Handwriting Analysis based on Segmentation Method for Prediction of Human Personality using Support Vector Machine

Shitala Prasad
Department of Information
Technology
Indian Institute of Information
Technology
Allahabad, India

Vivek Kumar Singh
Department of Information
Technology
Indian Institute of Information
Technology
Allahabad, India

Akshay Sapre
Department of Information
Technology
Indian Institute of Information
Technology
Allahabad, India

ABSTRACT
Handwriting analysis is a method to predict personality of an author and to better understand the writer. Allograph and allograph combination analysis is a scientific method of writer identification and evaluating the behavior. To make this computerized we 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. For this experiment 100 different writers were used for different handwriting data samples. The proposed method gives about 94% of accuracy rate with RBF kernel. In this paper an automatic method has been proposed to predict the psychological personality of the writer. The system performance is measured under two different conditions with the same sample.

Keywords
Image processing, Segmentation, Graphology, Handwriting Analysis, Support Vector Machine, Personality Traits, Human Behavior Analysis, Psychology

1. INTRODUCTION
Personality identification of a human being by their handwriting is an old technique. Earlier these were done manually by spending lot of time to predict the nature of the person. Handwriting is brain writing, representing the mental status of the person. And handwriting analysis is a projection technique as the body language that profiles the human behavior in areas of the social skills, achievements, thinking styles, or work habits. Handwriting also depicts the possible ways of person's transactions with stress. It is a system to study frozen graphic structures which is being generated in the author's brain and is placed on the paper in a cursive or printed handwriting style which is different with other authors comparing the personalities and their potential for problem solving.

This paper is based on the same old technique handwriting detection and prediction of the characteristic of the person whose handwriting is. For this the first step is to teach the computer what handwriting is and how to segment it. Compared to other methods, such as manual document handwriting analysis, automated handwriting analysis is very fast and accurate. The rule-based approach [15] is also used for the handwriting analysis and prediction of human personality. Collecting digital samples of handwriting and computer prediction is very low-cost and convenient method. One can easily give the digital sample of his/her handwriting to a computer and it calculates the features using the image processing techniques and predicts the nature of the writer.

In this paper an automated method has been proposed to predict the personal behavior of an individual from their handwriting analysis in digital form. To predict the actual personality of the individual there are various features [2], such as slant, size, pressure, upper zone or case(as in I, t, h, S, etc), lower zone (as in g, q, y, z, etc), word spacing, line spacing, page margins, middle zone or case(as in a, o, c, s, e, etc), arcade, garland, angle, thread, wavy line (written by authors who are mentally mature and are skilful), and many others [2, 10, and 11]. But the proposed system has only six features among them discussed above. These features are size of the letters, slant of words and letters, baseline, pen pressure, spacing between letters and words as they are enough to predict the behavior of the person. The main attention of this paper is feature extraction and its classification. All features are extracted automatically from the digital image of handwriting. These samples are then input to the support vector machine for classification. The algorithm proposed is simple and easy to implement and use. MATLAB is the tool used for the same.

2. PROPOSED METHODOLOGY
Handwriting analyst called graphologist analyzes the handwriting on a piece of paper written by the individual writer which is very time taking. And the level of accuracy in the result is totally depending on the knowledge and experience of the graphologist. There were many works in this proposed field [7, 8, and 9] which has three main steps: pre-processing, feature extraction, and classification. Following the same to atomize the procedure the various steps needed to be followed as mentioned in Figure 1.
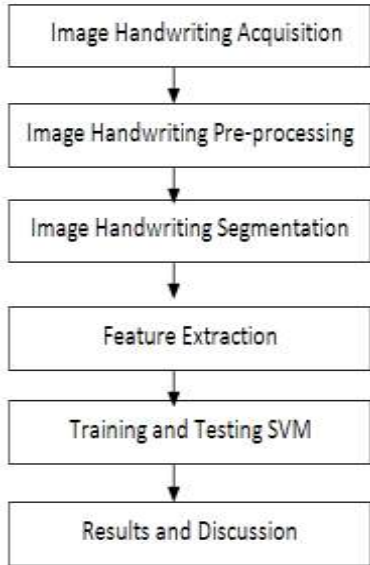
Figure 1. Block Flow Diagram of proposed system.

