CHAPTER-1

INTRODUCTION

Pattern Recognition is the scientific discipline that enables the learning progression. It is the process of automated recognition of regularities and patterns using machine learning techniques. It is used in cognate fields such as document analysis, image processing, speech recognition, text recognition, bioinformatics, military, remote sensing, etc. Biometrics encloses the science of measuring human body characteristics and authorizing the user based on biometric modalities. Physiological and behavioral biometrics identifiers are the two kinds of biometric traits. Based on the pattern recognition techniques and biometrics, document analysis and recognition has been a very appealing and exigent area of research. The proposed work carried out in this thesis addresses the novel and distinct concept of gender classification and writer identification based on the handwritten text in the Gurumukhi script. Gurumukhi script is the script used to write the Punjabi language and is the official script of Punjab. Applications of gender classification and writer identification systems are exigent, awe-inspiring, and appealing such as forensic investigations, forgery detection, autopsy determination, questioned documents, analyzing indented handwriting etc.

1.1 BACKGROUND AND MOTIVATION

Pattern recognition is the process of recognizing patterns, classifying the data and making analytical decisions based on the statistical and structural information extracted from the data (Samsuryadi *et al.*, 2021). A pattern can be defined as regularity or repeated arrangement in the digital world which can be visualized physically or mathematically and recognition is defined as exploiting knowledge and skills based on the training patterns (Guha *et al.*, 2019; Asht and Dass, 2012). The pattern may be anything like a signature, handwritten text, weather images, plant leaves, medical reports like ECG, brain print, etc [w1]. By using numerous approaches such as statistical, structural, template matching, fuzzy-based, hybrid model, and computational engineering techniques, novel, stimulating and pioneering applications can be developed for recognizing patterns in the document.

Document analysis and recognition (DAR) is one of the stimulating applications of pattern recognition. It is a constructive paradigm of artificial intelligence, machine learning, and pattern recognition that explores the ability of a machine to analyze and interpret the data [w2], extracting hidden patterns and affinities reside in the characters, words, handwritten texts, signatures, documents, etc. Interesting and challenging applications based on the handwriting are signature identification, gender classification, writer, age, handedness (left or right) and nationality, personality, mood and stress identification (Marinai, 2008). In the present scenario, there is no disbelief that the computers are smart and intelligent for reading, processing documents, and yielding an electronic reproduction, but an efficient process of extracting the features and classification is still a necessity and peculiar move to explore.

Document Analysis System (DAS) supports a computer system in encoding and perceiving online and offline handwritten documents for processing, extracting information, hidden linkages, and the associations in the large volume of data sets (Margner *et al.*, 2018). An increase in deceit in the handwritten content has been motivating the researchers to devise an automatic system to verify the deceptive behavior with ethical and security concerns. In the current scenario, there is no disbelief that the computers are smart and intelligent for reading, processing documents, and yielding an electronic reproduction, but an efficient process of extracting the features and classification to identify the hidden patterns from the offline handwritten text is still a stipulation progress.

The proposed title of the research is the "***Development of gender classification and writer identification systems for offline handwritten Gurumukhi text*",** deals with extracting, selecting and compiling hidden patterns and distinct informative features from the offline handwritten text in order to classify the gender with the name of the writer in the Gurumukhi script.

Automatic gender classification and writer identification based on the handwritten text are interesting, challenging, and stimulating accomplishments as handwriting always carries rich, unique, robust, and hidden information concerning male and female social perspectives. On the basis of handwriting in both Indic and non-Indic scripts, researchers have shown significant interest in developing the

systems for writer identification and gender classification and undoubtedly achieving surprising results.

1.2 BIOMETRIC SYSTEM

Biometric science is the branch of science that deals with life or body measurements by using physical and behavioral human characteristics (Dargan and Kumar, 2020). It is the statistical analysis of an individual for identification, authentication, investigation, surveillance, and security via biometric modalities (Kloppenburg and Ploeg, 2018). Biometric systems, also called Identity Verification System exploits biometric features and implement authentication techniques to validate an individual for handling security and safety issues in the digital and technological world [w3]. As these traits are always diverse and exclusive so these can be accommodated in the technological, operational, and definitional predictions of an individual. The general framework of biometric system consisting of the collection of samples based on the biometric trait, generating corpus, implementing Pre-processing, feature extraction techniques, classification, dimensionality reduction and hybridization of both feature extraction and classification methods.

Biometric traits also called as modalities or identifiers are of two types, namely, physiological and behavioral biometric traits (Bouchaffra and Amira, 2008). Biometric systems that are based on a single trait or modality are called unimodal systems. A multimodal system is a system that is based on the multiple biometric traits (Vijay and Indumathi, 2021; Joshi and Kumar, 2016). The performance of systems is greatly affected by noise and sensitivity of data and the challenges like high security, less accurate results, poor recognition, and robustness against spoofing attacks.

