International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

A Hybrid Set of Handwriting Features for Handwritten Recognition

Bramara Neelima K, S Arulselvi

Abstract*: Handwriting of each person is unique since each person has their own unique and different style of handwriting. Handwriting verification can be performed in two ways, dynamic and static. The dynamic verification process is the writer dependent whereas the static verification process is the writer independent procedure. The features can be spatial, structural, statistical, geometrical, graphological, and from other feature extraction techniques. In this work, we are considering the combination of multilevel feature set for writer recognition and identification purpose. A dataset of different handwriting samples collected from 100 different writers is used for this experiment. A decision tree classifier with random forest implementation is used for recognition and identification of writer with 98.2% accuracy.

Keywords: Handwritten document, writer recognition, feature extraction, decision tree classifier.

I. INTRODUCTION

Handwriting of each person is unique since each person has their own unique and different style of handwriting. Therefore, handwriting is considered as one of the biometric now a days, means it can be used to verify person by studying the handwriting. Handwriting verification can be performed in two ways, dynamic and static. The dynamic verification process is the writer dependent whereas the static verification process is the writer independent procedure. Some general handwriting characteristics are specific shape of letters, the slope of the letters, regular or irregular spacing between letters, the pressure to the paper, average size of letters, thickness of letters, and rhythmic repetition of the elements. The characteristics of handwriting extracted process from the handwritten image are referred as feature extraction process and the characteristics are mentioned as features. The image features can also considered since handwriting is on static image. Various feature sets can be extracted, as visual features, gradient features, space structural features, statistical features, graphometrical features, geographical features, and other, for handwriting analysis.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Bramara Neelima K*, Research Scholar, Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Chennai, India.
S Arulselvi, Research Supervisor, Associate Professor, Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Chennai, India.

To make a handwriting match, several characteristics are considered in handwriting analysis. They are height, width, size of letters, letter spacing, slant, separation, connecting strokes, beginning and ending of strokes, line quality, base line habits, unusual letter formation, diacritic placement, flourishments, embellishments, pen lifts, pen pressure and shading. The height, width and size of the letters simply analyses the proportions of the handwriting. Sometimes the letter is unusually tall or short. Letter spacing is referred to as amount of space put between letters. The letters could all be connected or spaced drastically. The way the writing tends to slant either to the right or to the left or straight (no slant) is referred as his/her slant. The most usual slant is to the right. The separation tells the way of writer. It tells about the writer usual way of connecting or giving space between letters. Similar to this the connecting strokes explain the strokes between the connection of words and the connection of lower letters and upper letters. The beginning of strokes explains where the writer begins the letter or word, and the ending of strokes explains how the writer ends the word or letter. The usual ending of the strokes may be either with a curl, upstroke or down stroke.

The thickness, flow and strength of the letters are measured by line quality. It explains whether the writer writes in shaky letters, flowing or amount of thickness in letters. Base line habits describe the writing tendency along with a reference line, generally called a baseline. The writer may tend to write above the line, on the line or below the line. An average writer would not add any extra curls, loops, unique capital or small letters. The peculiar writing style of a writer is considered in unusual letter formation. It expounds the unique style of writer in his writing habits. The individualities of the writer sometimes defined using his crosses and dots; this is termed the diacritic placement. The writing way and crossing of t's, and placement of dots on i's and j's styles the writer distinctiveness. The large loops of lowercase letters are referred as flourishments and the large swirls on lowercase letters are referred as embellishments. The pen lifts measures the way of the writer usual writing manner. It clearly explains where the writer lifts the pen; generally for which letter in a word or before writing a new letter in a sentence. For each writer it is unique; i.e. usually use the same pen lifts. Shading analyses the pen pressure given by the writer the most, either on the upstroke or on the down stroke. The script is thick or shaded where the writer applied the more pressure. In this work, the writer independent process is performed and the key characteristics of handwriting are considered as features, which are obtained from the static handwritten images.

