sciendo

A Study on Various Techniques Involved in Gender Prediction System: A Comprehensive Review

Payal Maken, Abhishek Gupta, Manoj Kumar Gupta

School of Computer Science & Engineering, Shri Mata Vaishno Devi University, Kakryal, Katra, Jammu & Kashmir – 182 320, India
E-mails: 17mms007@smvdu.ac.in   abhishekgupta10@yahoo.co.in   manoj.cst@gmail.com

Abstract: *Predicting gender on the foundation of handwriting investigation is a very invoking research area. Handwriting analysis also has numerous applications. It is useful for forensic experts to investigate classes of writers. This prediction is executed fundamentally by two steps: feature extraction and classification.* **As a whole, prediction solely depends upon the feature extraction.** **Distinct and most varied features make the classification accurate.** *This paper describes the problem comprehensively along with its foundation. It presents a survey of the available methods to solve the discussed problem. A comparative analysis of the discussed methods of feature extraction and classification along with the available databases is also presented in this paper.*

Keywords: *Handwriting recognition, gender, extraction, features, forensics.*

1. Introduction

Writing is a very common mode of communication between the individuals. Handwriting has a significant use from the past decade, as it is associated to an individual. HandWriting Recognition (HWR) has become now one of the important and interesting fields and has received a substantial attention from researchers of all over the world. Handwriting recognition has been mostly used in the field of signature verification [1] or identification in forensic investigations. Character recognition of handwritten text or printed document text is the ancient dreams. A basic method for character recognition is Optical Character Recognition (OCR) which is very common in current decade. D a f e and C h a v h a n [2] had given the details on optical character recognition method which includes optical scanning and the working of OCR. Character recognition is not only limited to recognizing characters of one specific language. Multiple languages recognition can be done by the algorithms designed for such problems. One approach was proposed by K a u r [3] demonstrating hierarchical centroid method and Support Vector Machine (SVM) for the recognition of Gurmukhi printed characters. Character recognition is branched into two headings distinctively, i.e., online character recognition system and offline character

recognition system [4]. Offline character recognition HWR system is further grouped as printed and HWR system [5].

In online system, the input is sensed through the tips of the pen based monitors whereas in offline system, the input is taken from a dataset, that dataset has acquired input images using optical scanning (optical character reader). Based on the literature [6] it was observed that a recognition system shall be divided into four blocks named as:

a) Input image/ Image acquisition,
b) Image preprocessing,
c) Features extraction,
d) Verification/Validation.

Researchers are working to evolve a robust method to solve the problem of HWR system [7], but progress is still ongoing in this direction and not converged yet.

A human population can be classified in the various classes of gender, age, races, nationality based on their handwriting [8]. The work is at initial stage to predict the gender, age, etc., of the writer automatically through their handwriting but still there are a few studies. Some of them are listed in Table 1.

Table 1 shows the details of the works done in this field in the past years, it portrays the objective of the work done by the authors, databases specifications (name of the dataset used, size of the dataset, language, etc.), feature extraction method, classification methods, results of their work, advantages and disadvantages of the methods used.

Liwicki system for classification purpose shows in [9, 10] even better results than the humans. It uses online features as the advantage of online features is that features like speed and direction of the writing is also measured which makes the system more efficient. There are some systems like [11, 12] which use combinations either features combinations or combine the classifiers using ensemble classifiers to increase their accuracy rates as compared to the single classifiers.

Handwriting is a blend of expression and adaptation of the person. Previously it was very tiring to predict the traits by the conventional methods. However, now it becomes straightforward to find the demographic traits of an individual with the use of different methods [26]. Graphology is also a study of handwriting [27]. It shows the method to predict the personality traits through handwriting. Graphology is taught as a subject in many European countries such as Germany, France, Switzerland, etc., Graphologists inspect and interpret all the elements of writer's handwriting separately and produce a thumbnail outline of the character traits, and by using this, many handwriting analysts can predict the character of the writers. One such system which can find the correlation between these traits and the handwriting have been presented by B o u a d j e n e k and A g e [21]. This work is based on the extraction of two gradient features which are the histogram of gradients and local binary patterns for each cell of the image and then were concatenated to constitute the image feature vectors. M a a d e e d and H a s s a i n e [17] presented another system which even predicts the nationality of the writer along with gender and age in offline handwriting; it made use of combination of several geometric features for classification purpose.

Table 1. Literature review

Reference and year	Objective of the paper	Language/size of the data set	Method of feature extraction/features extracted	Classification method	Implementation result	Advantages/ disadvantages
L i w i c k i et al. [9]	Classification of writer for demographic traits (gender and handedness) from the hand written data	DATABASE USED-IAM-OnDB [13]. Online, Roman handwriting. Images from more than 200 writers having eight texts per writer	16 features for each successive pair of points: 1. speed 2. writing in (x, y)-directions (2) 3. curvature 4. normalized (x, y)-coordinates (2) 5. speed in (x, y)-coordinates (2) 6. overall acceleration 7. acceleration in x and y-direction (2) 8. log curvature radius 9. vicinity aspect 10. vicinity curliness 11. vicinity linearity 12. vicinity slope-cosine angle 13. vicinity slope-sine angle 14. ascenders 15. descenders 16. context map	1. SVM (Support Vector Machine) 2. GMM (Gaussian Mixture Method)	In case of gender: GMM – 67.06%, SVM – 62.19%. **Conclusion:** The system resulted even better than experiments taken by humans for classification	**Advantages:** 1. Better results than humans 2. Temporal as well as offline both info of the input is present because of the features extracted method. **Disadvantages:** Feature extraction method is quite complex as because of online data
L i w i c k i, S c h l a p - b a c h and B u n k e [10]	Classification of the handwritten data with respect to the gender of the writer	**Database** USED- IAM-OnDB [13]. Roman handwriting, online data and offline representation of online data and both the combination. Images of handwritten text of 200 different writers with each writer having eight texts and having average of seven lines per text	16 features for each successive pair of points : 1. speed 2. writing in (x, y)-directions (2) 3. curvature 4. normalized (x, y)-coordinates (2) 5. speed in (x, y)-coordinates (2) 6. overall acceleration 7. acceleration in x and y-directions (2) 8. log curvature radius 9. vicinity aspect 10. vicinity curliness 11. vicinity linearity 12. vicinity slope-cosine angle 13. vicinity slope-sine angle 14.ascenders 15.descenders 16. context map	GMM	Gender classification rate for test set: Online – 64.25%, Offline – 55.39%, Combination – 67.57%. **Conclusion:** The classification method results were better with online features and the mixture of both two are significantly better	**Advantages:** Classification results were comparatively accurate
S o k i c, S a l i h - b e g o v i c and A h i c-d j o k i c [14]	Analysis of the male and female handwriting to indicate that the gender provides some discri-minative difference in the handwriting of the two	**Database** used-BHDH [13]. The database consists of 3766 handwritings, i.e., the handwritings of 300 students of Electrical Engineering Sarajevo	They had used shape descriptors such as curvature function, tangent angle function, etc., using two approaches: • Generic Fourier Descriptors • contour based approach: ✦ tangent angle function ✦ Contour curvature and Bending energy	(GFD) Generic Fourier Descriptors	Male-Grad. Osc. Parameter=0.7615, Average bending energy=0.5563, Female-Grad. Osc. Parameter=0.6200, Average bending energy= 1.1910. **Conclusion:** Stereotypic examples of male and female handwriting show quantitative differences in spectral bands of the handwritten sample, curvature, speed, etc.	**Advantages:** Used methods were fast, simple and not sensitive to noise. **Disadvantages:** 1. GFD were de-veloped for shape recognition, and it captured the overall structure of the text, not the local contours details. 2. The broader conclusion about the gender cannot be made using GFD
X i e and X u [11]	Gender prediction from handwriting	QUWI dataset [15]. English and Arabic both text independent and dependent images. Data used from the compe-tition held on QUWI data-set. They used only 282 writers for training set, with each writer con-tributed with four documents samples	Images were first binarized using Otsu thresholding and Probability Distribution Function (PDF) is extracted from the images. Features extracted were: • Tortuosity • Direction • Curvatures • Chain code • Edge detection	1. kNN (k-Nearest Neighbor) 2. l_1 regula-rized logistic regression 3. decision tree and random forest	Combined result-kNN – 71.54%, tree – 62.53%, random forest – 72.57%. **Conclusion:** Using single feature such as middle axis direction has shown the best classification rate using LASSO and then random forest for all the combined features followed by kNN	**Advantages:** Random forest method has shown the best results combining the classification methods. **Disadvantages:** Not all features separately (single feature) were used in LASSO technique

