

Feature Extraction from Handwritten Documents for Personality Analysis

Salankara Mukherjee
Khardah Priyanath Balika Vidyalaya
Kolkata 700117, India
Email: getsalu@gmail.com

Ishita De, Senior Member, IEEE
Barrackpore Rastraguru Surendranath College
Kolkata 700 120, India
Email: ishitade.ghosh@gmail.com

Abstract—Handwriting can be used to predict or analyze a person's behavioral or personality traits. Characteristics of a handwriting are studied for that. In this work various characteristics like size, spacing, slant, skew, pressure, etc are studied. Since signature reflects important characteristics of human being, we analyze them also. As our objective is to build an automated or computerized handwriting analysis system, we discuss the features from algorithmic view points. For experimental purpose, various handwriting of different writers are collected. Identical as well as different texts are used for that. Implementing these algorithms on the data sets, we obtain experimental results which are given here.

Index Terms—Handwriting analysis, Feature extraction, personality analysis.

I. INTRODUCTION

Human beings communicate among themselves in many ways. Handwritten text is one of the oldest ways of non-verbal communication. Till today it is in use, though the advancement of technology has provided humankind with many other tools for the latter. There are various reasons for its popularity. For example, it can be written by any literate person at any place, by using a simple tool like a pen on a paper and it can be conserved easily for a long time. Handwriting is often called brain-writing. Neuro muscular movements for handwriting are resulted from neurological brain pattern. It is observed that during writing nearly seventy to eighty percent cells of human brain work actively, whereas during talking only fifteen to twenty percent of it work actively. Throughout writing, the writer's brain gives directions to him for two things: what to write and how to write. The conscious part of the brain dictates him about what to write or the contents of writing; while both conscious and subconscious parts control the movements of hand to produce the writing.

Handwriting conveys important information about the physical, mental and emotional state of the writer during writing and also about his overall behavior or personality traits. Two persons cannot have identical hand writings though there may be striking similarities. Each writer has his own unique characteristics which may change depending on age, experience and the states mentioned above. Some external factors such as tools used for writing, environment in which the writing is produced also influence one's handwriting. The science of handwriting analysis to detect behavioral or personality traits is called Graphology. The word comes from the Greek word

“Graph”, which means writing and another Greek word logos, which means theory. Camillo Baldi, the Italian Professor and Physician analyzed handwriting in a systematic manner and wrote his first book on Graphology in 1622. In the next four hundred years, many people contributed to develop the subject.

In general, handwriting analysis by graphologists is done manually. Automatic handwriting analysis is a challenging task and an active research topic. This paper presents our work related to automatic analysis of handwritten documents in English script. It gives a brief literature survey, discusses on features of handwriting, provides algorithms for feature extraction, and finally it gives experimental results on collected data. The paper is organized as follows. Section II gives a brief review on related work; section III discusses on common features of handwritten English text and their meanings in the light of Graphology; section IV presents the algorithms used by us to extract the features, it also includes experimental results on collected data; and finally concluding remarks are placed in section V.

II. RELATED WORK

A number of work for automatic handwriting analysis has been published in recent years. A method of writer identification and behavior evaluation is discussed in [1]. For this the authors considered six main different types of features: (i) size of letters, (ii) slant of letters and words, (iii) baseline, (iv) pen pressure, (v) spacing between letters and (vi) spacing between words in a document to identify the personality of the writer. Segmentation is used to calculate the features from digital handwriting and is trained to SVM which outputs the behavior of the writer. In [2] feature extraction and analysis is done in a local approach. Writing is divided into sub-images and morphologically similar sub images are grouped together. Bayesian classifier is used in this paper. In the paper [3] the following features of handwriting are discussed upon in the light of graphology: baseline, slant, size, margin, pressure, speed, spacing, zones, print writing and cursive writing, connecting strokes and signature. Here algorithms and experimental results are not given. The main objective of [4] is to analyze the handwriting characteristics like Baseline, Slant, Pen-Pressure, Size, Margin and Zone to determine the emotion levels of a person. In [5] the authors have carried out research of the various state of the art technologies