2.1. Image Handwriting Acquisition and Database

Handwriting image samples of different individuals are used in this research which is digitally collected by scanning the handwritings of 100 different writers. Each of them was asked to write a text document of simple 70-80 words in running hand. Most of the handwritings are cursive but few of them are printed handwriting. The samples were written on a plane paper without any margin.

2.2. Image Handwriting Pre-Processing

In pre-processing stage, the handwriting image is pre-processed to remove the noise by applying local thresholding and followed by resizing the sample to the correct orientation. The pre-processing involves opening of the digital image and then smoothing it. Opening removes the unwanted characters (dots, etc.) in the digital document. The next process is to segment the digital document in various segments, such as word segments, letter segments, and line segments. In Figure 2, the word segmentation in the document is shown clearly.

Figure 2. Image Handwriting pre-processing.

2.3. Image Handwriting Segmentation

In Handwriting image segmentation digital handwriting is segmented into three different types of segments, i.e. word segmentation, letter segmentation and line segmentation, each used for different processing.

2.3.1. *Word Segmentation*

This segmentation process is used to segment the words in digital handwriting document to calculate different features related to words indicating the disposition towards criticism, and towards argument.

2.3.2. *Letter segmentation*

Here segmentation is performed on each letters in the word in digital handwriting document of each individual. This segmentation is used in feature calculation related to letter for the prediction of the personality of individuals. Figure 3, shows the letter segmentation which is future used to calculate various slants in letters indicating the openness of sentiment and consequently of the intelligence.

Figure 3. Letter segmentation after word segmentation.

2.3.3. *Line Segmentation*

This line segmentation is used to find baseline features to judge the writer's emotional stability and dispositions in the baseline of the writing.

Segmented image is then processed to calculate some numerical values mathematically to classify the personality of writer's on psychological factors.
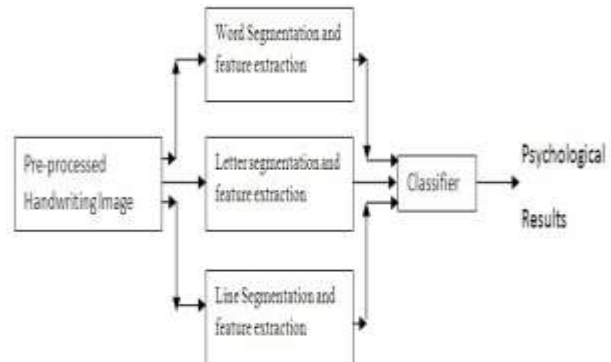

Figure 4. Segmentation and feature extraction.

2.4. Feature Detection and Psychological Representation

Feature extraction is a technique of dimensionality reduction from a high dimensional input data. This reduced output data is a transformation of the high dimension input data represented as feature vector. In our case, features are the six important factors on which a graphologist predicts the behavior of human being. These features are explained in more detail below.

2.4.1. *Size of the Letters*

The size of the letters in a document represents the author's desire to be noticed. In handwriting, large and bold writing indicates that the author wants to say "notice me". While in case of small writing author is less desired to be noticed. Middle writing is the normal writing indicating the desire to be fit in the running world.

26

The size of handwriting is judged by a benchmark of 3mm as normal writing and full height of 9mm. Other than this writing is classified as large or small writing. The letters are divided into three zones: lower case or zone (e.g. g, y), upper case or zone (e.g. b, I, t), middle case or zone (e.g. a, c, e). Figure 5, explains the same more clearly.
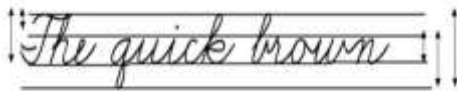

Figure 5. Different zone of letters

2.4.2. *Slant of Words and Letters*

Slant in handwriting indicates the emotional interactions of the author. There are three classes in this, right slant, left slant and vertical. Along with this some authors handwriting are varying. To calculate the slant in world and letter we use simple trigonometric formula, as below:

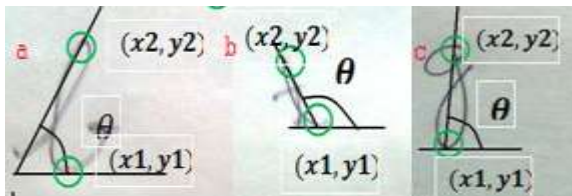$$\theta = \tan^{-1} \frac{y2-y1}{x2-x1} \qquad \text{--- (1)}$$

Figure 6. Slant of Letters (a) right slant, (b) left slant, (c) vertical.

