Predicting the Big-Five Personality Traits from Handwriting

Nidhi Malik(✉) and Ashwin Balaji

Computer Science & Engineering, Amity School of Engineering and Technology, Noida, India
nmalik@amity.edu, bashwin52@gmail.com

Abstract. This work is about how the personality can be derived from the image of handwriting which can be derived among any of the five big personality traits accompanied by the list of features or reflecting the nature of the handwriting image from seven features. We would be using handwriting images which decide the nature of the person based on big-five trait. In the feature list dataset, every features are mined from images to match images with the image dataset and creating a separate clean feature list. Training the data using feature list and tag list efficiently by iterating every image and generating list of 12 elements for every image. Data list is classified using various algorithms like support vector machine, K-nearest neighbor, and random forest to evaluate accuracy by displaying the features of the input essay along with five big traits.

Keywords: Personality detection · Behavior analysis · Handwriting · Computer vision · Graphology · Natural language processing

1 Introduction

Personality is a vital human characteristic, and it also portrays the individuality. It is one of the basic aspects, by which we get to know various types of people and their selfness in a better way. It is considered to be one of the long-term goals and a difficult task for psychologists to evaluate human selfness and its effects on human nature. A person's reaction to a certain conditions plays a major role on how the person is depending on the situation. But, in most of the time, people react with respect to their personality or nature. It is possible to extract someone's personality traits by text samples to automatically identify personality and predict their reactions and behavior. Humans used to identify someone's nature by analyzing them within a period such that they become expert in predicting their nature or probable reactions by just analyzing their face. Researchers around the world are working on this domain especially computational linguistics such as machine learning, natural language processing predominantly in artificial intelligence.

In recent times, the interest of the scientists is leaning toward personality recognition which is expanding quickly. There are some applications that can make use of personality recognition such as social network, recommendation/review systems, deception detection, authorship attribution, sentiment analysis/opinion mining, among others.

M. K. Sharma et al. (eds.), *Innovations in Computational Intelligence and Computer Vision*,
Advances in Intelligent Systems and Computing 1189,
https://doi.org/10.1007/978-981-15-6067-5_25

Working in this field is beneficial for many activities that are performing by means of online facilities on a daily basis like customer care support, suggestions of services and products, etc.

2 Related Work

In [1], image files are explored to extract features of images to convert in a word embedding feature list. They have proposed a methodology to extract image features by using three different layers to namely base, intermediate, and final.

In base layer, various steps were involved in Normalization of image which takes place by applying various noise reduction methods like Boolean filter to remove textured background, ramp width reduction filter for sharpening and adaptive unsharp masking to maintain contrast. Contour smoothing to retrieve relevant data to extract handwriting features, local weighted averaging method is used for this purpose. Compression of image is performed using global thresholding to convert a color image to binary image. Next [1] performed row segmentation to extract row-wise vital data by applying vertical and horizontal projections in order to obtain maximum valid/important pixel count. After we would try to obtain spacing between lines feature and baseline feature by applying horizontal projection such that graphology dataset is used as a reference to bifurcate the handwriting features based on obtain data. Writing pressure is also obtained by analyzing pixel counts based on the thickness of pressure applied by applying horizontal projection. Next [1] performed word-level segmentation to obtain vital features; for this, vertical projection is used to determine height, width, and spaces between context of words. Also, [1] tried to obtain word slant feature by taking the reference of graphology dataset. In next step, letter-level segmentation is done in order to finally map the obtained values in the form of vector list in intermediary level. In intermediate level, obtained values were mapped (Handwriting map) to feature list in the form of vectors. In top layer, classification step is used by using neural network by assigning node weights to facilitate the process of classification based on the data and weight being fetched from previous layer. Finally, [1] presented the final analyzed data in the form of big-five personality trait. [2] Proposed a different approach to obtain handwriting features of human by applying steps Firstly, polygonization method in which a polygon is drawn around a line to retrieve baseline features and slope of the flow of handwriting. Then, threshold is performed to convert a grayscale image to binary image.

The technique of template matching is implemented by using graphology dataset for the purpose of matching the various types of alphabets writing techniques and finally, classification using artificial neural network (ANN) is done.