available in analyzing an individual's behavior based on their handwriting and the effectiveness of predicting the character and personality of that individual. They tried to determine handedness, authorship and gender through analysis. In this paper [6] 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. A HMM-based recognition system is developed with 161 models that include a space model and a blank model. Handwriting database (for Arabic scripts) Links can be found from this paper. Names of online tools for handwriting analysis are mentioned in [7]. They analyze English and Korean handwriting using Fuzzy logic. Computer aided graphology concept is used in [8]. They describe a system which is able to analyze handwritten text without much human intervention. They also extract the features like margin, baseline, size, zone. Their report includes a personality type based on Myer Briggs dichotomies, a temperament based on Kerseys temperament sorter. Machine learning approach like KNN with incremental learning is used in [9] to improve the efficiency of the tool, which analyze the handwriting features like margins, baseline and T-bars. In [10] a method has been proposed to predict the personality of a person from the features extracted from his handwriting using Artificial Neural Networks. The personality traits revealed by the baseline, the pen pressure and the letter "t" as found in an individual's handwriting are explored.

III. FEATURES OF HANDWRITING

As mentioned in section I, each writer has unique handwriting characteristics which may change depending on age, experience and physical or mental conditions. The characteristics are related to important behavioral personality traits such as emotional steadiness, concentration, adaptability, motivation, honesty, intelligence, energy, fear, defense etc. The characteristics or features are expressed by the strokes of writing, structure or size of letters and even by the gaps between lines or letters. The topic of handwriting characteristics and their relation to various personality traits are discussed in detail in [11].

In this section, we discuss on various features of writing, viz. zone, size, spacing, skew, baseline, slant, pressure and margin. It is a well-known fact that signature of a person is also an important feature. Hence signature styles are examined as well. Among the above features, some are to be extracted from the document page as a whole, while others are to be extracted from the smaller units of writing such as lines, words, letters, and/or connected components. In our context, the first type of extraction is called global feature extraction, whereas the second type is called local feature extraction. And there are features which need to be extracted both globally and locally. Now technical details of the features and their implication in terms of Graphology are discussed in brief.

A. Zones of writing

For analysis purpose, a line of handwritten English text is divided horizontally into three zones, middle, upper and lower.

Distribution of a writing in these zones shows how the writer makes specific use of their mind, emotions and physical elements in environment. Large middle zone indicates demand for attention whereas small middle zone indicates reserved nature, intelligence, modesty and ability to concentrate. Variation in zonal distribution is illustrated in Fig. 1.

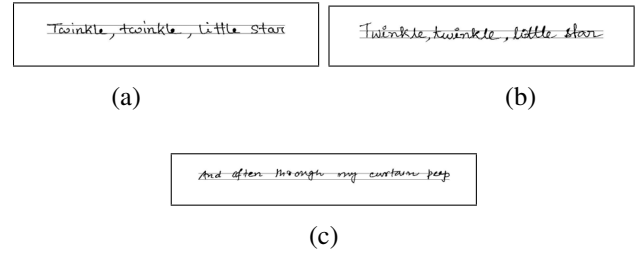


Fig. 1: Variation in zonal distribution of writings. (a) Small middle zone, (b) Large middle zone, (c) Almost all words are in upper zone

B. Size of writing

Size of a writing reveals nature, concentration and adaptability of a person. A writer having small handwriting is introvert in nature and possesses deep concentration. A writer having large handwriting is extrovert in nature, wants to draw attention and is a big planner. Medium size writing means average nature of expression and average ability to concentrate on things. A writer of with variable size writing is full of life and jovial in nature but not sure about his interest, rights and goal. It is also observed that in general children write bigger than adults, and among adults inexperienced or occasional writers write bigger than experienced or regular writers. A handwriting having a vertical length of three millimeter in middle zone and nine millimeter in three zones combined is considered to be a standard middle-sized writing. Otherwise, the writing is considered to be either large or small. Variation in writing-size is illustrated in Fig. 2.