Here θ is the calculated angle for points $(x_1 y_1)$ and $(x_2 y_2)$. We considered θ_0 a standard reference angle to find the slant. In simple words, θ_0 is 90°.

2.4.3. *Baseline*

The emotional stability and disposition of writer is judged by the baseline in the handwriting as in Figure 7. To calculation this equation (1) is used. In this, line spacing is also considered, which shows that the author wishes to stand back and take a large view of the life.
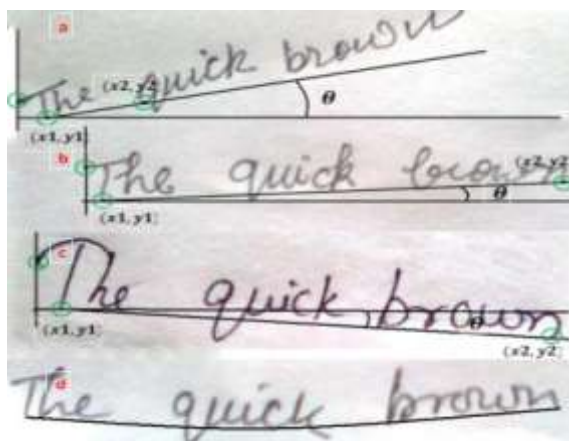

Figure 7. Type of baselines (a) rising, (b) straight, (c) falling, (d) erratic.

2.4.4. *Pen Pressure*

Pen Pressure in handwriting shows the physical and mental levels. Heavy pressure indicates commitment and taking things seriously. Light pressure shows sensitivity to atmosphere and also empathy to others. To measure this simple thresholding technique is used. The documents darker than the threshold value, th_0, the document is heavy pressure else it is light pressure.

2.4.5. *Spacing between Letters*

Letter spacing in handwritten documents represents the openness of sentiment and consequently of the openness of intelligence.

2.4.6. *Spacing between Words*

Similar to letter spacing word spacing also indicates the disposition towards criticism, and towards argument. It adds to person's personality other then baseline.
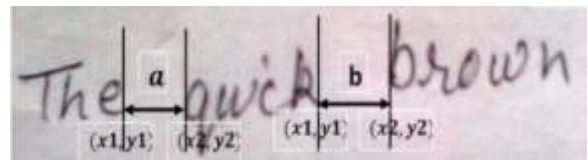

Figure 8. Represents word spacing in the document.

Each steps or process explained above, gives some numerical value, which is -1, 0, or +1 depending on the threshold and the value obtained. Combining all these numerical values together in an SVM format and forwarded to the SVM classifier to classify the result and predict the author's personality.

3. CLASSIFICATION

The classifier here used is SVM as it gives good result with more accuracy. SVM is time efficient to than neural network as tested. It analyses the data and recognize patterns in the data. Multiclass-SVM aims to assign labels to instances by using support vector machines, learning algorithm, where the different classes are drawn from a finite set of elements to clearly separate them. SVM takes training data x, of some space, R^D, where, $x = \{x_1, x_2 \ldots, x_N\}$ such that $x \in R^D$. We also give their class label $y = \{y_1, \ldots, y_N\}$ as to which x_i belongs to. In binary SVM $y \subseteq \{-1, +1\}$ while in case of multi-class SVM the y ranges 1 to N, where, N is the number of classes x_i can belong to.

The linear SVM classification function is represented by $f(x) = w.x + b$, which relates to a separating hyperplane $w.x + b = 0$. The hyperplanes separating the training data samples with the maximum margin such that the closest point to plane is $\frac{1}{||w||}$ [2].
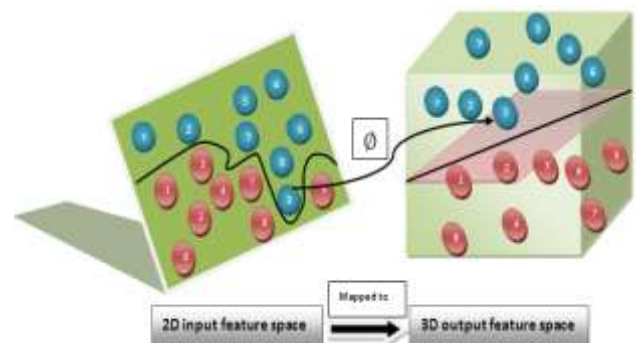

Figure 9. Principle of Support Vector Machine (SVM).