References [2–10] gave the insights of handwriting data retrieval by applying certain steps. First, they have performed preprocessing to remove noise and smoothing. Then, word-level segmentation, letter segmentation, and line segmentation are performed to input the word embedding in SVM (classifier). Some features of handwriting like, pen pressure, slant of words and letter, space between words, baseline, size of letters, and space between letters have been considered. Djamal and Darmawati [11] proposed a method to develop automated handwriting analysis system. Input image was segregated using RGB threshold and also cropping the image to retrieve handwriting feature based on user interest. They have eight handwriting characteristics to reveal the nature of the user. Characteristics used were writing pressure, baseline, writing slant, breaks while writing, word spacing, margins and analysis of speed of writing by observing stroke length, ink density, size of the letters.

3 Methodology

3.1 Big Five Model

Big Five Model is the most worked or analyzed metrics of personality domain in recent times. It is being evaluated by extracting and predicting certain patterns of texts repeatedly to arrive at a conclusion about a human. Patterns play an important role because one time or the other, person would definitely show his/her authentic nature which requires time and space. This model is being widely used personality trait structure. The human personality is computed as a list of five values with respect to bipolar traits. This is model is popular among the language and computer science researchers.

Personality is formally described in terms of the Big Five personality traits, in terms of (yes/no) values such as:

i. Extroversion (EXT): Is this person energetic, talkative, and outgoing or is he solitary and reserved.
ii. Neuroticism (NEU): Is this person anxious and sensitive or is he self-assured and secure?
iii. Agreeableness (AGR): Is this person straightforward, trustworthy, modest, and generous, or is he boastful, complicated, and unreliable?
iv. Conscientiousness (CON): Is this person efficient and organized or is he careless and sloppy?
v. Openness (OPN): Is this person creative and curious or is he dogmatic and cautious?

This work intends to automate the basic handwriting analysis tasks of graphology to determine a few important personality traits. Seven features/characteristics of handwriting are considered to be extracted from a sample handwriting image. Each of the seven resulting in unprocessed values will be matched into corresponding graphology rules of respective feature variations. A combination of these discrete values will be used to perform training each with classification algorithms (namely SVM, KNN, random forest) to determine personality traits based on values to predict the personality traits of the writer.

3.2 Dataset Available

IAM Handwriting Dataset

This dataset consists of images of scanned at a resolution of 300dpi with 256 gray levels in PNG format. It contains 657 writer's samples of their handwriting, scanned text of 1539 pages, 2685 labeled and isolated sentences, 13,353 labeled and isolated text lines and 115,320 labeled and isolated words. This dataset is used in this proposed methodology for the purpose of testing and training data.

4 Implementation Details

Our experiment, as discussed in the previous section, proposes a new method for predicting an individual's Big-5 personality traits based on handwriting. Our research is therefore based on two cognitive components: the Big five model, graphological review, and also considering image dataset. In the following sections, we have described the different phases of this novelty.

4.1 Flowchart

See Fig. 1.