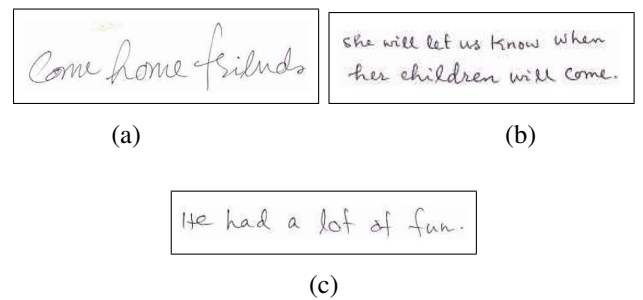


Fig. 2: Handwritings of different sizes. (a) Large writing, (b) Small writing, (c) Medium writing

C. Spacing in writing

Spacing in a writing is analyzed by blank spaces or gaps between two lines, two words and two letters. They are termed

respectively as line space, word space and letter space. Line space can be equal, unequal, wide, very wide, narrow, or very narrow. An ideal or standard word space is called “n-gap” or the width of the small letter “n”. Word space larger than “n-gap” is considered to be wide and less than that is considered to be narrow.

Equal spacing indicates confidence, regularity and maturity whereas unequal spacing indicates confusion, irregularity, risk-taking behavior. A person who uses wide space in his writing reflects has clear thought, pride, good taste, independence and ability to organize his work. A person who uses small or narrow space in his writing has passion, spends wisely but he is confused, impatient and unable to be alone. Variation in word spacing is illustrated in Fig. 3.

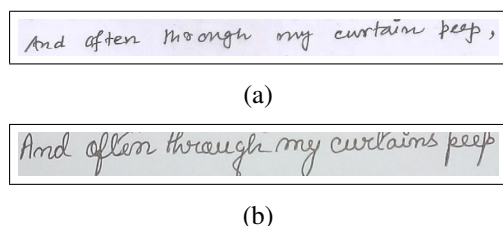


Fig. 3: Same text with different word space. (a) Large word space, (b) Small word space

D. Skew and baseline of writing

Skew of a writing is determined by the angle between a line of writing and the horizontal line. Skew-angle can be zero (horizontal), positive (upwards) or negative (downwards). It is related to the baseline of a writing.

Baseline is the pre-printed or imaginary line on which the letters reside. Depending on skew angle, it can be straight (horizontal), ascending or descending. In some writings baselines do not maintain a uniform skew angle. Rather they are convex, concave or erratic in shape.

Nature of a baseline helps to find out the emotional control and the reliability of the writer. Straight and true ascending are common baselines. They generally indicate positive mentality, hopeful behavior, disciplined and stable mind. Downward baseline indicates weakness of mind, pessimistic behavior, tiredness and sometimes suicidal tendency. Baseline feature is discussed in detail in [3].

E. Slant of writing

Slant of a writing is the angle in which the letters are inclined towards right or left. Vertical slant in writing indicates that the writer has independence, cool judgment and controlled emotions. The rightward slant indicates initiative character, social achievements, intense bonding of emotion. The leftward writing indicates selfishness, introversion and inexpressive emotion of the writer. Finally, slant of variable nature indicates the unstable behavior.

Slant is determined by the angle between the axis of the letter and the vertical axis. If each letter is at 0 degree with the

vertical axis or 90 degree with the horizontal axis, then writing has vertical slant. If each letter is at an angle less than 90 degree with the horizontal axis, then writing has rightward slant whereas if each letter is at an angle greater than 90 degree with the horizontal axis, then writing has leftward slant. Variation in slant is illustrated in Fig. 4.

