Proceedings of the 12th INDIACom; INDIACom-2018; IEEE Conference ID: 42835
2018 5th International Conference on "Computing for Sustainable Global Development", 14th - 16th March, 2018
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Gender Detection using Handwritten Signatures

J Mohit Reddy
National Institute of Technology Karnataka,
Surathkal, India
Email Id: mohitreddy1996@gmail.com

T Guru Pradeep Reddy
National Institute of Technology Karnataka
Surathkal, India
Email Id: gurupradeept@gmail.com

Samarth Mishra
National Institute of Technology Karnataka,
Surathkal, India
Email Id: samarthmishra95@gmail.com

Manjunath Mulimani
National Institute of Technology Karnataka,
Surathkal, India
Email Id: manjunath.gec@gmail.com

Shashidhar G. Koolagudi
National Institute of Technology Karnataka,
Surathkal, India
Email Id: koolagudi@nitk.edu.in

Abstract – In this paper, a method is proposed which uses both Image Processing and Machine Learning techniques which detects the gender of a person using handwritten signature. A photograph of a handwritten signature is given as input to the model which then extracts different features like pen pressure, slant angle, count external and internal contours etc. The features extracted from multiple images in the dataset are used to train the model, which then predicts the output of a new input given to it. Our objective is to collect unbiased datasets from a set of people and feed those signatures to the model, carrying out the statistical analysis and calculating the accuracy of the algorithm after every signature classification. We have used Adaboost classifier which gave a cross-validation accuracy of 73.2% compared to other classifiers like Gradient Boosting Classifier, Random Forest Trees and Multi-Layer Perceptron which gave 73.2%, 63.2% and 59.6% accuracies respectively.

Keywords – Classifiers, Gender Prediction, Handwritten Signatures, Image Processing, Machine Learning.

I. INTRODUCTION

It might appear a very mundane and trivial task of signing your name but it comprises of a very complex neural process taking place in our mind. The mental and physical coordination is carried out by the various areas in the brain like motor, cognitive and emotion areas and the signals are passed down to the hand. We call this Basic Neurological Pattern (BNP). Our assumption in this paper is that the BNP of an individual remains constant when he/she is writing or signing. Every human has a unique neuromuscular movement and those are closely related to the personality traits of that individual. Some of the features of the handwriting are closely linked to gender. The work could be used in the forensic or legal domain to focus on certain category of suspects presenting improved results in the verification of the writer.

The detection of gender through signature analysis could be found in very few literature. Few of the papers give the hypothesis of how a unique personality trait is identified through one's handwriting. In [1], macro features that classify the handwriting into demographic categories. Those features mainly concentrated on the stroke of writing, ratio of sizes of letters, pen pressure, direction and slant with the prediction of age and gender recognition accuracies of 86.6% and 72.5%. Although there were a few shortcomings of this paper is the same piece of the sample had to be collected and analyzed for all individuals but that is never the case in real forensics.

Srihari, Sargur N, et al. [2] carried out studies to analyze the handwriting of an individual for over 1500 data sets. The handwriting samples collected had a varied proportion of people from different backgrounds, age groups and ethnicities. Three handwriting samples were given by each individual and features were extracted from global level, paragraph level and word level. Further, some individual letters were used to extract some finer features. The machine learning approach was used to characterize each individual. The devised algorithm took few samples to learn the writer's handwriting features and others were used to test the learned models. Based on the macro and micro features that were extracted an individual was identified with an appreciable confidence. The setback was that it was not possible to establish the results on the basis of a single stroke of writing.

In our paper, we propose a methodology of detecting gender. We take into account certain features like ratio of black pixels, external and internal contours, ratio of the

 ISSN 0973-7529 ISBN 978-93-80544-28-1

minimum to the maximum size of a character in a signature, contour slope, slant angle, maximum length of strike in 8 directions and neighboring pixel correlation. The extraction gives out certain value and is stored for every signature. Classifiers are used for classification of the signatures using the training data set. Statistical analysis is carried out to find the closeness of the algorithm.

The major advantages of this paper are:

- The proposed model is easily deployable and needs no external assistance other than installation of prerequisites.
- The proposed model is inexpensive. Images to be given to the model as an input can be taken from any source. No special devices are required.
- The proposed model accuracy is nearly 73.2% which is better than a chance i.e 50%. Hence can be for forensics or other applications.