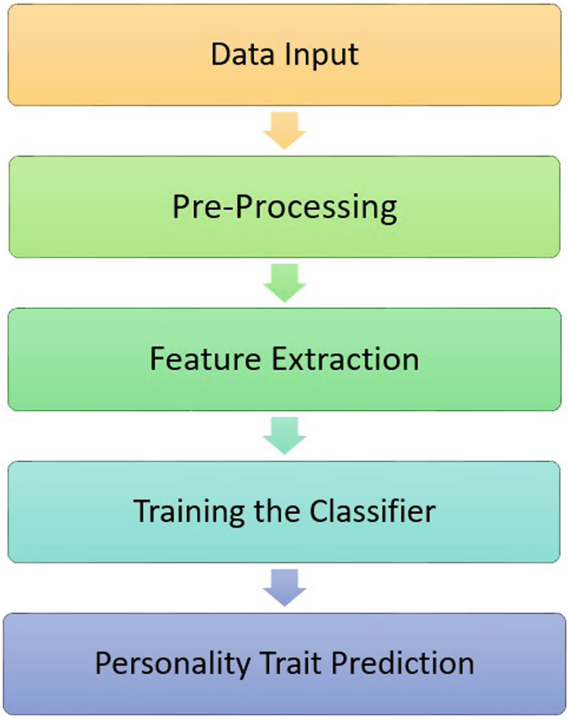


Fig. 1. Phases of the project

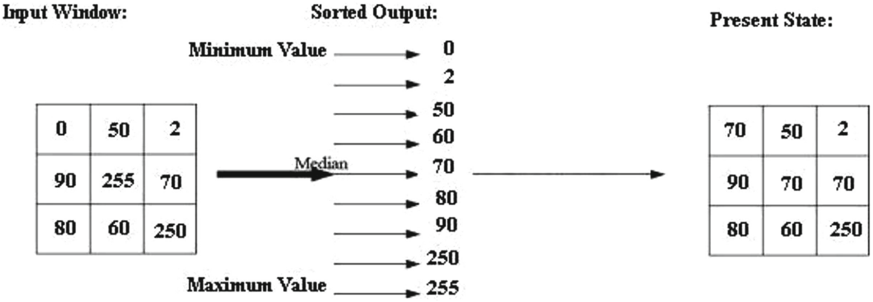
4.2 Preprocessing

During scanning process image could get distorted so to neglect the effect of distortion we need to do the preprocessing. The image is made suitable by applying various filtering methods, image thresholding, image transformations, and projections.

4.2.1 Noise Removal

In this, we would be using three different types of filters, namely median filter and bilateral filter.

Median Filter It is a digital/discrete and nonlinear filtering technique such that this filter chooses the average set of pixels or dense set of pixels to maintain integrity of image data (Fig. 2).



0	50	2
90	255	70
80	60	250

70	50	2
90	70	70
80	60	250

Fig. 2. Median filter [13]

Bilateral Filter It is an edge-preserving, noise-reducing, smoothing, and nonlinear filter. It adjusts the density of pixels by approximating the nearest pixel value (Fig. 3).

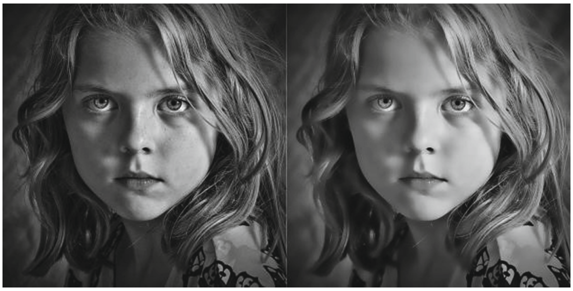


Fig. 3. Bilateral filter [13]

4.2.2 Grayscale and Binarization (Inverse Binary Thresholding)

In this, pixel density values are set to 0 below a threshold value and 1 otherwise. Output of this thresholding actually reverses the appearance of source image (Fig. 4).

4.2.3 Contour and Warp Affine Transformations

In this, image transformations are performed over a pixel matrix. Operations include rotations, translations, and scaling of the image (Fig. 5).

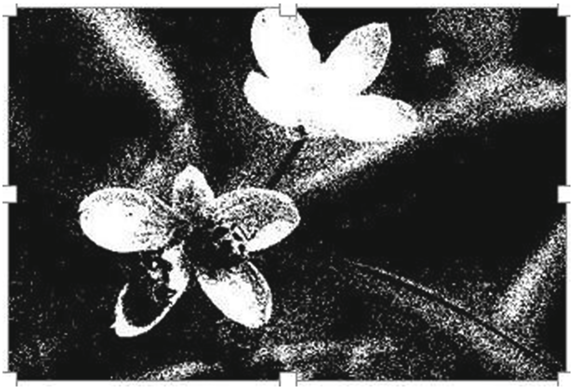


Fig. 4. Inverse binary thresholding

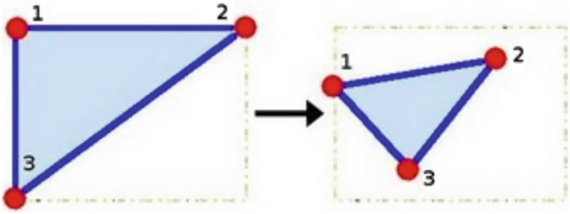


Fig. 5. Different phases [15]

4.2.4 Projection

Horizontal and vertical projections techniques are used to count pixels based on the alignment of text pattern. For example, baseline, line spacing, top margin are needed to be horizontally projected because we may get the maximum contour in the form of pixel counts. Similarly, word spacing, letter spacing, letter size, word size, etc. needed to be vertically projected to obtain maximum relevant data.