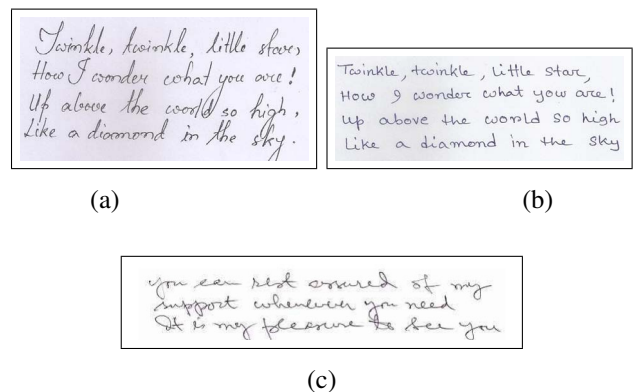


Fig. 4: Same text with different slants. (a) Slant to the right, (b) Almost vertical slant, (c) Slant to the left

F. Pressure of writing

Pressure of a writing indicates the force by which the writer has written, as well as his grip of pen. It reveals emotional energy of the person. Pen pressure may be heavy, light or medium. Heavy pressure implies self-assertive-ness, dynamicity, anger, energy, activeness, anxiety and alertness. Light pressure implies passivity, calmness, lack of intensity and illness. Medium pressure implies feelings of moderate intensity. Variation in pressure is illustrated in Fig. 5.

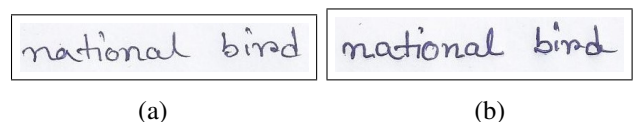


Fig. 5: Same text with different pressure: (a) Light Pressure, (b) Heavy Pressure

G. Margin of writing

White spaces surrounding a page of writing are called margins. On white pages the margins can be drawn as vertical lines on the left and right of the page and horizontal lines on top and bottom of the page. But most of the times, the lines are not drawn explicitly but kept in the mind of the writer. Margins indicate past and future, intelligence, adjustments, truthfulness and fastness. If there is constant left margin, then writer has good manners and constant behavior. If it is irregular, then writer is careless about his dressing and behavior. Constant right margin shows good judgment and the ability to take decision. Top margin tells about convention. Less space on bottom side shows laziness and indecision of writer.

H. Signature of a writer

Handwritten signature of a person conveys a lot of information about the person. Two signatures of a person are very much similar and signatures of two persons are considerably different. But signature of a person may change over time because of change in physical and mental states of the person. So considerable intra-personal structural variations in signatures may exist and they may not be repetitive in nature. Graphologists compare the signature of a writer with his handwriting, and examine for similarity, legibility, skewness, slant angle, pressure etc. The position of signature on the paper as well as special style of signing are also important. Feature extraction from signatures is not an easy and trivial task due to italicized and unconventional writing styles. Example of some personality traits with their corresponding signatures are given in [3].

IV. OUR WORK

Handwritten documents are collected from a number people. The documents are digitized by scanning and then scanned documents are binarized by thresholding. After that, denoising filters are applied on the binary images.

For global feature extraction the denoised image of a page of handwritten text is taken as input. For local feature extraction the page is partitioned into lines of texts and then every line is partitioned into a number of connected components. The components are used as units of local feature extraction. In this section, we present algorithms for extraction of various features along with illustrations.

A. Calculation of size

As discussed in section III-B, size of handwriting varies from writer to writer. It is noted that a writing with large sized letters occupies greater area and hence more number of black pixels than a writing with small sized letters. So area and number of black pixels can be used to calculate the size of writing.

In our work, size is calculated both globally and locally. It is calculated globally by (i) finding out the total area of the writing in a page, and (ii) counting the total number of black pixels in the writing. For local calculation of size, a page of writing is divided into lines of texts and each line of text is divided into connected components of words or their parts. So size is detected locally for each connected component by (i) finding out the area, and (ii) counting the number of black pixels in it.