We have used RBF kernel as our features are non-linear and it results better than linear kernels and other non-linear kernels such as polynomial kernel and sigmoid kernel. Radial Basis Function non-linearly maps the input space to higher dimensional unlike the linear kernel. RBF kernel is defined as:

$$K(x_i, x_j) = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) \forall \gamma > 0 \quad \text{--- (2)}$$

where, $K()$ is a kernel function mapping x_i and x_j of input space to high dimensional feature space and γ is the kernel parameter changing which may give better results [3]. It adds "bumps" with all input data sets which make it easy to separate the data. In Figure 9, $\emptyset$ is the mapping function of two different spaces as below:

$$\emptyset: R^D \rightarrow R^{D'} \quad \text{--- (3)}$$

where, D and D' are dimensions of the input space, I and the feature space, F respectively.

The classification can be performed in following three steps:

1. First, the input features are formulated as input vectors in some feature space.
2. Map these feature vectors to the higher dimension feature space using RBF kernel function.
3. Then a division global hyperplane is computed to separate the feature space optimally to the classes of the training vector samples.

4. EXPERIMENTS AND RESULTS

We conducted the experiments in two steps. The first step A, uses two-third of data samples as training data and one-third as the test data. And the second step B, involved one data as a test data and the rests where used to train the system. The result in both the cases is near about same with the difference of 3%.

In first step, each 100 writer's where asked to write the test documents and were saved in digital form. Then each individual's manual psychological analysis was made. The two-third of the handwriting data sample and its psychological results are used to learn the system and the rest one-third handwriting data samples are tested for the accuracy. It results in 90.3% of accuracy. While in the second step more or less we use the same technique to the same data sample to train the system but tested with only one sample as the test sample. In this case the accuracy graph jumped from 90.3% to 93.86% which is a very good result.

The Table 1 show the criteria to classify various writing styles and Table 2 below shows the handwriting styles and its corresponding psychological personality traits the writer may have in more detail as Champa H. N. and K. R. AnandaKumar [4], predict the personality of an individual form the letter't'.

Table 1. Criteria for handwriting style classification.

S.No.	Writing Category	Sub-category Criteria
1.	Size of Letters	Large: size > normal size Small: size < normal size Medium: size = normal size

2.	Slant of Words and Letters	Right: $\theta > \theta_0$ Left: $\theta < \theta_0$ Vertical: $\theta = \theta_0$
3.	Baseline	Raising: $\theta > 0$ Falling: $\theta < 0$ Straight: $\theta = 0$ Erratic: otherwise
4.	Pen Pressure	Light pen: threshd $< th_0$ Heavy pen: threshd $> th_0$
5.	Spacing between Letters	Far: > avg_standard_L Close: < avg_standard_L
6.	Spacing between Words	Far: > avg_standard_W Close: < avg_standard_W

Table 2. Psychological personality traits predicted by various handwriting styles of writer.

S.No.	Writing Categories	Psychological Personality Behavior
1.	Large Letters	Likes being noticed, stands out in a crowd
2.	Small Letters	Introspective, not seeking attention, modest
3.	Medium Letters	Adaptable, fits into a crowd, practical, balanced
4.	Right Slant	Sociable, responsive, interested in others, friendly
5.	Left Slant	Reserved, observant, self-reliant, non-intrusive
6.	Vertical	Practical, independent, controlled, self sufficient
7.	Light Pen Pressure	Strong emotion, successful, emotion last for long time
8.	Heavy Pen Pressure	Try to avoid energy draining situations
9.	Far Spacing Letters	Openness of sentiment and intelligence
10.	Close Spacing Letters	Closeness of sentiment and intelligence
11.	Far Spacing Words	"give me breathing space"
12.	Close Spacing Words	Wish to be with others, intrusive
13.	Raising Baseline	Optimistic, upbeat, positive attitude, ambitious, hopeful
14.	Falling Baseline	Tired, overwhelmed, pessimistic, not hopeful
15.	Straight Baseline	Determined, stays on track, self motivated, controls emotions, reliable, steady

16.	Erratic Baseline	Waving, lacks definite direction, emotionally unsettled, unpredictable

The performance of the proposed system is comparatively very good as it takes less time to train and test and the accuracy rate is also very high. Table 3 below pictures the accuracy rate of the proposed system.