Table 1 (c o n t i n u e d)

Reference and year	Objective of the paper	Language/size of the data set	Method of feature extraction/features extracted	Classification method	Implementation result	Advantages/ disadvantages
H u s s a - i n e et al. [16]	A competition was held on predicting gender from handwriting in 2013, i.e., ICDAR2013	Kaggle competition [57]. Subset of QUWI dataset, having both Arabic and English handwriting samples. Total of 475 writers produced four handwriting documents	They provided 30 features along with the images to the participants. the features include: chain code tortuosity, edge based features, etc.	Classification techniques used such as: SVM, random forests, gradient boosting machines, etc.	The edge based directional, chain codes and curvatures features results more discriminative than the other features. **Conclusion:** The best performance was achieved by a Gradient Boosting Machine both for feature selection and classification method	**Advantages:** Various methods were discussed for the prediction of gender based on handwriting
M a a d e e d and H a s s a - i n e [17]	Prediction of the age, gender and nationality of the writer using handwritten texts from offline images	**Database** USED-QUWI [15]. Offline data, both English and Arabic data. Dataset have handwritings of 1017 writers in both English and Arabic in both text-dependent and one text-independent	It binarized the data using otsu thresholding algorithm before feature extraction. The features were defined by the PDF extracted from handwritings images. Features extracted: 1. direction feature 2. curvature feature 3. tortuosity feature 4. chain code feature 5. edge-based directional feature	1. Random Forest (RF) 2. Kernel Discriminant Analysis (KDA)	Gender-KDA – 73.6% approx., RF – 74.8% approx. **Conclusion:** RF is generally ideal for prediction of age range and nationality whereas KDA is chosen for gender prediction	**Advantages:** 1. No need to measure the temporal information. 2. Chain code feature outperforms the features for prediction. **Disadvantages:** RF is not good with new data
S i d d i q i , D j e d d i and R a z a [18]	Study of a system for predicting the demographic classes such as gender from the handwriting of the writer	2 datasets were used: QUWI (English and Arabic) [15] and MSHD (French and Arabic) [19] **QUWI DATA-SET-**Writing samples of 475 writers, i.e., total of 1900 samples. **MSHD DATA-SET-**Dataset contains 87 writers' samples. Each writer contributed 12 pages, 6 for French and 6 for Arabic	Discriminative attributes of writings such as : Slant/orientation and Curvature/roundedness- were estimated from the contours of writing by: • Freeman chain codes • a set of approximating line segments (polygon zed contours) Neatness/legibility- • estimated by computing the fractal dimension of the writing Writing texture- • local binary patterns • Auto Regressive (AR) coefficients'	ANN (Artificial Neural Network) SVM	QUWI DATASET=SVM-68.75% ANN- 67.50%/ MSHD DATASET= SVM- 73.02% MSHD- 69.44% **Conclusion:** The two gender groups share some common attributes that are consistent even geographically	**Advantages:** 1. Two classifiers enhance the overall classification rates. 2. Slant and curvature features performs better than fractal and texture features. **Disadvantages:** The slant and curvature features being local features were more sensitive to the content of the image
B o u a - d j e n e , N e m m o u r and C h i b a n i [20]	A System to classify writers as male and female using textural features	Dataset used-IAM-OnDB [13]. Offline data, English. Dataset having 200 images from 200 writers	Features extracted – • Textural features using Local Binary Patterns (LBP) • Gradient features using histogram of gradients	SVM	First and second training set – LBP – 70% & 63%, HOG – 74% & 70%. **Conclusion:** HOG features outperforms all the other features with significant classification rates	**Advantages:** 1. Gradient features are better than local binary patterns. 2. Drawbacks of local binary patterns was removed with the help of gradients
B o u - a d j e n e k and A g e [21]	Classification of persons according to their biometric traits such as age, gender and handedness	**Database** used-IAM and KHATT [21, 22]. Offline, English and Arabic data images. 75 samples from each class were randomly selected from IAM dataset. There were 135 samples per class for gender prediction in KHATT dataset	Two Gradient features were extracted: • Histogram of gradients • Local binary patterns	SVM	Gender – HOG – 60% for IAM-1, –75.45% for IAM-2, –68.89% for KHATT, GLBP – 76% for IAM-1, –75.45% for IAM-2, –74.44% for KHATT. **Conclusion:** GLBP (Gradient local binary patterns) outperforms the HOG features for both IAM-1 and KHATT dataset but was same for IAM-2	**Advantages:** 1. SVM is well suited for pattern recognition purposes. 2. Gradient features such as HOG shown a high performance rates. **Disadvantages:** 1. HOG even being the robust still not provided the high performance as compared to GLBP. 2. HOG is bothered from the noise

Table 1 (c o n t i n u e d)

Mirza et al. [24]	Gender prediction of the writer from handwritten text in an image by applying the Fourier transform to the matrix	A subset of QUWI database [15]. QUWI database having 1000 writers' handwriting samples with four samples for each writer	A Gabor filters is applied to the handwritten images. Then, mean and standard deviation of the images are collected in a matrix and the Fourier transform of that matrix is used as the feature for classification	ANN (Artificial Neural Network)	The result for the 4 tasks taken are- TASK A – 70%, TASK B – 67%, TASK C – 69%, TASK D – 63%. **Conclusion:** The proposed system gives the better results or classification rates than those of the winner systems in the competition under the same experimental settings	**Advantages:** 1. Gabor filter is well adapted method, it allow us to exploit the textural features. 2. It is comparatively easy and well suited method. **Disadvantages:** Gabor filter is to be used separately from different angles and it is unsupervised
Ahmed et al. [12]	Prediction of gender from images (off-line) using multiple classifiers and boosting, voting and stacking techniques to combine the classifiers and to enhance the classification	A subset of QUWI dataset [15], i.e., in Arabic and English. Handwritten samples by 1017 writers. Each writer contributed total of 4 samples, 2 in Arabic and 2 in English	Textural features including: • Segmentation based fractal texture analysis (SFTA) • LBP • HOG • Gray Level • Co-occurrence Matrices (GLCM)	Individual classifiers used: Decision Trees (DT), SVM, kNN (k-Nearest Neighbor), Random Forest (RF) and ANN on a four features (LBP, HOG, GLCM, AND SFTA) and on the combination on these features	The results is different for individual and ensemble classifiers. **Conclusion:** 1. The traditional classifiers such as SVM and ANN gives higher classification rates. 2. Learners that gives best results individually also performs well in ensemble classifiers	**Advantages:** 1. With the use of the ensemble classifiers the result obtained was more efficient. 2. The textural features LBP and GLCM outperform the HOG and SFTA
Gattal et al. [25]	Classification of the gender and boosting the feature extraction step using oriented Basic Image Features (oBIFs)	QUWI database [15] Language used – Arabic and English. Each writer contributed four samples to the database, i.e., two in Arabic and two in English both in independent and dependent text	It starts with the binarizing (preprocessing) images using global thresholding and then extracting the oBIFs and • making the oBIF histogram and • oBIF column histogram. These are then added together to make the feature vector for further processing	SVM having kernel parameter range [0, 100] and the soft margin parameter C fixed at 10	For script dependent – 77.07%, 79.50% and 75%. For script independent – 71.37%, 76% and 68%. For different subsets of experiments. **Conclusion:** Attributes likes neatness and homogeneity (similarity) can be effectively use to characterize the gender	**Advantages:** 1. This method is best among the other experiments on the same subsets. 2. oBIF features considered to be more effective in finding the correlations between handwriting and gender. **Disadvantages:** oBIF histograms and oBIF column histogram were not found better as a single features