II. RELATED WORK

Good Online handwriting analysis and verification is the simple, easiest procedure [1] than offline handwriting analysis. The handwriting features extracted from handwriting can be classified as local features, global features, geometric features, directional features, statistical features and structural features. Heutte [2] combine different structural and statistical features for handwritten character recognition. Aurora [3] combine different feature extraction methods such as shadow features, chain code and curve fitting features. Kimura [4] finds the suitable combination of features by genetic algorithm method. The structural features [5] represents the shape of the character capture the corners, vertical, diagonal and horizontal lines in gradient image. The gradient features informs about the flow orientation and variations using gradient directions. Leedham [6] extracted global features for the identification of handwritten digits. Wang [7] proposed directional features and analyze these features for Chinese character images. Rajiv Kumar Nath [8] describes the techniques to extract textual features from the handwritten document images. Zrei [9] proposed a new approach for feature extraction from handwritten images by obtaining the normal vectors of outer contour points of each connected component. Dynamic and contextual information [10] is extracted, combined and processed through hidden markov models. Shah [11] presented various feature selection and feature extraction techniques. Sumedha [12] proposed twelve directional features are used for handwritten text recognition and verification. The state of art of feature selection and extraction methods [13] on handwritten text are discussed, which are based on character level, word level, line level, and paragraph level. Graphology is the study of the physical traits and patterns of handwriting [14], research focuses on computer assisted handwriting analysis by considering the nine graphological features of handwriting. A novel approach [15] of machine learning technique to implement the automated handwriting analysis tool is proposed based on graphometric features.

III. RESEARCH METHODOLOGY

Handwritten is considering as one of the biometrics for writer identification, since it has rich set of information about the person such as uniqueness, emotional, mental characteristics. The identification can occur in different levels of characterization based on various set of features. The features can be spatial, structural, statistical, geometrical, graphological, and from other feature extraction techniques (like SIFT, SURF, FAST features). In this work, we are considering the combination of multilevel feature set for writer recognition and identification purpose. The proposed methodology contains regular image processing steps; such as handwritten image acquisition, pre-processing, feature extraction, and classification. The handwritten document samples are collected from 100 individuals and form a dataset for our work. The fixed size image is acquired by proper scanning of handwritten document by 300dpi scanner. The pre-processing on acquired image is carried out as sequence of following operations: noise removal by a median filter; removing unwanted characters by opening operation; smoothing the image; segmenting the image into multi-level segmentation methods as line segmentation, word segmentation and letter segmentation techniques. Feature vector is calculated from various features from segmented images and given to classification technique for writer recognition. The feature set consists of following features.

Size of letters: The size is defined with the height and width of the letter. The normal writing size is 9mm height and 3mm width. The style is considered as normal writing, large writing, or small writing based on simple comparison of size with regular norms. The size calculation is shown in figure1.

Aspect ratio: The aspect ratio is calculated as the ratio of width and height of the letter. The letter occupies in three zones upper zone, lower zone and middle zone. The aspect ratios of different zones are considered as feature.

$$Aspect\ ratio = \frac{width}{height}$$

***Baseline*:** The imaginary line draws connecting the bottoms of the middle zone letters is baseline of person's handwriting. The slope of the baseline reveals the person characteristics. Using polygonalization method, a polygon is drawn around a single line of handwriting. The slope of the polygon is the slope of baseline and that can be ascending baseline, descending baseline or straight baseline. The concept is shown in figure1.

***Slant of letters and words*:** Slant represents the writing style of the writer and also emotional direction. To find slant, a line is drawn between the lowermost and uppermost points of word or letter. The angle is calculated between points using the following formula. There are extreme right slant, right slant, extreme left slant, left slant and vertical slant for a word or a letter.

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

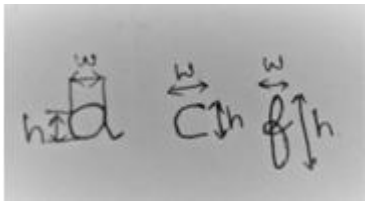
$$Extreme\ Right\ slant: \theta \gg \theta_0$$
$$Extreme\ left\ slant: \theta \ll \theta_0$$
$$Right\ slant: \theta > \theta_0$$
$$left\ slant: \theta < \theta_0$$
$$vertical\ slant: \theta = \theta_0$$

Where, Θ is calculated between points (x_1, y_1) and (x_2, y_2). The reference angle to compare is Θ_0 is 90^0, which is considered as vertical slant.