In comparison to unimodal systems, multimodal systems are generally more secure from the spoofing attacks and are highly reliable and robust. Dhieb *et al.* (2020) presented the security, reliability, and robustness in dynamic environmental conditions (Bhalla, 2020; Chapran, 2006; Nguyen *et al.*, 2018). Applications of biometric recognition systems include forensic investigations, driver authentication, access control, human-computer interaction, criminal identification, identification of Aadhar card, military access control, personal identification, login authentication,

medical diagnosis, etc. Figure 1.1 shows applications based on the physiological and behavioural biometric trait.

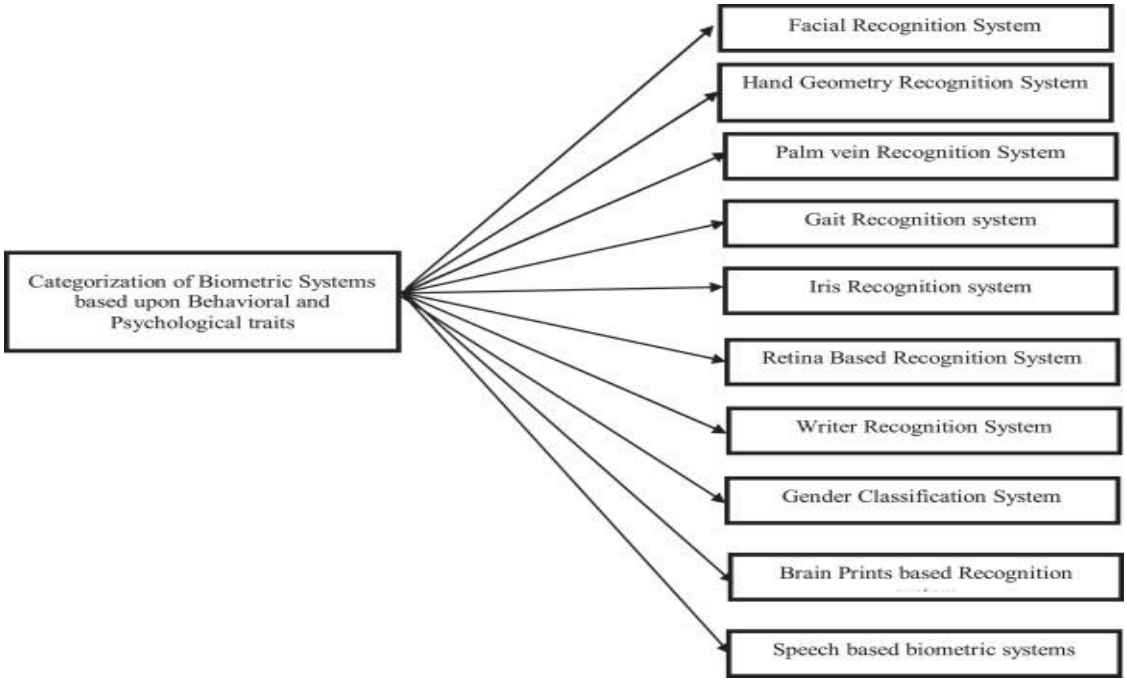


Figure 1.1. Biometric based Recognition Systems

1.2.1 Physiological and Behavioral Biometric Traits

Based on the physiological and behavioral biometric traits, numerous novel developments have been evolving successfully. Faundez-Zanuy *et al.* (2020) presented the concept of behavioral and physiological biometric traits focusing on various application domains. Physiology is defined as the characteristics of the body and thus varies from one body to another body. It refers to static physical features and related to the shape of the body such as face, retina, iris, fingerprint, hand geometry, palm print, etc. (Sabhanayagam *et al.*, 2018). Physiological biometric traits are permanent, unique and inexpensive to collect and are hardly subject to change due to aging (Alsaadi, 2015). It is the characteristic of the body that varies from person to person. Once the physical features are revealed, they can be reused multiple times by the fraudsters. Numerous physiological biometric authentication techniques have been exploited to build a secure system based on such modalities. Jain *et al.* (2016) presented introductory concepts on biometrics with applications. Figure 1.2 shows some physiological biometric traits.

Figure 1.2. Physiological Biometric Traits

Behavior biometric is the branch of measurement that is based on behavioral characteristics such as handwriting, mouse motion, key stroke dynamics, navigation pattern, gait analysis, and so on. It refers to any pattern of behavior that is specific to the user. Bouadjenek *et al.* (2016) presented a biometric-based system based on handwriting, which is dynamic and variable over time, and hence more security lies with behavioral traits (Handa *et al.*, 2019). Figure 1.3 shows behavioral biometric traits. Behavioral biometric traits are more secure and can increase the level of confidence. Fallah and Khotanlou (2016) presented a survey on human personality parameters such as behaviour, anger, personality, happiness, mood, and stress based on the handwriting trait.

