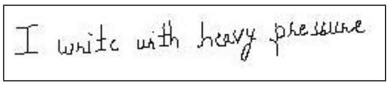
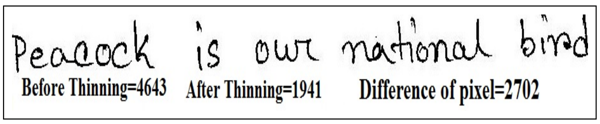


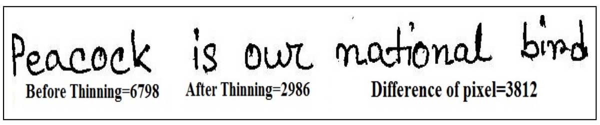
Fig. 8: After Thinning of Fig.7(c)

of pressure classification by grayscale thresholding [17]. At first the RGB image is converted into grayscale and then median filter is applied for noise removal. Pixel-intensities of the noise-free image are in the range [0,255]. A text written with high pressure has more number of pixels with low intensity values because black pixels have value zero. In our work, the average gray-value of the pixels with intensity-value less than 150 is calculated. Less average value means darker writings and hence high pressure. Percentage of the above average-value with respect to the highest intensity value 255 is calculated. When the average is less then the percentage is also less and the writing is dark. When the average is high, percentage is also high and the writing has less or medium pressure. An output of second method is given here for images in Fig. 7. In this figure, image (a) shows light pressure and the percentage value is 44, image (b) shows medium pressure and the percentage value is 37, whereas image (c) shows high pressure and the percentage is 28. The algorithms for pressure calculation are given now.

- First Approach :
  1) Start
  2) Read the pre-processed binary image as input.



(a)



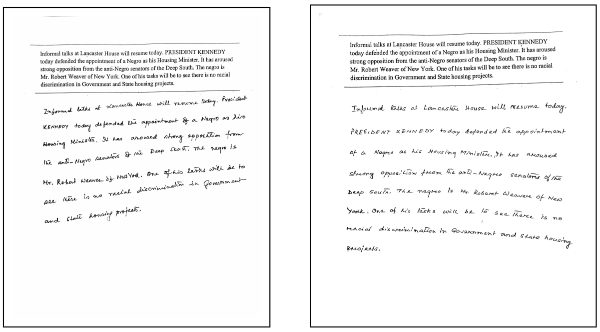
(b)

Fig. 9: (a) Low pressure writing, (b) High pressure writing

  3) $Before\_thin$ =Total number of black pixels calculated before thinning
  4) Use ZS algorithm for thinning
  5) $After\_thin$ =Total number of black pixels calculated after thinning
  6) $Diff = After\_thin - Before\_thin$
  7) End
- Second Approach :
  1) Start
  2) Convert the RGB image to grayscale.
  3) Apply median filtering to the grayscale image for noise removal.
  4) Let m=Number of rows , n=Number of columns
  5) for $i = 1$ to m
     for $j = 1$ to n
     if (Gray value of(i,j)-th pixel of the image array < 150)
     $cnt = cnt + 1$
     $s = s + Image(i, j)$
     $[End\ of\ if]$
     $[End\ of\ inner\ for\ loop]$
     $[End\ of\ outer\ for\ loop]$
  6) $avg = s/(m * n)$
  7) $percnt = (avg * 100)/255$
  8) End

IV. Experimental Result

In our work we use the renowned IAM Handwriting Database [6] which contains forms of handwritten English text. All of our program code accept scanned text as input from this database. Then as a pre processing step we crop the input image to extract the handwritten text part. After that lines are separated from the whole text. Algorithms described in this work are applied on individual line. MATLAB version R2014a(32bit) is used as image processing software. Intel corei5 processor, 4GB RAM and 64 bit operating system are used as hardware and software specification. Due to limited space, we present a few of our results in Table I. Four features like Skew angle detection and Baseline calculation, Zone detection, Slant angle calculation and pen pressure calculation are described in this table with respect to minimum value, maximum value and average value. Frequency of baseline types is also presented here. As we present two approaches of pressure calculation, our resultant table shows only the output of the first approach. In IAM database, handwritten text samples are almost skew corrected(parallel to x-axis), so all the



(a) (b)

Fig. 10: (a)No.of ascending baseline=5; No.of leveled baseline=2, (b)No.of descending baseline=3; No.of leveled baseline=4

sample image have zero degree skew angle and baseline type is leveled. In Fig. 10 we present two samples (which are not a part of IAM database)to show the variety of baseline type and skew angle.

V. Conclusion

This work concentrates on feature extraction from handwritten document images for individuality detection. The features analyzed are skew angle, baseline, zone, slant angle and writing pressure. The features are described from algorithmic view-point and images are treated as collection of black and white pixels. Experimental results are given on images from IAM handwriting data set which is a controlled data set in the sense that pre-printed text-samples are copied by the writers. In our experiments, it is observed that skew-angle and baseline do not show much variation, whereas zone, slant and pen pressure show enough variation and hence individuality. In future, more features will be extracted, analyzed and they will be applied on large no of controlled and random data set. Finally our aim is to design an automated system which can efficiently and quickly read handwritten images, extract and analyze different features and these results can be applied to support large no of image processing applications.