4.3 Feature Extraction

In this phase, features of the preprocessed image are being discovered and stored in the list format.

Seven features based on which the extraction would take place are:

[Baseline, Top Margin, Letter Size, Line Spacing, Word Spacing, Pen Pressure, Slant Angle].

Now, these seven features are obtained in this phase accompanied with IAM dataset to generate image values and stored in list in runtime (as shown in Table 1) by extract routine module. In the below tables, sample outputs are evaluated for first ten images.

The file obtained is unfavorable because it may contain negative as well as positive values with no value range constraints, which makes it difficult to use for classification purposes. So these lists of data are re-organized and approximated to obtain a clean

Table 1. Unprocessed feature list generated in runtime

Image set	Baseline	Top margin	Letter size	Line spacing	Word spacing	Pen pressure	Slant angle
000-1.png	−0.39	5.7	12.29	4.37	2.46	171.38	−15.0
000-2.png	0.05	3.07	17.29	2.89	1.74	194.71	−15.0
000-3.png	−1.1	2.02	16.36	3.14	1.65	170.29	−15.0
000-4.png	−0.01	1.91	15.73	3.32	1.6	165.56	180.0
000-5.png	−0.0	2.1	15.7	3.25	1.69	171.25	−15.0
000-6.png	−0.27	2.14	16.33	3.16	1.54	174.16	−15.0
000-7.png	0.29	2.62	14.1	3.82	1.93	170.93	−15.0
000-8.png	−0.06	2.8	14.3	3.58	1.92	171.9	−15.0
000-9.png	0.11	2.43	14.0	3.78	1.78	171.28	180.0
000-10.png	−0.05	1.36	16.22	3.16	1.42	164.27	180.0

feature list which is generated in runtime (as shown in Table 2) using feature routine module which imports graphology rule module to categorize the seven features supported by unprocessed feature list file.

Table 2. Feature list generated by “graphology rules” module

Image set	Baseline	Top margin	Letter size	Line spacing	Word spacing	Pen pressure	Slant angle
000-1.png	1	0	1	0	0	2	1
000-2.png	2	0	2	2	2	0	1
000-3.png	1	0	2	2	2	2	1
000-4.png	2	0	2	2	2	2	6
000-5.png	2	0	2	2	2	2	1
000-6.png	2	0	2	2	2	2	1
000-7.png	0	0	2	0	2	2	1
000-8.png	2	0	2	0	2	2	1
000-9.png	2	0	2	0	2	2	6
000-10.png	2	1	2	2	2	2	6

Training Data

First it will fetch data from a file (feature list file generated by graphology rule module) as shown in Table 2. Then it will analyze words from fetched data, and it will map the words with output of feature routine module and based on the mapping it will assign the value to seven features iteratively to produce a file tag list file containing every image features iterated till the end of file.

Likewise, tag list file is also generated which would be taken into account during classification.

In the next step, we would try to obtain list of 12 vectors as shown in Table 3.

[Baseline, Top Margin, Letter Size, Line Spacing, Word Spacing, Pen Pressure, Slant Angle, Neuroticism, Agreeableness, Openness, Conscientiousness, Extroversion] is generated in tag list file.

This step is important because the classification is dependent upon these generated files as if now we have obtained features and tags to implement supervised learning.

4.4 Classification

In classification, first we will split the vectors which were obtained in the previous phase where there was a list of 12 elements vector length.

The first seven are features, and rest of the five values are personality big traits.

```
Variable_1=Classification (function)
Variable_2=Variable_1.fit (train_data, train_tags)
Variable_3=Variable_2.predict (train_data)
ShowOutput (train_tags, Variable_3, "Statement")
```

4.4.1 Random Forest

Random forest is a group of machine learning technique in which collection of decision trees (i.e., forest) or the multitude of decision trees are taken during the model training and the output is classified using either classification or regression techniques.