Global area is calculated by product of the height and width of the bounding rectangle of the writing in a page. This rectangle is generally called the bounding box. Height of a bounding box is determined by the difference of row numbers of top and bottom rows where the writing starts and ends in a page respectively. Width of the bounding box is determined by the difference of column numbers of leftmost and rightmost columns of the writing. Area of a connected component is calculated by product of the height and width of the bounding box of the component.

The algorithm to calculate the height of a connected component is given below.

- 1) Start
- 2) Labeled connected components are detected
- 3) For each connected component
 - a) mx_rw = The number of the row containing the bottom most black pixel of the component
 - b) min_rw = The number of the row containing the topmost black pixel of the component
 - c) $height = mx_rw - min_rw$
[End of for loop]
- 4) End

The algorithm to calculate the width of a connected component is given below.

- 1) Start
- 2) Labeled connected components are detected
- 3) For each connected component
 - a) mx_col = The number of the column containing the rightmost black pixel of the component
 - b) min_col = The number of the column containing the leftmost black pixel of the component
 - c) $width = mx_col - min_col$
[End of for loop]
- 4) End

The algorithm to count the number of black pixels in an area of m rows and n columns is given below.

- 1) Start
- 2) $black_pxl_cnt = 0$
- 3) for $i = 1$ to m
 - for $j = 1$ to n
 - if (pixel at position (i,j) is black)
 $black_pxl_cnt = black_pxl_cnt + 1$
- [End of if]
- [End of inner for loop]
- [End of outer for loop]

4) End

Now we illustrate size calculation with figures. In Fig. 6, (a) and (b) show a line of text written by two writers. It is manually verified that the writing in Fig. 6(a) has smaller size than one in Fig. 6(b). Our calculation shows that there are 3953 and 6698 black pixels in Fig. 6(a) and Fig. 6(b) respectively. Hence the first one has less number of black pixels than the second one (the difference is 2745). So experimental results agree with physical verification.

In Fig. 6, (c) and (d) are used to compare size-variation in writing of a single person. In this figure, the word "Twinkle" is written twice by the same person. In (c) the writing is slightly smaller than that in (d). The difference in count of black pixels in (c) and (d) is small $2031 - 1881 = 150$.

It is mentioned in subsection III-B that the size of a normal writing is 3mm in middle zone (eg. 'a', 'e', 'o') and 9mm in three zones combined for full length letters (eg. 'f'). We have mapped the millimeter scale to pixel scale and obtained that

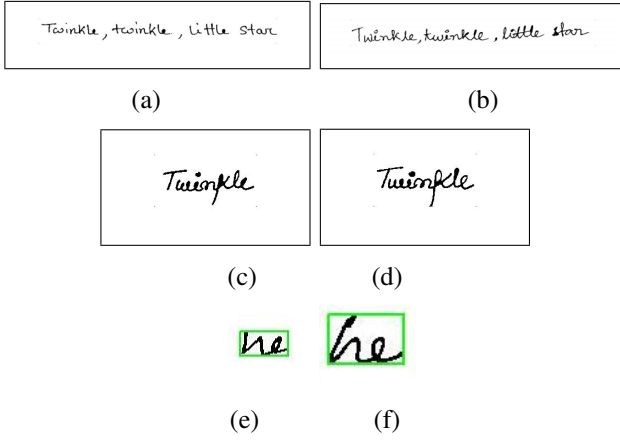


Fig. 6: Illustration of inter-personal and intra-personal variation in handwriting-size. (a) and (b) A line of text written by two writers, (c) and (d) Same word written twice by same writer, (e) and (f) Same word written by two writers

on average 3mm correspond to 25 rows and 9mm correspond to 75 rows.