Thus, such types of systems were used in literature to predict the gender and any other personality traits of the writer through handwriting named as handwriting recognition system. HandWriting Recognition (HWR) is the application of pattern recognition area. Pattern recognition is a classification process where given patterns are assigned in required order. The unique writing style of every individual makes this type of classification popular in several areas such as, forensics, handwriting recognition, biometrics, signature verification. In addition to this, processing of each category separately can narrow down the search to some extent and helps to provide improved results. Curiosity is more focused on determining the gender among all the traits. To predict writer's gender is very significant task for its applications. There are a few researches, e.g., [27, 28] which have shown that the handwriting of an individual contains a rich information about the writers, thus the handwriting can be used for predicting some personality traits such as age, gender, handedness. Researchers also had looked into the relationship between sex hormones with the

handwriting, and found that prenatal sex hormones (inherited from parents) do affect the women handwriting [29]. Studies, e.g., [30-32] have also shown the key differences between the male and female handwritings through the physical appearance of the handwriting. It was resulted that male handwritings are tending to be more hurried, sloppy, untidy, spiky, etc. On the other hand, handwriting of females was more decorative, neat, consistent, regular, delicate, etc., than males. One example of handwriting is shown in Fig. 1. Another example of writing is shown in Fig. 2 for the identification traits of writing of nurse, social worker and teacher.

Even though this type of research is gaining popularity and interest among researchers due to its significant applications but still, it has certain limitations as following:

1. The handwriting of individual can be different if the conscious state of that person results emotional experiences or thoughts and so the result may be erroneous.

2. There may be a situation when a left-handed child is taught or forced to write from right hand, which results in to modify the childhood personality hormones.

3. The handwriting samples should be independent from any drugs, diseases, anxiety, fatigue, etc., as these conditions modify the personality traits.

Hence, the prediction of gender from the handwriting becomes more challenging and need more research in continuation of the current advancement. In this section, various studies were cited based on their contribution in the field and compared comprehensively to understand the prior art in the field. Many types of databases used for solving the problem are also cited in Table 1 so that future advancement in the field can also be validated on the available datasets. Next section lists the methods used in the field, and also discuss them comprehensively.
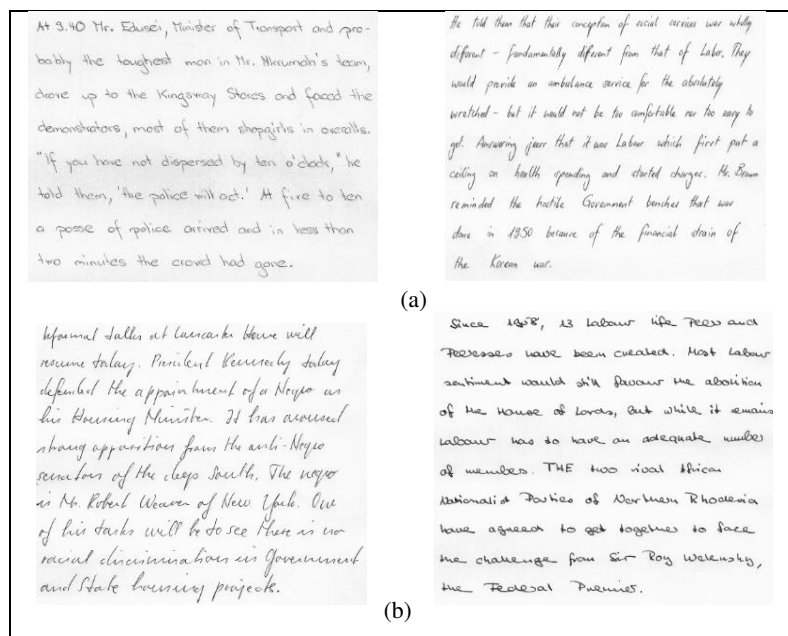

Fig. 1. Sample handwriting of females (a); sample handwriting of males (b)

Handwriting recognition has mostly been used in the field of signature verification or identification in forensic investigations. It is useful for forensic experts to investigate classes of writers. This article is written with a motivation of understanding the complete process for prediction of writer's sex based on his/her handwriting. This article describes the various challenges associated for solving the problem of prediction of writer's sex. It also includes several techniques conducted by various researchers for solving the problem. It compares the features extracted by the different experiments so that readers may understand the contribution of a feature in the handwriting-based prediction system.
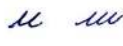
'u' shaped 'm' and 'n' – diplomacy and adaptability	'i' dots directly above stem – Concentration and precision	Letters 'm', 'n' and 'r' flat from top – Technical proficiency	Size of handwriting small – Concentration and good memory
u uu	*i i i*	*m n r*	*small handwriting*
Final strokes long and round – humanity and charity	Letters 'a', 'o', 'd','g' mostly closed – diplomacy and discretion	Vertical slant – controlled emotions	Connected letters seven or more – fluent speaking or writing
a d s o	*o a d g*	*vertical forward*	*implicit word*
't' crosses low – sympathy, prompt action	Initial strokes omitted – self-dependence	Spacing between words are equal – consistency	Legible writing – accuracy
t t t	*a w k*	*Even*	*Clear, Simple*

Fig. 2. An example of writing samples of nurse, social worker and teacher

2. Related work

Most of the results depend on the methods of selection and extraction of features. It needs to adapt the features that are suitable for text classification. Before extracting the features one must acquire the input image in the required circumstances and the suitable preprocessing can be done on the images.

2.1. Image acquisition

Images can be acquired in two ways: (a) online, and (b) offline images. Image acquisition depends upon the type of data. For online data, e-beam interface is used. It generates the x, y-coordinates sequence representing the position of the tip of the pen and the counter for each location. Fig. 3 shows the recording process of online data through any e-beam interface. For offline data, optical character reader is used. It converts the handwritten or printed text into machine codes using optical scanning.

Fig. 3. Recording of online data

2.2. Datasets

There are various data sets available for handwriting analysis consists of thousands of handwriting samples in different languages. Some of the mostly used datasets are two.

2.2.1. QUWI dataset [15]

This dataset is developed by Qatar University. QUWI stands for Qatar university writer identification dataset. This is an offline dataset. QUWI dataset contains the writing samples from the 1017 writers. Writer's samples were taken by both blue and black pen as well as pencil. To avoid any misleading the volunteers (writers) were first asked to fill the information page including age, gender and handedness, writer's profession, education level and nationality. However, the name is kept confidential due to privacy policy. Writers were then asked to write four pages, i.e., first page consists of six handwriting lines in Arabic language of writers imagination, second page consists of three paragraphs to be copied by volunteers in Arabic language. Third and fourth page consists of English language samples. Third page consists of six handwriting lines of writer's imagination and fourth page consists of three paragraphs to be copied by them. Thus, first and the third page are used for text dependent tasks and second and fourth page – for text-independent tasks. Not all volunteers know Arabic and some were the beginners to the English language. This dataset was constructed in 1017 folders with total of 4068 documents in it. It is available for free for commercial as well as for non-commercial research work.

2.2.2. IAM dataset [13]

This dataset was developed by research group at computer vision and artificial intelligence at Bern University. IAM dataset contains handwritten English text which

is used for handwriting writer identification purposes. This dataset was first published in ICDAR 1999 in [33]. It contains unconstrained forms of handwritten text that can be scanned at 300 dpi resolution and can be saved as image with .png extension. This database is structured as: 657 writers' handwriting samples were contributed in it out of which 1,500 pages having scanned text, 5,600 unique sentences and labeled, 13,300 is unique and labeled text and 11,532 are isolated and labeled words. IAM Handwriting DB 3.0 has offline images from the website. IAM On-line handwriting database data were acquired on whiteboard and it contains: writing samples from 221 writers with 1,700 and more forms of 13,049 writing samples that is different and labeled text lines in on-line and also in off-line format and 86,272 word samples from a 11,059 words dictionary. IAM OnDB is freely and publicly available for non-commercial research uses.