The paper has been divided into 4 sections. Section I is a brief introduction and insight about the work, Section II explains the feature extraction and importance of each feature in classification of gender, Section III is about the classifiers used on the training dataset and Section IV is finally summed up as results and discussions which also explains about the various classifiers used.

II. PROPOSED METHODOLOGY

We start with data collection where the data is collected from various sources. It is essential to preprocess and normalize the collected data. Processed data is used for feature extraction and the features extracted are stored in a Comma Separated Value(CSV) file along with the respective writer's gender. The dataset is used to train the Machine Learning models and classifiers. 4-fold cross validation algorithm is incorporated in the calculation of accuracy.

A. Data Collection

Signatures were collected equally in number from male and female individuals to avoid skewed dataset. 90 Male signatures and 82 Female signatures were collected, forming a data set of 172 handwritten signatures. Images for all the signatures were taken from a 16-megapixel camera in the same environment. For instance, randomly selected male and female signatures are shown in Fig.1.

(a) (b)

Fig. 1 : Signatures (a) Male signature (b) Female signature

B. Preprocessing

- Draw Contours of the given image.
- Coordinates of various points on the contours are known.
- Consider the minimum x-coordinate and minimum y-coordinate of the contours which act as the top-left corner of the bounding box.
- Consider the maximum x-coordinate and maximum y-coordinate of the contours which act as the bottom-right corner of the bounding box.
- Given top-left and bottom-right points, a bounding box can be made surrounding the signature (foreground) in the image (see Fig.2). The image in the bounding box can be cropped out.

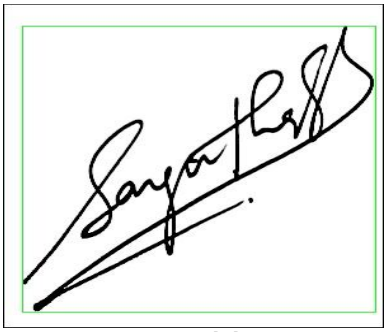


Fig. 2 : Bounded image

In grayscale every pixel will have discrete values of intensities in the range 0 to 255 inclusive. Binarization technique is applied to the grayscale image so that the feature extraction algorithms have to deal with only 2 intensities i.e either 0 or 255. The threshold value is calculated as follows:

- Plot a Graph representing Number of Pixels vs Intensity as shown in Fig.3.
- Start from intensity zero. Increment the count with the number of pixels at each intensity.
- Whenever the count becomes equal to 75% of the total number of pixels, store the intensity value.
- Intensity value at the 75% mark is the threshold value.

Because the signatures are taken on a plane sheet, it can be assumed that the intensity will be less only for the signature. White pixels are the one whose pixel intensity exceeds the threshold value and their intensity is set to 255 whereas the other pixels are categorized as black and their intensity is set to 0.

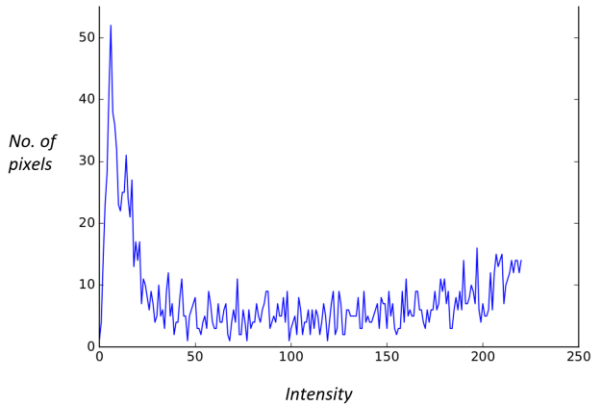


Fig. 3: Intensity vs Number of pixels histogram

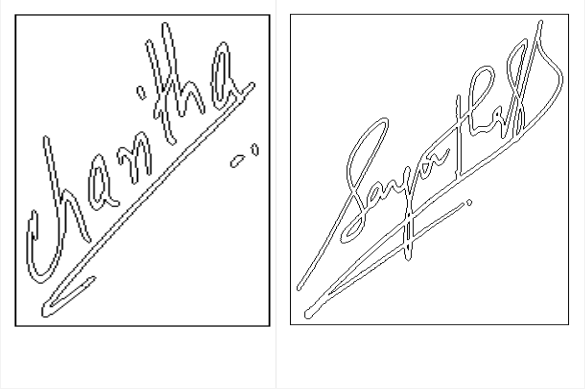
C. Feature Extraction

1. *Ratio of black pixels:* This feature is mostly used to indicate how much relative pen pressure is associated with the signature sample. It is calculated by dividing the number of foreground and background pixels in the processed image as given in (1). We can derive an idea about the stroke thickness, the size of the handwriting and pen pressure from it.