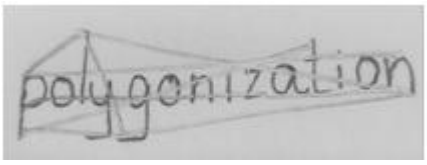
Spacing between letters and words: Regular or irregular spacing, between letters in a word, and between words, is calculated by applying grids on lines. Euclidian distance is measured between letters and words, which is used as classifying parameter.

$$Euclidean\ distance\ d = \sqrt{(x2-x1)^2 + (y2-y1)^2}$$
$$if\ dl > d_0\ then\ irregular\ distance$$
$$if\ dl \leq d_0\ then\ regular\ distance$$
$$if\ dw > d_0\ then\ irregular\ distance$$
$$if\ dw \leq d_0\ then\ regular\ distance$$

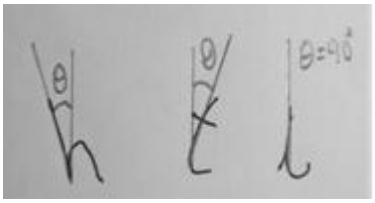
Where, dl is the Euclidean distance between letters, dw is the Euclidean distance between words, and d_0 is the standard distance measure.



(a)

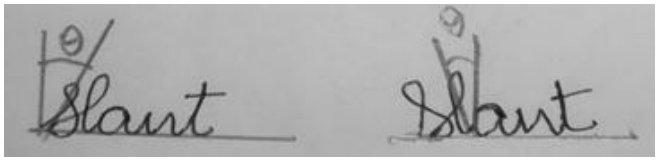


(b)

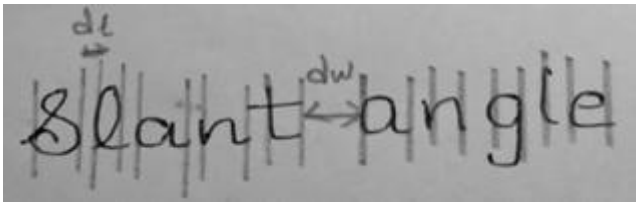


(c)

(d)



(e)



(f)

Fig.1 (a) size (height and width) of letters, (b) Drawing polygon around word, (c) slant of the letters, (d) baseline type; straight, ascending, descending, (e).slant of words, (f) spacing between the letters (dl) and spacing between words (dw)

***Pressure to paper*:** Pressure can be heavy pressure or light pressure and is calculated by applying thresholding technique. The heavy pressure has high threshold value results darker area and the light pressure has low threshold value results brighter area on paper. The local thresholding is considered for initial threshold (th_0) value.

$$Heavy\ pressure : threshold\ t > th_0$$

$$Light\ pressure : threshold\ t < th_0$$

Mean, variance, and Standard deviation of letters and words:

The mean is defined as the average value of the region (word image or letter image) and is mathematically expressed as follows.

$$mean\ (\mu) = \frac{\sum_{x=0}^{w-1}\sum_{y=0}^{h-1} I(x,y)}{w \times h}$$

Where, I(x,y) is the word or letter image and w,h are the size coordinates of the image.

The standard deviation is the measure of the dispersion of a set of data from its mean and is mathematically expressed as follows.

$$standard\ deviation\ (\sigma) = \sqrt{\frac{1}{w \times h}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(I(x,y)-\mu)^2}$$

Where, μ is the mean of the image region.

The variance is defined as the expectation of the squared deviation and is expressed as square of standard deviation.

$$variance = \sigma^2 = \frac{1}{w \times h}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(I(x,y)-\mu)^2$$

Where, μ is the mean of the region and σ is the standard deviation.

***Elasticity*:** one of the shape energy features is the elasticity, which is defined as the first order derivative of each pixel, mathematically expressed as follows.

$$el = \sum_{i=0}^{n-1}[(x_{i+1}-x_i)^2 + (y_{i+1}-y_i)^2$$

Where, i is the order of the pixels

***Curvature*:** Another shape energy feature is the curvature, which is defined as the second order derivative of each pixel, mathematically expressed as follows.

$$cr = \sum_{i=2}^{n-2}[(x_{i+1}-2x_i+x_{i-1})^2 + (y_{i+1}-2y_i+y_{i-1})^2$$

***Skewness*:** Skewness is the measure of the degree of asymmetry of the distribution of a variable about its mean. If Skewness is positive, the data are spread out more to the right than to the left. If skewness is negative, the data are spread out more to the left than to the right.