Sqrt and Log2 were the two different approaches in random forest.

Accuracy of random forest with Neuroticism, Extroversion, Openness, Agreeableness, and Conscientiousness was observed to 86.304, 85.869, 80.652, 85.217, and 88.043%.

4.4.2 K-Nearest Neighbor

K-nearest neighbor is an instance-based machine learning technique used widely for pattern recognition or analysis in which the algorithm tries to find the nearest instances for a particular data. Nearest neighbors can be varied by adjusting "K" value to maintain a set of close instances for classification and regression. It has four different algorithms, namely Ball tree, KD trees, Brute force and Auto.

Accuracy of K-nearest neighbor with Neuroticism, Extroversion, Openness, Agreeableness, and Conscientiousness was observed to 84.782, 81.304, 79.782, 86.739, and 88.260%.

4.4.3 Linear Support Vector Machine

For regression and classification of data points, support vector machines are widely used. It is a supervised learning algorithm. Hyperplanes or divide column is drawn over the

Table 3. Tag list of 12 features

Image set	Baseline	ToP Margin	Letter size	Line spacing	Word spacing	Pen pressure	Slant angle	Neuroticism	Agreeableness	Openness	Conscientiousness	Extroversion
000-1.png	1.0	0.0	1.0	0.0	0.0	2.0	1.0	1	1	0	1	0
000-2.png	2.0	0.0	2.0	2.0	2.0	0.0	1.0	1	1	1	1	1
000-3.png	1.0	0.0	2.0	2.0	2.0	1.0	1.0	1	1	1	1	1
000-4.png	2.0	0.0	2.0	2.0	2.0	2.0	6.0	1	1	1	1	1
000-5.png	1.0	2.0	0.0	2.0	2.0	2.0	2.0	1	0	0	1	1
000-6.png	1.0	2.0	0.0	2.0	2.0	2.0	2.0	0	0	1	0	0
000-7.png	2.0	0.0	0.0	2.0	0.0	2.0	2.0	0	1	0	1	0
000-8.png	1.0	2.0	0.0	2.0	0.0	2.0	2.0	1	1	0	1	0
000-9.png	2.0	0.0	2.0	0.0	2.0	2.0	6.0	1	1	0	1	1
000-01.png	2.0	1.0	2.0	2.0	2.0	2.0	6.0	1	0	0	1	1

dimension to classify data points efficiently with a complexity of $O(n)$. It classifies data points by categorizing instances making it to be a nonprobabilistic linear classifier using binary values. It is a linear as well as nonlinear classification which can be adjusted using Kernel methods of SVM. In our project, Radial Basis Function (RBF) kernel is used because this kernel is used when there is nonlinear data, provided that we need data in linear format.

Accuracy of support vector machine with Neuroticism, Extroversion, Openness, Agreeableness, and Conscientiousness was observed to 87.391, 88.043, 78.260, 87.391, and 88.913%.

5 Experimental Results and Discussion

We have collected graphology information from the nearest graphology department to ensure correctness of the novelty. We have conducted questionnaire and interviews to collect as much as information possible regarding the handwriting characteristics or patterns produced by various individuals. Apart from graphology information, we have used IAM Dataset containing 600+ handwriting images for various individuals. Also, the experiment was purely conducted on inter-subject dataset rather than considering intra-subject dataset [1], to ensure scalability of this novelty. Based on [1], computation overhead is reduced considerably because [1] divided this problem into smaller chunks, i.e., layers to analyze handwriting. Even though [1, 12] used neural network approach but here we have performed this based on the experimental working of [11] considering SVM, KNN and included random forest algorithm as this has been never used before for experimental or prediction purposes.

Training has been done using IAM Dataset accompanied by graphology rules to evaluate various features of handwriting. Testing phase included handwriting samples of (in paragraphs or document) various individuals collected and uploaded the image in JPG format to test whether the performance is optimal. Surprisingly, out of 20 samples predictions, 14 cases were right or partially correct. Finally, we've concluded that false positive rate is approximately 32% which is better compared to the prior art.