References

[1] R. Tolosana, R. Vera-Rodriguez, J. Ortega-Garcia, and J. Fierrez, "Preprocessing and feature selection for improved sensor interoperability in online biometric signature verification," in *IEEE Access*, May 2015.

[2] I. A. Siddiqi and N. Vincent, "Writer identification in handwritten documents," in *9th International Conference on Document Analysis and Recognition ICDAR*, (Brazil), 2007.

[3] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, "Individuality of handwriting: A validation study," in *6th International Conference on Document Analysis and Recognition*, September 2001.

[4] S. N. Srihari, L. Meng, and L. Hanson, "Development of individuality in children's handwriting," *Journal of Forensic Sciences*, vol. 61, July 2016.

[5] S. Mukherjee and I. De, "Feature extraction from handwritten documents for personality analysis," in *International Conference on Computer Electrical and Communication Engineering*, (Kolkata), December 2016.

[6] "Iam handwriting database." Website:http://www.fki.inf.unibe.ch/databases/ iam-handwriting-database. Last modified Feb 05, 2019.

[7] R. Plamondon, "Neuromuscular studies of handwriting generation and representation," in *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, (Kolkata, India), Novenber 2010.

[8] S. Prasad, V. K. Singh, and A. Sapre, "Handwriting analysis based on segmentation method for prediction of human personality using support vector machine," *International Journal of Computer Applications*, vol. 8, pp. 25–29, October 2010.

[9] N. Li, X. Xie, W. Liu, and K. M. Lam, "Combination of global and local baseline-independent features for offline arabic handwriting recognition," in *21st International Conference on Pattern Recognition (ICPR)*, (Tsukuba, Japan), November, 2012.

[10] A. R. M, D. Varghese, and M. K. G, "Habit: Handwritten analysis based individualistic traits prediction," *International Journal of Image Processing (IJIP)*, vol. 7, no. 2, 2013.

[11] S.V.Kedar, D.S.Bormane, A. Dhadwal, S. Alone, and R. Agarwal, "Automatic emotion recognition through handwriting analysis: A review," in *International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 811–816, 2015.

[12] P. Joshi, "Handwriting analysis for detection of personality traits using machine learning approach," *International Journal of Computer Applications (0975 – 8887)*, vol. 130, November 2015.

[13] M. Gavrilescu and N. Vizireanu, "Predicting the big five personality traits from handwriting," *EURASIP Journal on Image and Video Processing*, July 2018.

[14] P. Slavk and V. Govindaraju, "Equivalence of different methods for slant and skew corrections in word recognition applications," *IEEE Transactions On Pattern Analysis and Machine Intelligence*, vol. 23, March 2001.

[15] U. Pal and B. B. Chaudhuri, "Indian script character recognition: A survey," *IEEE Transactions On Pattern Analysis and Machine Intelligence*, vol. 23, March 2001.

[16] L. B. Mahanta and A. Deka, "Skew and slant angles of handwritten signature," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 1, November 2013.

[17] R. Coll, A. Fornes, and J. Llados, "Graphological analysis of handwritten text documents for human resources recruitment," in *12th International Conference on Document Analysis and Recognition*, p. 1081–1085, July 2009.

Sl no.	Img name	Skew angle in deg	Occurance of Baseline types			Zone width in no. of rows(Avg)			Slant angle in deg		Pen pressure in no. of pixels		
		Avg	Leveled	Ascend	Descend	Up	Mid	Low	Min	Max	Min	Max	Avg
1	a01-003.png	0.00	10	0	0	24.40	21.50	6.90	45.02	45.02	3617	6488	5284.30
2	a01-030.png	0.00	8	0	0	16.13	37.50	6.38	26.57	-26.57	5581	9457	8216.13
3	a01-053.png	0.00	11	0	0	53.00	52.00	16.45	26.57	45.02	7600	10766	9303.45
4	a01-087u.png	0.00	9	0	0	13.00	26.78	7.44	-26.58	0.00	5307	6314	5744.00
5	a01-132.png	0.00	11	0	0	14.45	29.45	8.00	-26.58	26.58	4817	6475	5667.27
6	a02-000.png	0.00	7	0	0	4.20	37.80	2.20	45.02	45.02	1662	4367	3389.38
7	a02-008.png	0.00	9	0	0	5.13	28.00	3.38	26.58	26.58	1749	5542	4521.78
8	a02-017.png	0.00	8	0	0	9.14	30.85	5.43	0.00	45.02	2320	4514	3747.75
9	a02-053.png	0.00	8	0	0	2.17	34.50	1.83	0.00	0.00	2819	3828	3380.50
10	a03-037.png	0.00	8	0	0	8.75	43.75	4.50	-56.34	0.00	4640	8278	6963.63
11	d06-008.png	0.00	10	0	0	21.30	36.10	8.40	26.58	45.02	5361	6973	6239.78
12	g01-008.png	0.00	9	0	0	21.00	29.33	5.89	0.00	0.00	4800	8232	7061.70

TABLE I: Experimental Results