Now in Fig. 6, (e) and (f) show the word “he” written by two different writers. The width and height of writing in (e) are 47 columns and 23 rows, whereas that in (f) they are 55 columns and 36 rows respectively. If we compare with the benchmark size both of the handwriting lie within the normal scale.

B. Calculation of spacing

As discussed in subsection III-C, spacing or gaps in a writing can be analyzed by line space, word space or letter space. In our work, spacing is calculated by finding out the gaps between connected components in a line of writing. For that the connected components in a line are detected at first. Then the number of white columns (columns with no black pixels) between consecutive components are counted by tracing the image matrix column wise. The algorithm is given below.

- 1) Start
- 2) Labeled connected components are detected
- 3) For each connected component C_i
 - a) X_i =Column number of the left most pixel of component C_i
 - b) X_{i-1} =Column number of the right most pixel of component C_{i-1}
 - c) $Space_i = X_i - X_{i-1}$ [End of for loop]
- 4) End

Now we illustrate space calculation with Fig. 7. In this figure, (a) and (c) show the same line of text written by two writers, and (b) and (d) show these lines broken into connected components. Significant end points of components are designated in both (b) and (d) by $a, b, c, d, e, f, g, h, i$ and j . Calculated number of white columns between consecutive points are also shown.

It is easily seen that writing in (a) has wider gaps than the

writing in (c). The observation is in accordance with numerical calculation. For example, in (b) the gap between the points a and b is 75 white columns, whereas in (d) the same gap is 16 white columns. If all the gaps are noted in this manner, it is easily concluded that the first writer gives more spaces between each word while writing than the second one.

C. Detection of skew angle

As mentioned in previous section, skew angle is the angle between the baseline and the horizontal line. In [12] a method to find skew angle is proposed. The method is used in our work and given below as an algorithm.

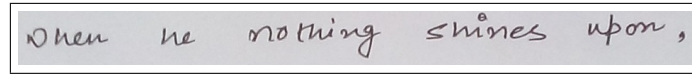
- 1) Start
- 2) for $i = 1$ to no.of rows
 - for $j = 1$ to no.of columns
 - if (image(i,j)pixel is black)
 $sum_u(i) = sum_u(i) + u(i)$
 $sum_v(i) = sum_v(i) + v(i)$
- [End of if]
- [End of inner for loop]
- [End of outer for loop]
- 3) $U = sum_u(i) / \text{Total no of black pixels}$
 $V = sum_v(i) / \text{Total no of black pixels}$
- 4) for $i = 1$ to no.of rows
 - for $j = 1$ to no.of columns
 - if (image(i,j)pixel is black)
 $sum_sq_u(i) = sum_sq_u(i) + (u(i) - U)^2$
 $sum_sq_v(i) = sum_sq_v(i) + (v(i) - V)^2$
- [End of if]
- [End of inner for loop]
- [End of outer for loop]
- 5) $U^2 = sum_sq_u(i) / \text{Total no of black pixels}$
 $V^2 = sum_sq_v(i) / \text{Total no of black pixels}$
- 6) for $i = 1$ to no.of rows
 - for $j = 1$ to no.of columns
 - if (image(i,j)pixel is black)
 $mul = mul + (u(i) - U) * (v(i) - V)$
- [End of if]
- [End of inner for loop]
- [End of outer for loop]
- 7) $UV = mul / \text{Total no of black pixels}$
- 8) The following matrix will be formed now:

$$M = \begin{bmatrix} U^2 & UV \\ UV & V^2 \end{bmatrix}$$

- 9) Orientation of the least eigen vector of the previous step matrix is calculated and this gives the required skew angle.

10) End

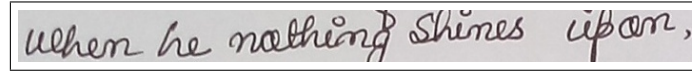
In the Fig. 8 two different handwriting are provided with two different skew angle. These two skew angles are positive angles and in the Fig. 9 the image is represented with a negative skew angle.