2.2.3. KHATT dataset [21, 22]

KHATT stands for **KFUPM (King Fahd University of Petroleum and Minerals) Handwritten Arabic TexT.** KHATT dataset originally known as King Fahd University of Petroleum and Minerals (KFUPM) Handwritten Arabic text database, now known as KHATT was published by [21, 22] head of research group of KFUPM recently on November 16, 2015. This dataset consists of a free handwritten Arabic text from 1000 writers. All Writers are from diverse countries, age, gender, education level and handedness. It contains 2000 unique-text images and 2000 similar text images. The KHATT version 1.0 is freely available for academic research work.

After acquisition of input image, next important step is to apply some preprocessing technique on the data. There are various factors that may compromise the quality or the accuracy of the text through OCR such as scanner quality, resolution, fonts, paper quality, etc. Therefore, the importance of preprocessing is to overcome these problems before feature extraction.

2.3. Preprocessing techniques

Preprocessing techniques applied in handwriting recognition are as follows.

1. The normalization operations were applied before feature extraction as the recorded online data contains the noise and gaps within stroke in online data [9].

2. To make the text pen independent it binarized the data using Otsu thresholding algorithm [34] before feature extraction[17].

3. Another preprocessing technique is to apply on handwritten images to make them of equal size which may be achieved by grids to divide it into a number of cells [21].

2.4. Feature extraction

Feature extraction is the processing of finding or computing influencing and differentiating attributes which may help to classify the particular item in a given set. This section will explain three different feature extraction methods.

2.4.1. Local and global features

The feature set contains the features extracted from online data as well as the offline version of the online data. J a e g e r et al. [35] explained 18 features extracted for the text recognizer in their online handwriting recognition system named as the NPen++. The 18 features were computed for a given stroke, between a consecutive pair of points. The following online features were computed in their study:

- **Speed (1).** To calculate how fast a writer can write which shows the amount of energy a person has.
- **Writing direction (2).** The writing direction in x and y coordinates.
- **Log Curvature radius (2).** Length of the circle which defines the curvature at that point.
- normalized x and y-coordinates (2)
- speed in x- and y-coordinates (2)
- overall acceleration (1)
- acceleration in x- and y-direction (2)
- **Log curvature radius (1).** To find the curvature radius of the writing.
- **Vicinity aspect (1).** To find the feature of the path followed by the point in the surroundings of the point.
- **Vicinity Curliness (1).** To measure the deviation in the neighborhood of the point from the straight line.
- **Vicinity Linearity (1).** The average square distance between every point in the surroundings and the straight line connecting the first and the last point in the neighborhood.
- **Vicinity slope (2).** Cosine and the sine angle of the straight line from the first to the last region point

The offline features were extracted using two-dimensional matrix, representing the offline version of the data that are:

- **Ascenders/Descenders (2).** The number of points whose x-coordinates are in the area of the point having minimum distance to the corpus/base line and that points are above/below the base line,
- **Context map (9).** The 2-dimensional area of the point is divided into three parts for each dimension.

The number in their brackets gives the number of each feature. Thus, there will be 29 total numbers of features in this feature set.

2.4.2. Geometric features

Geometric features are the features that are manufactured by a set of arithmetic attributes such as curves, lines, points or surfaces. These features are ridges, corners and blobs, etc. are defined for a PDF which can be extracted from a handwriting text images. M a a d e e d et al. [36] have used these geometric features for writer identification. There are four of the geometric features.

2.4.2.1. Tortuosity feature

Tortuosity is the property of any shape being curved or twisted (having many turns). This feature helps in determining the fast writers who produce twisted handwriting or the slow writers who produce smooth handwriting. As, fast writers will have more curves in their text and the slow writers will have less curves. In case of text, for each pixel, some features were considered related to tortuosity and it represents the length of the greatest line segment that is completely inside the text and which traverses the pixel. The PDF features represent the direction of the text. Generally, the simplest method to calculate the tortuosity of a 2D image is the arc-chord ratio as shown in Fig. 4 and listed in the following equation:

$$(1) \qquad\qquad\qquad\qquad T = \frac{L}{C},$$

where L is the curve length, and C is the distance among the ends of the curve.
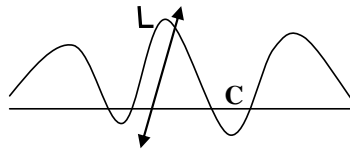
Fig. 4. Tortuosity of a line

2.4.2.2. Curvature features

In general term, curvature at any point in the circle is basically known as the reciprocal of the radius. It is mostly used in forensic examinations [9]. It is used to find the curvature of the contours. In images, for each pixel in the image, belonging to the contours, a neighboring window is considered of fixed size. Inside the window, the number of pixels n_1 belonging to the background and number of pixels n_2 belonging to the foreground are computed. If the difference $n_1\text{-}n_2$ of all the contour pixels is positive then the contour will be convex if $n_1\text{-}n_2$ is negative then the contour will be concave. Therefore, in this way the local curvature of the contours is detected and also shown in Fig. 5 and listed in the following equation:

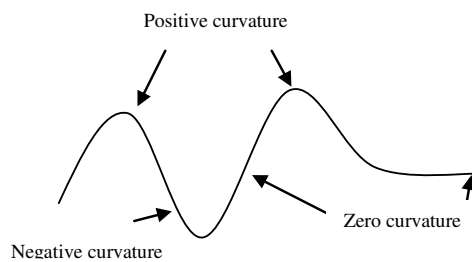$$(2) \qquad\qquad\qquad\qquad C = \frac{n_1 - n_2}{n_1 + n_2}.$$

Fig. 5. Curvature values at different points in a curve

2.4.2.3. Chain code features

It is the method of representing the boundary of the shape by a straight line segment of fixed length. Chain codes are generated by scanning the outline of the text and

passing on each pixel in the curve a number according to its previous pixel. Chain codes can be applied at different orders such as 4-dimensional directions or 8-dimensional directions as shown in Fig. 6. This feature gives a detailed description about the curvature of the text. It has been used in writer identification [38].
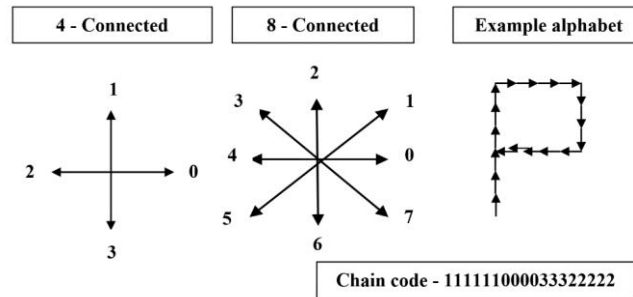

Fig. 6. 4-directional chain code (a); 8-directional chain code (b);
Example for 4-directional chain code (c)

2.4.2.4. Edge based directional features [39]

This geometric feature gives a detailed description about the directions of the shape. This is done by allocating a window of different sizes having center at each contour pixels and then counting the occurrences of the direction as the pixel moves one by one. This can be applied not only to the outline of the moving window but also of the entire window.

2.4.3. Gradient features

Image gradient means directional changes in the concentration (intensity) or color of an image. For calculating the gradient changes two feature descriptors were computed. Feature descriptor is a symbol of an image which is a patch from an image and useful for removing the extraneous information of an image. In these methods, textural information was used for feature extraction and gradient information was used as weight for histogram. Two gradient algorithms (HOG and GLBP) cited in the literature [40].

2.4.3.1. Histogram of Oriented Gradients (HOG)

HOG is a one of the feature descriptors which is used for object detection. This was introduced by D a l a l et al. [41]. This technique counts the gradients (changes) areas in the local parts of an image. HOG descriptor shows high performance in various signature verification, identification [42] or face recognition application [43, 44]. HOG is developed from SIFT algorithm [45]. In HOG descriptor, directions of oriented gradients were used as a feature, as the amounts of gradients were large around the corners and edges. Generally the corners and edges have more information about the shape of the image than the flat surface. Following algorithm is used for finding HOG:
- Preprocessing: Segment the images into a number of cells by applying a grid.