It was observed that overall 85.12% accuracy was achieved considering inter-subject data. Moreover, as per the prediction experiment of the prior art, accuracy of [1] 80.5% in inter-subject dataset, [13] accuracy was 78.9% and accuracy of [12, 14] were 72.8% and 68%, respectively (Figs. 6 and 7).

6 Conclusion

Everything in this world has some positive as well as negative attributes, how to overcome those negative attributes is the real challenge. This project successfully predicts the nature of human being by using their handwriting images. Any set of images of the users can be employed to predict nature based on big-five personality. It is extremely helpful in future as we may need program instructions to decide human's nature especially in the field of psychology, recruitment process, etc. There may be less chances that the human may fake the document because the mind of human being is always trained to reveal patterns in anyway, and one such thing is handwriting. This project may reveal the authentic nature

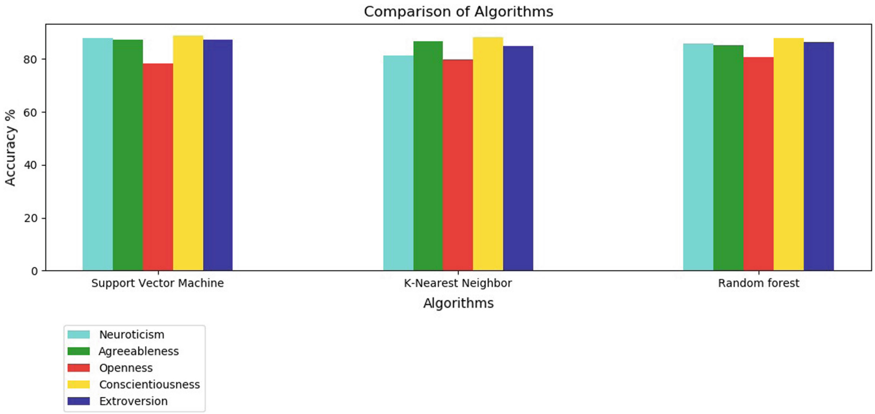


Fig. 6. Experimental performance of algorithms on Big-5 traits

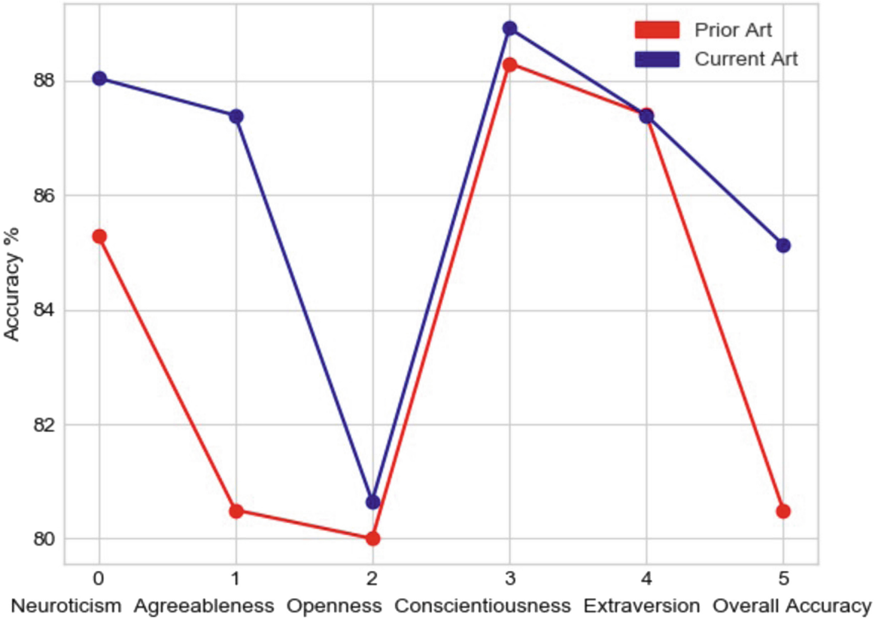


Fig. 7. Comparison graph

of human beings in most cases. As far as the project is concerned, seven handwriting features are extracted from handwriting sample and five personality traits are predicted. For each of the personality trait, separate SVM classifier is trained. We have also trained random forest and KNN classifiers. After sufficient and satisfactory training, we would be able to predict personality traits on new handwriting image samples with greater accuracy and efficiency.