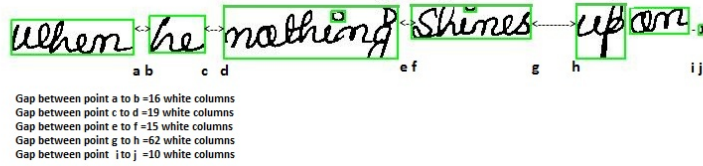
(a)



(b)



(c)



(d)

Fig. 7: Illustration of inter-personal and intra-personal variation in spacing. (a) A line written by writer 1, (b) Connected components in writing 1 (c) Same line written by writer 2, (d) Connected components in writing 2

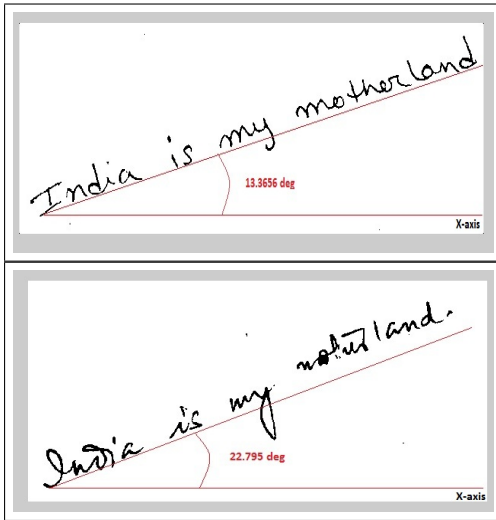


Fig. 8: Handwriting with positive skew angle

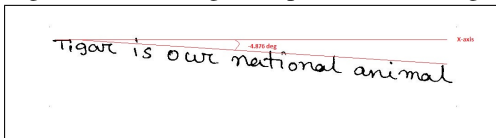


Fig. 9: Handwriting with negative skew angle

D. Detection of slant angle

Slant angle is determined by shear transformation procedure. In [13] the slant angle detection procedure is described. That procedure is represented here in a pixel level view.

- 1) Start
- 2) Perform vertical shear using affine transformation for few angles on the image.
- 3) Calculate vertical projection profile by finding maximum vertical frequency.
- 4) The shear factor for which the maximum vertical frequency is found, is taken into consideration.
- 5) Finally the slant angle is calculated by using the following formula:

$$\text{slant_angle} = \tan^{-1}\left(\frac{1}{\text{sh_fac}}\right)$$
- 6) End

In the Fig. 10 two different handwriting with two different slant angles are described. Both the writings are creating angles with Y-axis.

E. Calculation of Pressure

Pressure of handwriting basically depends both on the writer and the pen or pen gripping. While analyzing manually, this is done by examining the back side of the page. If the handwriting is done with much pressure, the back side of the page will show all the curvature of writing very clearly (we can also feel it by touching our fingers), otherwise the writing will not have much effect on the back side.

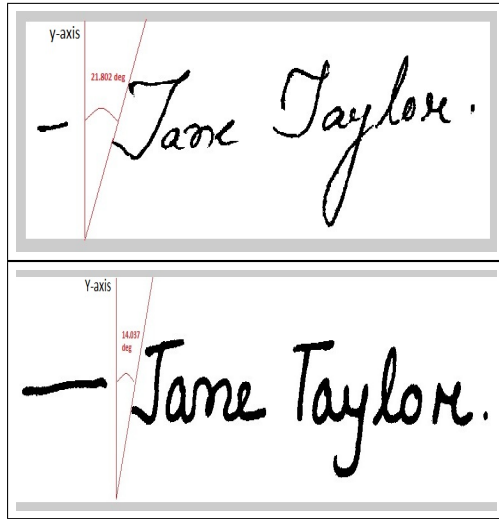


Fig. 10: Handwriting with different Slant angle

For automatic calculation of pressure, writing area is used. It is noted that the writing area is large when the pressure is high, whereas the writing area is small when the pressure is low. Total number of black pixels of 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. The algorithm is as follows.