$$Ratio = \frac{Number\ of\ foreground\ (black)\ pixels}{Number\ of\ background\ (white)\ pixels} \quad (1)$$

The intensity v/s number of pixels of an image in the dataset is shown in Fig.4. The intensity closer to 0 denotes that the pixel is part of foreground i.e signature, and intensity closer to 255 denotes the pixel is part of the background. From the graph, we can infer about the dark pixels which are indicative of pressure applied by the person and the size of the handwriting. High spikes at the beginning of the histogram imply that there are more number of black pixels and vice-versa.

2. *External contours:* The number of contours is indicative of writing movement. The number of external and internal contours are extracted from the contour hierarchy. External contour is the one which does not have any parent. They can be used to measure writing movement. For example, the signatures of the people who write in a disconnected manner will have more number of external contours. Example of this contour connectivity feature is shown in Fig. 4a. As we can see the number of external contours for the image are more compared to internal contours indicating that the signature is more disconnected.
3. *Internal contours:* Internal contours are the one which has the parent in hierarchy representation of the contours. They will be nested in some external contours. Even they can be used to estimate the writing style of the person. For example, a person who writes in a curvy and connected manner will have a higher degree of internal contours. Example of this contour connectivity feature is shown in Fig. 4b. As we can see here the number of internal contours are 14 and the no of external contours are 3 which clearly suggests that the signature is more connected or curvy in nature.



(a) (b)

Fig. 4: Contours (a) Number of external contours = 8, Number of internal contours = 2; (b) Number of external contours = 3, Number of internal contours = 14.

4. *Maximum to minimum size ratio:* Size of characters is one of the important characteristic of an individual's signature. But the size of the signature may vary from person to person, so if we take the actual size of each contour we may misinterpret this information. To avoid this we use ratio of height and width of the contours. To measure the size of a signature, we take height and width of contours into consideration. As a feature, we calculate and store the ratio of minimum and maximum height and ratio of minimum and maximum width among all contours of the signature.
5. *Contour slope:* Slope is calculated from each contour using its height and the width. The global contour slope is the average of all the slopes of each contour. This is the indicative of the nature of the stroke formation. To get the entire picture of stroke

formation of the signature, we are considering the average of slopes of all contours as given in (2).

$$Average\ of\ slopes\ of\ all\ contours = \sum s_i / N \quad (2)$$

where S_i denotes the slope of i^{th} contour, calculated using (3) and N is the number of contours.

$$Si = \frac{Height\ of\ the\ ith\ contour}{Width\ of\ the\ ith\ contour} \quad (3)$$

6. *Slant angle:* Slant angle is used to know the overall inclination of the signature. To calculate this we find the lowest left point where the signature starts P1(x1, y1) and rightmost lowest point P2(x2, y2). Angle made the line joining P1, P2 and the line y = y1 will be the slant angle. The line joining the both will give us the slope of inclination of the signature. Example of the slant angle is given in Fig.5.

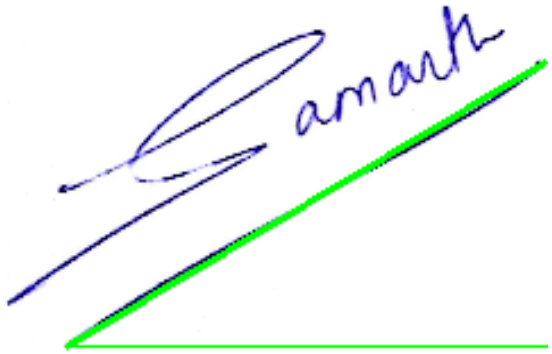


Fig. 5: Slant angle of signature

7. *Maximum length of strike in 8 directions:* This can help us distinguish between fast writers who tend to produce smooth handwriting and the slow writers who tend to produce twisted handwriting. Here we calculate the maximum length of the line segment in each of the eight directions i.e right, top-right, top, top-left, left, bottom-left, bottom, bottom-right. For smooth handwriting usually, a number of twists or turns in their handwriting will be less so they tend to produce long line segments in one direction. For slow writers, they tend to change the direction very frequently so the length of line segments in any direction will be less. To estimate this, for each pixel, we determine the longest line segment passing through that pixel. The line segment can be in any one of the eight directions. After performing the above procedure for all pixels, we will take the maximum length line segment in each direction. To eliminate the effect of size of the signature (normalization), we divide the maximum length line segment in each direction by maximum among all these line segments.