Table 3. Accuracy rate table.

	A	**B**
Polynomial kernel	87.9%	89.1%
RBF Kernel	**90.3%**	**93.86%**

5. CONCLUSION AND FUTURE WORK

This paper proposes an automated method to predict the personality of an individual by his/her handwriting sample using SVM machine learning algorithm. SVM takes the size of the letters, pen pressure, baseline, letter spacing and word spacing and the most important slant of the letter and word in a document as an input and gives their psychological behavior as the output for each individual writer. The various parameters are calculated by simple use of trigonometry and thresholding techniques.

This is a real-time handwriting analysis system which is writer independent. For future, the features can be increased for more correct results such as page margins, arcade, garland, thread, etc. We want to enhance the proposed system be used as lie detector on the basis of writer's handwriting along with writer identification.

6. REFERENCES

[1]. Tripathy N. and Pal U. 2006,"Handwriting segmentation of constrained Oriya text", Sadhna, Vol.31, Part 6, pp. 755-769.

[2]. Graphology – Handwriting Analysis http://www.businessballs.com/graphologyhandwritinganalysis.htm.

[3]. Lei Zhang, Fuzong Lin, Bo Zhang, "Support Vector Machine Learning for Image Retrivel", Image Processing, IEEE 2001 International Conference, page(s): 721 - 724 vol.2.

[4]. Hua Hu, Jing Ye, Chunlai Chai, "A Talent Classification Method Based on SVM", International Symposium on Intelligent Ubiquitous Computing and Education 2009, Page(s): 160 – 163.

[5]. Champa H. N., Dr. K. R. AnandaKumar, "Artificial Neural Network for Human Bahavior Prediction through Handwriting Analysis", International Journal of Computer Applications (0975-8887) Volume 2 – No.2, May 2010.

[6]. Ameur BENSEFIA, Ali NOSARY, Thierry PAQUET, Laurent HEUTTE, "Writer Identification By Writer's Invariants", Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02) 0-7695-1692-0/02 $17.00 © 2002 IEEE.

[7]. S.H. Cha, S. Srihari, "Multiple Feature Integration for Writer Verification", 7th International Workshop on Frontiers in Handwriting Recognition; IWFHR VII, Amsterdam, The Netherlands, pp 333-342, 2000.

[8]. U.V. Marti, R. Messerli, H. Bunke, "Writer Identification Using Text Line Based Features", Proc. ICDAR'01, Seattle (USA), pp 101-105, 2001.

[9]. A. Srihari, S. Cha, H. Arora, S. Lee, "Individuality of Handwritig: A Validity Study", Proc. ICDAR'01, Seattle (USA), pp 106-109, 2001.

[10]. N Mogharreban, S Rahimi, M Sabharwal, "A Combined Crisp and Fuzzy Approach for Handwriting Analysis", IEEE Annual Meeting of the Fuzzy Information, 2004, vol.1, pp. 351-356.

[11]. Champa H N, K R AnandaKumar, "A Scientific Approach to Behavior Analysis through Handwriting Analysis", National Conference on Research Trenda in Information Technology, S R K R Engineering College, Bhimavaram, Andhra Pradesh, 2008.

[12]. Deodhare Dipti, Suri NNR Ranga, Amit R. 2005, "Preprocessing and Image Enhancement Algorithms for a Form-based Intelligent Character Recognition System", International Journal of Computer Science and Application, Vol 2 No.2, pp. 131-144.

[13]. Daekeun You and Gyeonghwan Kim, "An approach for locating segmentation points of handwritten digit strings using a neural network", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003).

[14]. Champa H N, K R AnandaKumar, " Writer's Personality Prediction through letter „y" using Generalized Hough Transform(GHT) ", 3rd International Conference on Information Processing, organized by the Society of Information Processing, Bangalore, Defence Institute of Advanced Technology and University Visveswaraya College of Engineering, Bangalore, 2009.

[15]. Champa H N, K R AnandaKumar, " Rule Based Approach for Personality Prediction Through Handwriting Analysis", 2nd International Conference on Biomedical Informatics and Signal processing, organized by Sai's BioSciences Research Institute Pvt. Ltd., 2009.