- 1) Start
- 2) $Before_thin$ = Total number of black pixels calculated before thinning
- 3) Use ZS algorithm for thinning
- 4) $After_thin$ = Total number of black pixels calculated after thinning
- 5) $Diff = After_thin - Before_thin$
- 6) End

The Fig. 11 shows a handwriting with medium pressure and Fig. 12 shows a handwriting with high pressure. The pixel count also reflects the difference.

F. Signature analysis

Following features have been extracted and analyzed from handwritten signatures: aspect ratio, occupancy ratio, signature area, number of white columns, maximum horizontal frequency, maximum vertical frequency, number of straight line segments in the signature.

V. CONCLUSION

This work concentrates on feature extraction from handwritten document images for personality analysis. Space, size, slant, skew and pressure are some of the features analyzed here. The features are described from algorithmic view point and images are treated as collection of black and white pixels. Results are also given here with pixel wise explanation. In

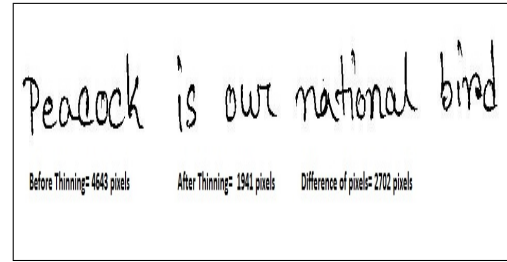


Fig. 11: Handwriting with low pressure

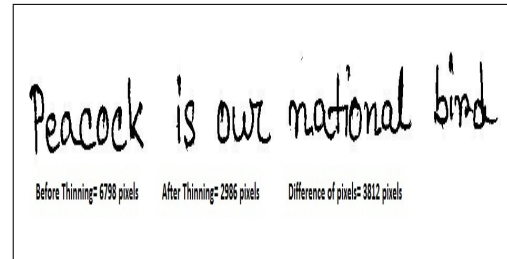


Fig. 12: Handwriting with high pressure

future, more features will be extracted and analyzed in detail and a classifier will be used to classify the results. Finally our aim is to design an automated system which can efficiently and quickly read handwritten images and analyze personality traits the writer.

REFERENCES

- [1] 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.
- [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. Kedar, V. Nair, and S. Kulkarni, "Personality identification through handwriting analysis: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, pp. 548–556, January 2015.
- [4] 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.
- [5] J. Fisher, A. Maredia, A. Nixon, N. Williams, and J. Leet, "Identifying personality traits and especially traits resulting in violent behavior through automatic handwriting analysis," in *Proceedings of Student-Faculty Research Day, CSIS, Pace University*, May 2012.
- [6] 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.
- [7] Aarti, B. Valsang, and C.M.Jadhav, "Human character recognition by handwriting using fuzzy logic," *International Journal of Engineering Sciences and Research Technology*, pp. 513–517, June 2014.
- [8] R. Kacker and H. B. Maringanti, "Personality analysis through handwriting," *GSTF Journal on Computing (JoC)*, vol. 2, pp. 94–97, April 2012.
- [9] P. Joshi, "Handwriting analysis for detection of personality traits using machine learning approach," *International Journal of Computer Applications (0975-8887)*, vol. 130, November 2015.
- [10] C. H. N and D. K. R. AnandaKumar, "Artificial neural network for human behavior prediction through handwriting analysis," *International Journal of Computer Applications (0975-8887)*, vol. 2, May 2010.
- [11] D. J. Antony and O.F.M.Cap, *Personality Profile Through Handwriting Analysis*. Tamil Nadu, India: Anugraha Publication, 2008.

- [12] M. K. Kalera, S. Srihari, and A. XU, "Offline signature verification and identification using distance statistics," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 7, pp. 1339–1360, 2004.
- [13] 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.