For any pixel of the image, we consider 8 directions or neighbors as shown in Fig.6. If any of the neighboring pixel is part of foreground, we increment the segment length in that direction with respect to pixel Pi by one. Now the foreground pixel is considered as center and the maximum length segment in that direction is calculated. It works similar to depth first search and by the end of it gives us the maximum length line segment in all eight directions which passes through some point on the signature. But for a pixel, the line segment in the directions top-left and bottom-right must be added to get the line segment length. Similarly for others directions, lengths must be added accordingly.

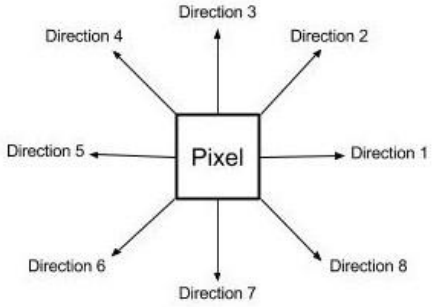


Fig. 6: Directions with respect to center pixel

8. *Neighboring pixel correlation:* This is edge based directional feature. This is used to find the distribution of the pixels in all directions which gives us the estimate about the direction in which person signs. This can also be applied at several sizes by posting the window at the center of the pixel. Here we considered 8 directions as shown in fig 6. To estimate this for we calculate the number of pixels in each direction by fixing every pixel at the center of the window, then we vary the sizes of the window.

For window size 1 we just need to consider eight pixels in all the eight directions. But as we increase the size of the window we need to consider more pixels. So we need to map every pixel in the window to some direction (among the eight). To do this we may need to map a pixel to more than one direction. If a pixel is along only one direction we map that pixel to that direction. But if the pixel is in between two different directions, we map the pixel to both the directions. So if there is a foreground pixel in

that direction we increment to the pixel count for both directions.

The same method can be extended to map all the pixels in the window of size 3 to 8 eight directions and using those we can calculate the no of pixels in each direction.

III. CLASSIFIERS

Different classifiers used in the proposed methodology are discussed below in brief.

A. Random Forest Classifier

Random forest classifier is one of the most widely used classifier because of its ability to fit a number of decision trees on various sub-samples of the dataset. The algorithm uses averaging to improve the accuracy and decrease the tendency of overfitting. In the model, 150 estimators are used for training. Estimators are the number of trees in the forest. In the model, we consider 7 features when looking for the best split. The nodes will be expanded until all leaves are pure or all leaves contain less than 2 samples. Here 2 is the minimum samples split which is a minimum number of samples required to split any internal node.

The rank of the output of a decision tree is given by comparing it to the already examined correct output from the training data. The features are then ranked accordingly. This is how the Random Forest averages decision tree output. The features in the tree will be given more importance and thus some of the decision trees will outperform. When the tree is injected with higher dimensional data the tree might be split by less relevant features and Random Forest method might fail.

B. Gradient Boosting Classifier

Gradient Boosting algorithm is a model which follows the level wise fitting method. Every stage a decision tree like structure is fit to the existing model. First, the base model is built and then fitted into the original target. Gradient Boosting classifier is robust to overfitting. Here we consider 250 boosting stages. As it is robust, higher values tend to give better accuracies. Maximum depth of the individual estimator is set to default value of 3. Decision trees are quite flexible and they can handle mixed data types. Loss functions can be easily implemented and residual gradient can be computed easily. We try to minimize loss functions.

C. Adaptive Boosting

AdaBoost (Adaptive Boosting) uses decision trees and iteratively improves the building of the learning of the algorithm by taking into account the incorrectly classified signatures in the given training datasets. The maximum number of estimators at which boosting is terminated is set to 200. In case of perfect fit, the learning procedure is stopped early.

Adaptive Boosting acts a strong classifier over a weak classifier. Adjusts the values to learn from the failures during the training phase of the model. The algorithm begins with assigning equal weights to all the signatures and some kind of decision tree is selected. Then the weights of the incorrectly classified signatures are increased. It is iterated n times, every time the weights on the training dataset is updated. The final result is the weighted sum of the n learning signatures. Any classifier can be used as base estimator model on which the boosted ensemble model is built. We used Gradient Boosting classifier as is not prone to overfitting. The parameters which were used to tune stand-alone Gradient Boosting classifier were used here.