62

- For each cell, HOG features is computed as follows:

For each pixel $I(x, y)$, calculate the vertical and horizontal gradient information as:

(3) $$g_x(x, y) = I(x + 1, y) - I(x - 1, y),$$

(4) $$g_y(x, y) = I(x, y + 1) - I(x, y - 1).$$

Calculate the magnitude and direction of each pixel as:

(5) $$M(x, y) = \sqrt{g_x{}^2 + g_y{}^2},$$

(6) $$\theta = \arctan \frac{g_y}{g_x}.$$

Calculate the histogram of gradients: Histogram is calculated over the patches instead of the whole image as to make it robust to noise. For the histogram, Q bins are chosen for example $Q=9$ (i.e., between $0°$ to $180°$).

2.4.3.2. Gradient Local Binary Patterns (GLBP)

In GLBP method the textural information and the gradient information are combined, similar to any other gradient and texture algorithms, this is more powerful for edge detection as it removes the pixel noise, and it gives different weights to each pixel by focusing the important portions of the images. It was also used for human detection [40]. The algorithm was used as follows:

- Initially, original pixel value and its 8-binary code are read from memory.
- The input image window is divided into several blocks (Cells).
- The 8-binary code is computed by comparing the value of the middle pixel with its 8-neighboring pixels value one by one. If the value of the middle pixel is bigger, then "1" is the resultant binary value otherwise "0" is the resultant binary value.
- Then the binary value of neighbor pixels is put together to get the several "1" area and several "0" area.
- Check the pattern is uniform or non-uniform;
- If, the "1" area and "0" area appears only once in the pattern of binary code then it is uniform pattern otherwise non-uniform pattern. For example; binary pattern"00111000" is a uniform pattern and patterns "01110101" are non-uniform pattern.
- Else, ignore this pixel and calculate binary code for another pixel.
- For each pixel with the uniform pattern in the block, calculate the gradient value (angle value) and GLBP table position value (width value).Width value is the number of "1" in the binary code pixel. There are 8-direction codes from 0 up to 7 in 8 directions of the eight neighbor pixels and angle value is the direction code of the middle pixel in the "1" area of its binary code. If the width value is even, then the angle value is smaller value of two direction values except the case when the middle direction of "1" is between 7 and 0 directions, in this case we set angle value as 7.
- Calculate the gradient value for each pixel, as:

(7) $$g_x(x, y) = I(x + 1, y) - I(x - 1, y),$$

(8) $$g_y(x, y) = I(x, y + 1) - I(x, y - 1),$$

(9)
$$M(x, y) = \sqrt{g_x{}^2 + g_y{}^2},$$

- Position of bin in the GLBP table is mapped from angle value and width value calculated in the step above. The gradient value is written into the bin.
- Then, GLBP histogram is derived within each cell using L2 normalization.

2.5. Classification methods

Classification and identification problems are closely related to one another but the only difference between the two is in case of gender classification. It is a two-class problem whereas gender identification is an N-class problem. The features that are extracted for the identification problem will perform well in the classification problem. However, some of the classification and identification methods are listed below which were used for gender prediction based on handwriting.

2.5.1. Support Vector Machine (SVM)

This is a discriminative classifier [46], which maximizes the discrimination between the classes by using the maximal margin hyperplane. This is a binary classifier[47]. It is used for both classification and regression applications [48]. However, applications are mostly in the field of classification. It is a supervised learning algorithm (that is for a given labeled training data, its output is maximum margin hyper plane which separates the two classes). In this algorithm, each data item (input) is plotted as a point in n-dimensional space (n is a number of features) and the feature value is the value of the coordinate. In a 2D space, hyper plane is a line which distinguishes the two classes by dividing the plane in two parts and each class is in one part of the plane as shown in Fig. 7. This algorithm outputs the hyper plane as a result of classification; the data item in either class belong to that class. It gives the hyper plane with maximum margin and the one for which the distance to the closest point in either class is maximum.

If the two classes are non-separable then to handle such a situation SVM has a technique known as kernel trick. It transforms the low dimensional input space into high dimensional input space, i.e., one more dimension is added called as z-axis. Now, when z and x-axis is plotted it transforms the non-linear problem into linear problem. The z-axis is calculated as; $z = x^2 + y^2$. the function of predicting the location of the new data item is calculated as in (10):

(10) $F(x) = B(0) + \text{sum} (a_i * (x, x_i))$,

Which is the linear product of input x_i and the support vectors x_i and $B(0)$, a_i are the coefficients which must be computed by the training data.

Two classes can never get classified without any error because of the outliers or the noise in the data, to solve such problems SVM makes use of Tuning parameters as given below. D o n g and J i a n [49] gives a method for parameter selection based of a support vector machine.

- Regularization parameter: This tells the SVM how much to avoid the outliers (misclassifications) in training data. It is denoted by C. For large values of C, SVM choose the smaller margin hyperplane, i.e., the one which do the classification correctly.

64

For the smaller value of *C*, SVM choose the larger margin hyperplane, even if it misclassifies the data.

• Gamma parameter: It defines how far the effect of training data reaches, i.e., how far margin should be included. For the low values of gamma, far away points are also included in decision of separating hyperplane, whereas for high values, close points are only considered for deciding hyperplane.

SVM is used as a classifier by [25] for gender classification using oriented basic image features where SVM is used with kernel parameter [50] that is selected in the range[0, 100] while the soft margin parameter *C* that is selected as 10.

For further work as a classifier in text recognition, it takes each word or sentence or block of a handwritten text and find the presence of the particular feature (curve, direction, edge, etc.) and the value of each feature (i.e., histogram values, tortuosity, curvature value, edge directional feature, etc.), is encoded as a data point in *n*-dimensional space. Then SVM will give a hyper plane to classify the features into two classes (male and female) and for nonlinearly separable classes it uses either Radial Basis Function (RBF) [51] or a kernel function[25].

2.5.2. Random Forest (RF)[17]

Random forests are the ensemble learning method for classification and regression. Random forests are made of a decision trees. Decision trees are easy to build, easy to use and interpret. However, its only drawback is that it is not accurate in practice which means they are good for the training data but not flexible for the classification of the new sampling data. Therefore, as now random forests came into picture here, they combine the simplicity of the decision trees with the flexibility to improve the accuracy. Flowchart of the algorithm of random forests is shown in Fig. 8 and also explained in steps as follows.

Step 1. Create a Bootstrapped dataset: Create a Bootstrapped dataset from the original dataset of the same size as that of original. For creating a bootstrapped dataset random samples are chosen from the original dataset. Samples can be chosen more than once as duplicity is allowed.

Step 2. Create a decision tree: Create decision trees from the bootstrapped dataset by selecting a random subset of variables (features considered) at each step. For example, in case of total of four variables randomly considered two variables (columns) at each step. The root of the decision tree is selected by checking the impurity of the columns (variables) using famous GINI method [52], the one with the minimum is selected as a root.

Step 3. Create a random forest: repeat from Step 1 by making a new bootstrapped dataset and then decision tree by using a different subset of variables (Columns). This will result in a wide variety of trees. This variety makes the random forest better and more efficient than individual decision tree.

Step 4. Evaluation of the random forest: Evaluation of RF is done by using the "Out-Of-Bag (OOB)" dataset. The dataset that is not included in the bootstrapped dataset is known as OOB dataset. Since, the OOB dataset is not used to create the decision trees, we run it through all of it and check whether it correctly classifies the data or not. Random forest accuracy is measured by the proportion of the OOB

samples that were correctly classified and the OOB samples that were not correctly classified are known as OOB error.