7 Future Work

There are rooms for future enhancement in this work. Some of the enhancements that can be made are listed as follows.

I. This experiment is working on supervised learning, it would be appreciated by the authors if this experiment is done using unsupervised learning by means of clustering segmentation techniques or other as it is challenging to apply unsupervised learning in an ambiguity domain like NLP.
II. This handwriting recognition should also be extended to other languages to increase the scope of this experiment by decreasing computational overhead.
III. This invention is purely experimental, and forward citations are greatly appreciated for upcoming research works and may contain some computational errors because graphology dataset may vary considerably because graphology dataset is not available in Internet. Generalized dataset for future computation is widely accepted and appreciated.
IV. More image processing techniques can be applied because image processing itself is a wide domain to enhance this experiment like by using watershed segmentation or DCT or by phase shifting of rasters, image tiling techniques for computation. Moreover, PDE-based experiment is also appreciated as it is comparatively faster than other image processing techniques.

References

1. M. Gavrilescu, N. Vizireanu, Predicting the big five personality traits from handwriting. EURASIP J. Image Video Process.
2. V.P. Dhaka, Offline language-free writer identification based on speeded-up robust features international. J. Eng. (IJE) **28**(7), 984–994 (2015)
3. The Seventh International Conference on Advances in Computing, Electronics and Communication—ACEC 2018
4. S. Prasad, V.K. Singh, A. Sapre, Handwriting analysis based on segmentation method for prediction of human personality using support vector machine. Int. J. Comput. Appl. (0975 8887) **8**(12) (2010)
5. V.P. Dhaka, Offline scripting-free author identification based on speeded-up robust features. IJDAR **18**, 303–316 (2015). https://doi.org/10.1007/s10032-015-0252-0
6. V.S. Dhaka, Segmentation of handwritten words using structured support vector machine. Pattern Anal. Appl. (2019). https://doi.org/10.1007/s10044-019-00843-x
7. V. Chanderiya, Writer identification using graphemes. Sādhanā **45**, 42 (2020). https://doi.org/10.1007/s12046-020-1276-9
8. V. Kamath, N. Ramaswamy, P. Navin Karanth, V. Desai, S.M. Kulkarni, Development of an automated handwriting analysis system. ARPN J. Eng. Appl. Sci. **6**(9) (2011)
9. H.N. Champa, AnandaKumar K.R., Artificial neural network for human behaviour prediction through handwriting analysis. Int. J. Comput. Appl. (0975–8887) **2**(2) (2010)
10. D. Prashar, P.K. Grewal, Behaviour prediction through handwriting analysis. IJCST **3**(2) (2012)

11. E.C. Djamal, R. Darmawati, Application image processing to predict personality based on structure of handwriting and signature, in *2013 International Conference on Computer, Control, Informatics and Its Applications*
12. Z. Chen, T. Lin, Automatic personality identification using writing behaviors: an exploratory study. Behav. Inform. Technol. **36**(8), 839–845 (2017)
13. Image compression and image processing.http://cs248.stanford.edu/winter19/lecture/imageprocessing/slide_066
14. P.S. Kedar, M.V. Nair, M.S. Kulkarni, Personality identification through handwriting analysis: a review. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(1), 548–556 (2015)
15. Improving the Effectiveness of the Median Filter. https://www.researchgate.net/figure/Impulses-switched-withthe-effective-median_fig5_280925268
16. L. Nader, A. Mohamed, M. Nazir, M. Awadalla, Identification of writer's gender using handwriting analysis. Int. J. Sci. Res. Publi. **8**(10) (2018)
17. S.K. Singh, Hemlata, M. Sachan, Personality detection using handwriting analysis: review
18. B. Fallah, H. Khotanlou, in artificial intelligence and robotics (IRANOPEN), in *Identify Human Personality Parameters Based on Handwriting Using Neural Networks (April 2016)*
19. M.S. Hemlata, S.K. Singh, Personality identification based on handwriting. Int. J. Emerg. Technol. Comput. Appl. Sci. (IJETCAS) **12**(3), 231–235 (2015)