IV. DISCUSSION

All these handwriting features have numerical values and are combined to form a feature vector. These features are used in classification stage to recognize or identify the writer. The classifier used in this work is the random forest decision tree classifier. The features obtained from feature extraction stage are given to random forest decision classifier to do various comparisons and being trained. Once the dataset is trained; it will verify and identify the new handwriting sample. The accuracy of the proposed methodology is 98.2%.

V. CONCLUSION

Handwritten is considering as one of the biometrics for writer identification, since it has rich set of information. Various feature sets can be extracted, such as visual features, gradient features, space structural features, statistical features, graphometrical features, geographical features, and other, for handwriting analysis. In this work, we are considering the combination of multilevel feature set for writer recognition and identification purpose. The proposed methodology contains regular image processing steps; such as handwritten image acquisition, pre-processing, feature extraction, and classification. The handwritten document samples are collected from 100 individuals and form a dataset for our work. The features obtained from feature extraction stage are given to random forest decision classifier to verify and identify the writer and obtained the 98.2% accuracy.

REFERENCES

1. M.Arif M, "A review on feature extraction and feature selection for handwritten character recognition", *International journal of advanced computer science and applications*, vol.6, no.2, 2015.
2. L. Heutte, J.V. Moreau, T. Paquet, Y. Lecourtier, C. Olivier. "Combining structural and statistical features for the recognition of handwritten characters."*13th International Conference on Pattern Recognition, IAPR-ICPR'96, Vienna, Austria, IEEE Proceedings, vol. 2,* pp. 210-214, 1996.
3. S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu and M. Kundu, "Combining multiple feature extraction techniques for handwritten Devnagari character recognition," *IEEE Region 10 Colloquium and 3rd International Conference on Industrial and Information Systems,* Dec. 2008.
4. Y. Kimura, A. Suzuki, K. Odaka, "Feature selection for character recognition using genetic algorithm," *IEEE Fourth International Conference on Innovative Computing, Information and Control (ICICIC), Kaohsiung* , pp. 401-404, Dec. 2009.
5. Zhang B., S. Srihari, and S. Lee, "Individuality of handwritten characters," in *International Conference on Document Analysis and Recognition, (Edinburgh, Scotland),* pp. 1086–1090, August 3-6 2003.
6. G. Leedham and S. Chachra, "Writer identification using innovative binarised features of handwritten numerals," in *International Conference on Document Analysis and Recognition,* 2003.
7. R. Li, H. Wang and K. Ji, "Feature Extraction and Identification of Handwritten Characters," *2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS)*, Tianjin, 2015, pp. 193-196.
8. R. K. Nath M. Rastogi "Improving various off-line techniques used for handwritten character recognition: a review*" IJCA vol. 49 no. 18* 2012.
9. A. R. Zarei R. Safabakhsh "A new approach for feature extraction with applications to automatic writer recognition" *ICCKE 2014 4th Internationale Conference* pp. 13-17 2014.
10. A.-L. Bianne-Bernard F. Menasri R.-H. Mohamad C. Mokbel C. Kermorvant L. Likforman-Sulem "Dynamic and contextual information in hmm modeling for handwritten word recognition" *TPAMI IEEE Transactions on vol. 33 no. 10* pp. 2066-2080 2011.
11. F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET),* Chennai, 2016, pp. 2264-2268.
12. Sumedha B. Hallale, Geeta D. Salunke, "Twelve directional feature extraction for handwritten English character recognition", *International journal of recent technology and engineering, vol.2, no.2,* 2013.
13. Khaled Mohammed, SitiZaitonMohd, azazkamilahMuda, "Feature extraction and selection for handwriting identification: A review", *Conference proceedings*, 2009.
14. Manimala, S & Gopal, Meghashree.G&Gokhale, Poornima&Chandrashekar, Sindhu.,"Automated Handwriting Analysis For Human Behavior Prediction", *ISRASE conference*, 2017.
15. Joshi, Prachi, Aayush Agarwal, AjinkyaDhavale, RajaniSuryavanshi and Shreya Kodolikar. "Handwriting Analysis for Detection of Personality Traits using Machine Learning Approach." *International journal of computer applications,* vol.130, no.15, 2015, pp.40-45.