RF is an ensemble of CART (Classification And Regression Tree) [53]. It is used for a text classification as it creates a bootstrapped dataset using n cases from the original dataset having N cases such that $n < N$ which is obtained after feature extraction (edges, direction, speed, angles, etc., of the text) with replacement. This will give rises to a decision trees using out of bag features from the above mentioned features from dataset. These decision trees use the RF classifier using some of the input (handwritten text images features) for training. Further the classification result is obtained by evaluating each decision tree of a random forest and by counting the voting for the result.
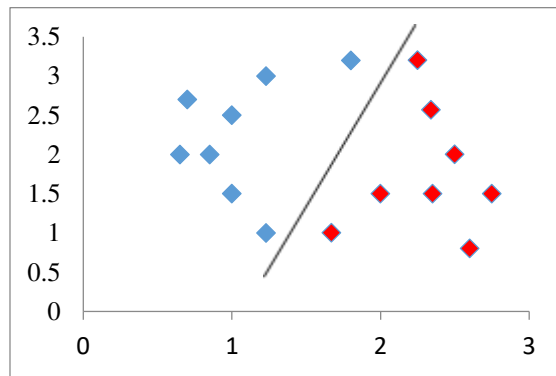

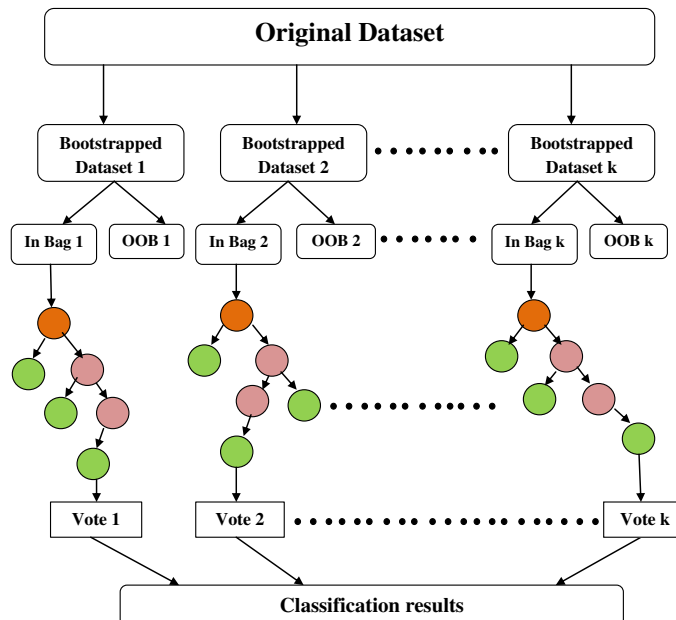Fig. 7. Hyperplane separating two classes in SVM classifier

Fig. 8. Random forest algorithm flowchart

66

2.5.3. kNN (k-Nearest Neighbor) method

It is non-parameterized method, it is used for classification, regression [45] and as a clustering process [54]. Its objective is to predict the class of the new input using the database in which data points are already classified into classes (trained). It is also known as lazy algorithm. It is supervised learning algorithm and has applications in pattern recognition [55], data mining [56], etc. It classifies the input on the origin of the distance between the new input data items and the nearest training samples. To find out the nearest neighbor of the data points, the most common used distance is *Euclidean distance* (if the input is of similar type). It is calculated as

(11) $$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2},$$

where $q_i - p_i$ is the difference between the new point and the existing point. Suppose, the data points have 2-dimensional coordinates then the Euclidean distance can be calculated as

(12) $$d(p,q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2},$$

where point p_i has the coordinates as (x_1, y_1) and point q_i has the coordinates as (x_2, y_2), where i is as per the number of coordinates.

Other popular distances are the following that may also be used as per the application:

- *Hamming Distance*: Distance between binary vectors;
- *Manhattan Distance*: Distance between the real vectors, also called City Block Distance (if the input variable are not of similar types);
- *Minkowski Distance*: Simplification of Euclidean and Manhattan distance.

Algorithm:

Step 1. Select the value of k (number of nearest point to be taken around the new input point), for larger values of k noise from the data can be removed but it makes the classification less distinct. There are various techniques to decide the value of k such as hyper parameter optimization, tuning or ensemble learning approach [57].

Step 2. Find the distance between the new input points and the neighbors of it.

Step 3. Sort the distances.

Step 4. Pick the distances from all the sorted distances (number of distances will be determined by the value of k).

Step 5. The one data point having the minimum distances will become the class of the new point.

Since, this is a supervised algorithm, it classifies the new input (handwritten text as of male or female writer) using a trained classifier (which is trained using the text features, i.e., curves, edges, acceleration of writing, speed, curliness, roundedness, etc. of the text). Then further, the new input (text image) is classified by selecting the k value (up to which nearest points should be taken) for classification comparison. It is used for text categorization by [58].

3. Discussion

There are a large number of feature extraction and classification methods which are applicable to different applications. This survey has included many of them and found two gradient methods better for solving the problem of gender classification. GLBP and HOG show high classification rates for tasks having signature verification or identification. It uses the feature descriptors for extracting gradients which in result removes all the unnecessary information from the image.

Even HOG being great feature descriptor GLBP still outperforms the HOG. As HOG being a robust classifier does not give the best results as compared to the GLBP but still it is the best from others. The disadvantage of HOG is that it gets troubled by the noise that even one pixel having noise will influence the near (neighboring) pixels a lot and it may have a large value to the histogram [21].

Among all classification algorithms, the highest performance rate is given by decision trees. And within decision trees, the highest accuracy rate was of RF. SVM also performs well but the classification rate depends upon the features extraction method used.

Table 2 describes the overview of the gender prediction systems that is extracted from the literature. It shows the generalized steps of solving the problem of gender classification based on handwriting. The problem is divided in three major steps which were used almost every other method available in the literature. Preprocessing, feature extraction and classification methods were used but with different types of methods, respectively results were obtained.

Table 2. Generalized steps of gender prediction system with some methods used in literature

Reference No	Preprocessing technique applied	Feature extraction method used	Classification method	Result	
[21]	Grid is applied to images	Gradient features	SVM	For script dependent: 77.07%, 79.50% and 75% For script independent: 71.37%, 76% and 68%	
[12]	None	Textural features	SVM, DT, RF, ANN, kNN and ensemble classifiers	INDIVIDUAL ANN – 74% SVM – 74% DT – 67% RF – 64% kNN – 68%	ENSEMBLE ANN – 74% SVM – 74% DT – 71% RF – 71% kNN – 68%
[17]	Binarized the data using Otsu thresholding algorithm [34, 59]	Geometric features	Random Forest (RF) and Kernel Discriminative Analysis (KDA)	Gender: KDA – 73.6% app RF – 74.8% app	
[20]	None	Textural features: Local Binary Patterns (LBP) and gradient features using histogram of gradients	SVM	First training set: LBP – 70%, HOG – 74% Second training set: LBP – 63%, HOG – 70%	

Five different methods are listed in Table 2 for solving the gender classification problem. These methods give comparatively better accuracy. Each method is explained in Table 2 in terms of methods applied for a particular step along with the result. Fig. 9 gives a summative overview of the solutions of the problem. Accordingly, researchers working in the field may be benefited with these results to evolve a comparatively better algorithm. Table 3 is also included to demonstrate the available databases for the validation of the algorithms developed for the prediction of gender using handwriting recognition.
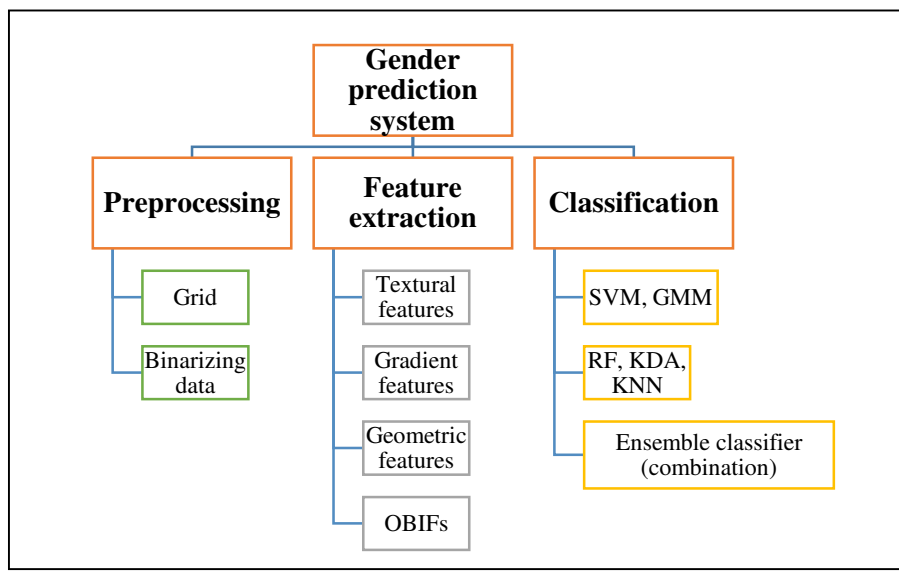