D. Multilayer Perceptron

It is a kind of an algorithm which learns on its own. It uses a training dataset and uses a set of features. There are continual multilevel nodes in a Directed Acyclic Graph (DAG), where each level is completely connected to the adjacent one. This model maps the input data to the set of the output data obtained [3]. There might be more than one non-linear layers they and they are called hidden layers. For our purpose, we used network with three hidden layers with 22, 11 and 5 neurons respectively. As they are fully connected the no of parameters to be learned increases drastically with increase in no of neurons in each layer. So, in order to learn all these parameters, we need a large training data. Considering the limited amount of training data, only three hidden layers are used.

IV. RESULTS AND DISCUSSION

For testing k-fold cross validation algorithm is used. A sample of data is collected and partitioned, this is called the training set, which is used to carry out the validation and analysis on the testing data set. The test is carried out in multiple rounds of cross-validation using different partitions, the result thus elicited and averaged over the different rounds. The split of the training set and Validation set depends upon the K value chosen. We used K = 4, in this case the training set will contain the data and validation will have the data. Total of four rounds will have happen, each time the split will be different and the final result will be average of this four rounds. Each round, 75% of the dataset is used as a training dataset and remaining 25% is used as the test data set. Similar process is carried out 4 times with different (random) splits. As the no of features and no of training samples are less cross-validation does not add much computational overhead.

Table I shows the accuracy values of different Classifiers. Confusion Matrix of different classifiers is shown

in Table II. All the models we have used other than MLP model are ensemble-based models. These models use techniques like bagging(subsampling and training data) or boosting(focussing on misclassified training instances and relearning with different weights). As they tend to take the weighted average of multiple models, even if one of the models is not up to the mark it won't affect the overall performance of the model much. And they also tend to reduce and very unlikely to overfit. So we preferred them over single independent models. It has been observed that Adaboost Classifier with gradient boosting gives better accuracy compared to others. The reason for this it learns from several weak classifiers and it will overcome fitting. It combines the advantages of gradient boosting (which is giving the second highest accuracy) too as it is the base estimator model. Due to limited amount of data and large number of trainable parameters, MLP (Multilayer Perceptron) classifier is not the suitable learning model for this data set.

	Random Forest		Gradient Boosting		Adaboost		Multilayer Perceptron	
	F	M	F	M	F	M	F	M
F	20	4	20	7	19	4	17	7
M	8	11	7	9	7	13	9	10

M : Male F : Female

TABLE II : Confusion Matrix of different Classifiers

V. CONCLUSION

This paper is an attempt to identify the gender using the signatures of an individual. The accuracy achieved through our method is 73.2%. Implementation of the proposed methodology is hosted on github [4]. However, graphologists who are good at hypothesis and theory, suggest that the prediction of gender with close to 100% accuracy is not possible. They prefer to categorize signatures/handwriting on the basis of personality not gender.

Therefore, this can be assumed as an indicative effort. There is room for more concepts and well-defined hypothesis to extend the work. Some of the directions that can be explored is by increasing the size of the data set (of the order of 10000 signatures), the data collected could be varied like different demographics or areas, more prominent features to be embedded, Deep Neural Networks like CNN can be used for images. Although this work highlights the classification on the basis of gender another work can be taken up on the same lines to identify forged signatures.

Classifier	**Accuracy (in %)**
Random Forests	63.21
Gradient Boosting	73.24
Adaboost	73.28
Multilayer Perception	59.61

TABLE I : Percentage Accuracy of Classifiers

REFERENCES.

[1] Al Maadeed, Somaya and Abdelaali Hassaine, "Automatic prediction of age, gender and nationality in offline handwriting". *EURASIP Journal on Image and Video Processing February 2014*.

[2] Srihari, Sargur N., et al. "Individuality of Handwriting." *Journal of forensic science* 47.4 (2002).

[3] Champa H N, K R Ananda Kumar. "Artificial Neural Network for human behavior prediction through handwriting analysis." *International Journal of Computer Applications, Vol. 2, May 2010*.

[4] Github link: https://github.com/mohitreddy1996/Gender-Detec tion-from-Signature.