Fig. 9. Summative overview of the methods involved to solve the problem

Table 3. Summative list available databases for the training and testing of the algorithms developed for the prediction of gender based on handwriting recognition

No	Name of database	Source of database
1	IAM database [13]	**http://www.fki.inf.unibe.ch/databases/iam-on-line-handwriting-database/download-the-iam-on-line-handwriting-database**
2	BHDH database	**http://people.etf.unsa.ba/esokic/BHDH/BHDH.html**
3	AHDB dataset [60] ICDAR 2011 dataset, ICHFR 2012, QUWI Handwriting dataset[15]	**http://handwriting.qu.edu.qa/dataset/**
4	ICDAR2013, Kaggle database	**https://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting**
5	KHATT database[21, 22]	**http://khatt.ideas2serve.net/KHATTAgreement.php**

4. Result

The various methods for the prediction of gender based on handwriting were studied and listed in the table highlighting the advantages and disadvantages of the methods. The databases used for such type of studies were also listed along with the attributes

of each database to further use in future by the readers. This survey article comments on the methods used for solving the mentioned problems and aggregate the result of each methods so that reader may understand the performance of each method along with the pros and cons of the particular method.

5. Conclusion

With the implementation of various mentioned methods, there is still a large scope to develop a robust algorithm to solve the challenges associated with the problem. The problem is divided into a number of steps to understand it well, and methods available in literature were also discussed to solve each step's problem. The accuracy of the whole system exclusively depends on two major steps: (a) extracted features and (b) classification based on those features. A large number of features were considered in the literature while a few were contributing and others may not be contributing. To classify accurately, it is important to consider the most influencing features which includes the writer's traits. With the variability of the writer's language, these traits may also vary. A classification method is applied on the extracted features to predict the correct class. The classification problem may also be extended to predict more number of classes not only based on the gender but also identify persons based on the available traits. Many databases are available for validating the results of a newly developed algorithm for the prediction of gender in different languages which may be useful for the researchers working in the field. This survey article provides a new insight to the researchers for understanding of the problem and helps to evolve with an improved methodology for the same.

R e f e r e n c e s

1. B o y a d z h i e v a, D., G. G l u h c h e v. A Combined Method for On-Line Signature Verification. – Cybernetics and Information Technologies, Vol. **14**, 2014, No 2, pp. 92-97.
2. D a f e, S. G., S. S. C h a v h a n. Optical Character Recognition Using Image Processing. – IRJET, Vol. **5**, 2018, Issue 3, pp. 962-964.
3. K a u r, S. Gurmukhi Printed Character Recognition Using Hierarchical Centroid Method and SVM. – International Journal of Computer Applications, Vol. **149**, 2016, No 3, pp. 24-27.
4. S h a r m a, A., S. K h a r e, S. C h a v a n. A Review on Handwritten Character Recognition. – IJCST, Vol. **8491**, 2017, Issue 1, pp. 71-75.
5. F a k h r, M. W. Arabic Optical Character Recognition (OCR). – Systems Overview, Vol. **2011**, January 2011.
6. R u b e e n a, L. R. B. A Comprehensive Study on Handwritten Character Recognition System. – IOSR Journal of Computer Engineering Ver. IV, Vol. **17**, 2015, No 2, pp. 2278-661.
7. Y a d a v, P. Handwriting Recognition System – A Review. – International Journal of Computer Applications, Vol. **114**, 2015, No 19, pp. 36-40.
8. G o l d b e r g, L. R., D. S w e e n e y, P. F. M e r e n d a, J. E. H u g h e s. Demographic Variables and Personality: The Effects of Gender, Age, Education, and Ethnic/racial Status on Self-Descriptions of Personality Attributes. – Person. Individ. Diff., Vol. **24**, 1998, No 3, pp. 393-403.
9. L i w i c k i, M., A. S c h l a p b a c h, P. L o r e t a n, H. B u n k e. Automatic Detection of Gender and Handedness from On-Line Handwriting. – Journal of Social Psychology, 2007, No March, pp. 179-183.

10. L i w i c k i, M., A. S c h l a p b a c h, H. B u n k e. Automatic Gender Detection Using On-Line and Off-Line Information. – Pattern Analysis and Applications, Vol. **14**, 2011, No 1, pp. 87-92.

11. X i e, Q., Q. X u. Gender Prediction from Handwriting. – Data Mining Course Project. – Fall 2013. 2013, pp. 10-13.

12. A h m e d, M., A. G h u l a m, H. A f z a l, I. S i d d i q i. Improving handwriting Based Gender Classification Using Ensemble Classifiers. – Expert Systems with Applications, Vol. **85**, 2017, pp. 158-168.

13. M a r t i, U. V., H. B u n k e. The IAM-Database: An English Sentence Database for Off-line Handwriting Recognition. – International Journal on Document Analysis and Recognition, Vol. **5**, 2003, No 1, pp. 39-46.

14. S o k i c, E., A. S a l i h b e g o v i c, M. A h i c-D j o k i c. Analysis of Off-Line Handwritten Text Samples of Different Gender Using Shape Descriptors. – In: IX International Symposium on Telecommunications (BIHTEL), 2012, pp. 1-6.

15. A l M a a d e e d, S., W. A y o u b y, A. H a s s a ̈ ı n e, J. M o h a m a d A l j a a m. QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification. – In: IIEEE International Conference on Frontiers in Handwriting Recognition QUWI, Vol. **95**, 2012, No 15, pp. 742-747.

16. H a s s a ï n e, A., S. A l M a a d e e d, J. A l j a a m, A. J a o u a. ICDAR2013 – Competition on Gender Prediction from Handwriting. 2013.

17. A l M a a d e e d, S., A. H a s s a i n e. Automatic Prediction of Age, Gender, and Nationality in Off-line Handwriting. – Eurasip Journal on Image and Video Processing, 2014, pp. 1-10.

18. S i d d i q i, I., C. D j e d d i, A. R a z a. Automatic Analysis of Handwriting for Gender Classification. – Pattern Analysis and Applications, Vol. **18**, 2015, No 4, pp. 887-899.

19. D j e d d i, C., A. G a t t a l, L. S o u i c i-M e s l a t i, I. S i d d i q i, Y. C h i b a n i, H. E l A b e d. LAMIS-MSHD: A Multi-Script Offline Handwriting Database. – In: Proc. of International Conference on Frontiers in Handwriting Recognition (ICFHR'14), Vol. **2014**, December 2014, pp. 93-97.

20. B o u a d j e n e k, N., H. N e m m o u r, Y. C h i b a n i. HOG and LBP Features for Writer's Gender Classification. – International Conference on Electrical Engineering and Control Applications, 2016, pp. 317-325.

21. B o u a d j e n e k, N. H. N., Y. C. A g e. Gender and Handedness Prediction from Handwriting Using Gradient Features. – Complexity, 2015, pp. 1116-1120.

22. M a h m o u d, S. A., I. A h m a d, M. A l s h a y e b, W. G. A l-K h a t i b, T. U. B r a u n s c h w e i g, S. A. M a h m o u d, I. A h m a d, M. A l s h a y e b, W. G. A l-K h a t i b, M. T. P a r v e z, G. A. F i n k, V. M a r g n e r, H. E. A b e d. KHATT: Arabic Off-line Handwritten Text Database. – In: International Conference on Frontiers in Handwriting Recognition (ICFHR'12), 2012, pp. 449-454.

23. M a h m o u d, S. A., I. A h m a d, W. G. A l-K h a t i b, M. A l s h a y e b, M. T a n v i r P a r v e z, V. M ä r g n e r, G. A. F i n k. KHATT: An Open Arabic Off-line Handwritten Text Database. – Pattern Recognition, Vol. **47**, 2014, No 3, pp. 1096-1112.

24. M i r z a, A., M. M o e t e s u m, I. S i d d i q i, C. D j e d d i. Gender Classification from Off-line Handwriting Images Using Textural Features. – In: Proc. of International Conference on Frontiers in Handwriting Recognition, ICFHR, 2017, pp. 395-398.

25. G a t t a l, A., C. D j e d d i, I. S i d d i q i, Y. C h i b a n i. Gender Classification from Offline Multi-Script Handwriting Images Using Oriented Basic Image Features (oBIFs). – Expert Systems with Applications, Vol. **99**, 2018, No February, pp. 155-167.

26. G r e w a l, P. K., D. P r a s h a r. Behavior Prediction through Handwriting Analysis. – InternatIonal Journal of Computer Science and Technology, Vol. **3**, 2012, No 2, pp. 520-523.

27. R i z v i, F. Personality Prediction through Offline Handwriting Analysis. 2017, No February.

28. T e t t, R. P., C. A. P a l m e r. The Validity of Handwriting Elements in Relation to Self-Report Personality Trait Measures. – Personality and Individual Differences, Vol. **22**, 1997, No 1, pp. 11-18.

29. W a r n e r, R. M., D. B. S u g a r m a n. Attributions of Personality Based on Physical Appearance, Speech, and Handwriting. Vol. **50**, 1986, No 4, pp. 792-799.

30. B u r r, V. Judging Gender from Samples of Adult Handwriting: Accuracy and Use of Cues. – Journal of Social Psychology, Vol. **142**, 2002, No 6.

31. R u b i n, D. L., K. G r e e n e. Gender-Typical Style in Written Language. – Source: Research in the Teaching of English, Vol. **26**, 1992, No 1, pp. 7-40.
32. K o p p e l, M., S. A r g a m o n, S. A. R a c h e l. Automatically Categorizing Written Texts by Author Gender. – Literary and Linguistic Computing, Vol. **17**, 2002, No 4, pp. 401-412.
33. M a r t i, U.-V., H. B u n k e. A Full English Sentence Database for Off-Line Handwriting Recognition. – Icdar, Vol. **1999**, pp. 705-708.
34. B a n g a r e, S. L., A. D u b a l, P. S. B a n g a r e, D. S. T. P. Reviewing Otsu's Method for Image Thresholding. – International Journal of Applied Engineering Research, Vol. **10**, 2015, No August 2016, pp. 21777-21783.
35. J a e g e r, S., S. M a n k e, J. R e i c h e r t, A. W a i b e l. On-line Handwriting Recognition: The NPen ++ Recognizer. – Int. J. Doc. Anal. Recognit., March 2001, No May 2017, pp. 169-180.
36. A l-M a a d e e d, S., A. H a s s a i n e, A. B o u r i d a n e, M. A. T a h i r. Novel Geometric Features for Off-Line Writer Identification. – Pattern Analysis and Applications, Vol. **19**, 2016, No 3, pp. 699-708.
37. K o p p e n h a v e r, K. M. Forensic Document Examination: Principles and Practice. – Forgery, 2007, pp. 55-60.
38. S i d d i q i, I., N. V i n c e n t. Text Independent Writer Recognition Using Redundant Writing Patterns with Contour-Based Orientation and Curvature Features. – Pattern Recognition, Vol. **43**, 2010, No 11, pp. 3853-3865.
39. H a s s a ï n e, A., S. A l-M a a d e e d, A. B o u r i d a n e. A Set of Geometrical Features for Writer Identification. – Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. **7667 LNCS**, 2012, pp. 584-591.
40. J i a n g, N. Research on Gradient Local Binary Patterns Method for Human Detection. 2013, No February .
41. D a l a l, N., B. T r i g g s, N. D a l a l, B. T r i g g s. Histograms of Oriented Gradients for Human Detection. – IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886-893.
42. Y i l m a z, M. B., B. Y a n i k o g l u, C. T i r k a z, A. K h o l m a t o v. Offline Signature Verification Using Classifier Combination of HOG and LBP Features. – In: 2011 International Joint Conference on Biometrics, (IJCB'2011), 2011, No October 2015 .
43. D é n i z, O., G. B u e n o, J. S a l i d o, F. D e T o r r e. Face Recognition Using Histograms of Oriented Gradients.pdf. Vol. **32**, 2011, pp. 1598-1603.
44. B a i, L., Y. L i, M. H u i. Face Recognition Based on Wavelet Kernel Non-Negative Matrix Factorization. – Cybernetics and Information Technologies, Vol. **14**, 2014, No 3, pp. 37-45.
45. L o w e, D. G. Distinctive Image Features from Scale-Invariant Keypoints. – International Journal of Computer Vision, Vol. **60**, 2004, No 2, pp. 1-28.
46. U l u s o y, I., C. B i s h o p. Comparison of Generative and Discriminative Techniques for Object Detection and Classification. – Toward Category-Level Object Recognition, 2006, pp. 173-195.
47. K u m a r, R., S. K. S r i v a s t a v a. Machine Learning: A Review on Binary Classification. – International Journal of Computer Applications, Vol. **160**, 2017, No 7, pp. 11-15.
48. S t r e c h t, P., L. C r u z, C. S o a r e s, J. M e n d e s-M o r e i r a, R. A b r e u. PER-08: A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. – Portugal, 2015, pp. 392-395.
49. D o n g, H., G. J i a n. Parameter Selection of a Support Vector Machine, Based on a Chaotic Particle Swarm Optimization Algorithm. – Cybernetics and Information Technologies, Vol. **15**, 2015, No 3, pp. 140-149.
50. H u a n r u i, H. New Mixed Kernel Functions of SVM Used in Pattern Recognition. – Cybernetics and Information Technologies, Vol. **16**, 2016, No 5, pp. 5-14.
51. S c h ̈o l k o p f, B., K.-K. S u n g, C. J. C. B u r g e s, F. G i r o s i, P. N i y o g i, T. P o g g i o. Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifier. Vol. **45**, 1997, No 11, pp. 2758-2765.

52. I m a n d o u s t, S. B., M. B o l a n d r a f t a r. Application of k-Nearest Neighbor (kNN) Approach for Predicting Economic Events : Theoretical Background. – Int. Journal of Engineering Research and Applications, Vol. **3**, 2013, No 5, pp. 605-610.
53. C u t l e r, A. Random Forests for Survival, Regression, and Classification. – Utah State University. Ovronnaz, Switzerland, 2010, pp. 1-129.
54. K i s h o r, D. R., N. B. V e n k a t e s w a r l u. Hybridization of Expectation-Maximization and k-Means Algorithms for Better Clustering Performance. – Cybernetics and Information Technologies, Vol. **16**, 2016, No 2, pp. 16-34.
55. P a r v e z, T. M., S. A. M a h m o u d. Arabic Handwriting Recognition Using Structural and Syntactic Pattern Attributes. – Pattern Recognition, Vol. **46**, 2013, No 1, pp. 141-154.
56. G u p t a, B., P. U t t a r a k h a n d, I. A. R a w a t. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. – International Journal of Computer Applications, Vol. **163**, 2017, No 8, pp. 975-987.
57. H a s s a n a t, A. B., M. A. A b b a d i, A. A. A l h a s a n a t. Solving the Problem of the *k* Parameter in the kNN Classifier Using an Ensemble Learning Approach. – International Journal of Computer Science and Information Security (IJCSIS), Vol. **12**, 2014, No 8, pp. 33-39.
58. G u o, G., H. W a n g, D. B e l l, Y. B i, K. G r e e r. Using kNN Model for Automatic Text Categorization. – Soft Computing, Vol. **10**, 2006, No 5, pp. 423-430.
59. P a t i l, A. B., J. S h a i k h. OTSU Thresholding Method for Flower Image Segmentation. 2016, pp. 1-6.
60. A l-M a'a d e e d, S., D. E l l i m a n, C. A. H i g g i n s. A Data Base for Arabic Handwritten Text Recognition Research. – In: Proc. of International Workshop on Frontiers in Handwriting Recognition, IWFHR, Vol. **1**, 2002, No 1, pp. 485-489.