Based on both physiological and biometric characteristics, numerous novel applications have been developed for security and surveillance perspectives. The choice of the biometric trait always depends upon the availability of data samples, nature of the application, level of confidence, enhancing security issues and kind of tolerance, etc. Also, comparing two-dimensional (2D) and three-dimensional (3D) views of traits, biometric technologies are more secure and produce relatively higher accuracy with the 3D vision of modalities. From the survey findings, it has been pointed out that in comparison to physiological biometric traits such as the face, fingerprint, iris, retina, behavioral biometric traits such as handwriting, strokes, and signature, etc., have been less explored and considered for the development of novel systems.

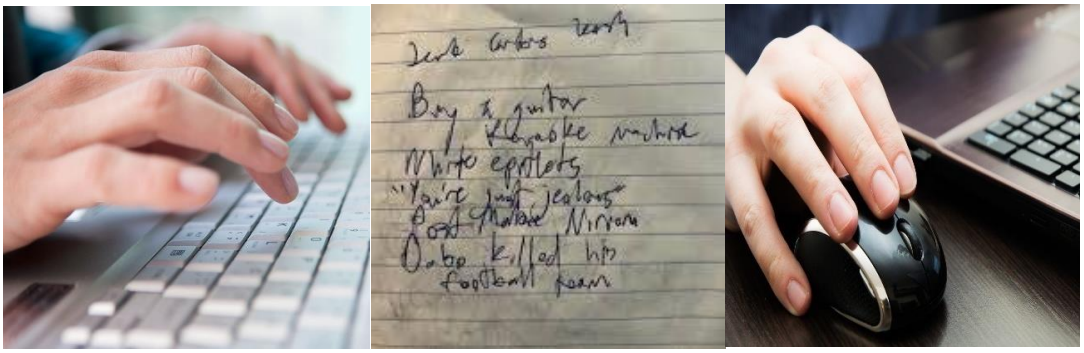


Figure 1.3. Behavioral Biometric Traits

1.2.2 Handwriting: A Behavioral Biometric Trait

The handwriting of an individual is a stable, robust, and strong behavioral biometric modality that shows the state of the heart and mind and gives a glimpse of our personality. It is believed that

*"**If eyes are a mirror to your soul, then handwriting must be a window**"*

Handwriting is an art that deals with the sentiments, personality, attitude, nature, intelligence, and brain. It may vary because of geographical location, cultural, social, and temporal aspects, such as age, illness, writing position, lack of concentration, and physical disturbance during the writing period [w4].

An individual cannot exactly replicate his/her own writing even after a few seconds and the behavior is known as call variation, i.e., a natural deviation occurred in an individual's writings (Khushboo *et al.,* 2020; Ghosh *et al.,* 2020a). Successful application has been accomplished using handwriting recognition tools and methodologies based on syntactical, structural and statistical approaches, as presented by Nguyen *et al.* (2018).

It is believed that indistinguishable twins who mostly share appearances and hereditary qualities do not have the same handwriting, and are called an inter-class variation [w6]. Handwriting-based applications are character, word, and postal name recognition, gender classification, writer's name, age, stress, personality, handedness (left or right), signature identification, and nationality identification. Handwriting is a strong and robust physical proof which is subsequently useful for forensic applications and autopsy determination.

"Handwriting is a spiritual designing, even though it appears by means of a material instrument." (Euclid- A Greek Mathematician)

Handwriting is the art of writing by hand and can be generated offline and online. Sometimes the handwriting of an individual can be changed due to many reasons such as illness, age, mood, etc. This property is called Dysgraphia. Each person has their own unique style of handwriting, whether it is a textual form or a personal signature. Even identical twins who share appearances and genetics have different handwriting. Yang *et al.* (2020) presented a study on the handwriting posture prediction based on unsupervised learning and convolution neural network and achieved an accuracy of 93.3%.

1.2.2.1 Features of Handwriting

Actually, handwriting was originated earlier to expand human memory and to facilitate communication [w5]. It consists of many hidden attributes such as line quality, word, and letter spacing, size consistency, pen lifts, connecting strokes, cursive and printed letters, pen pressure, baseline habits and slant, diacritic placement through which one can easily trace identity and authenticate a person (Miller *et al.*, 2017). Here are a few more handwriting features that can be analyzed and affect the accuracy rate.

- The specific shape of the character, i.e., roundness, robustness, or sharpness.
- Irregular or regular spacing between characters.
- Skelton and slope of the character.
- The rhythmic repetition of the elements or arrhythmia.
- Pressure on the paper.
- Graphemes and shape.
- The average size of letters and thickness of letters.

Because handwriting is comparatively stable, a change in the handwriting can be analytic of the anxiety and intoxication of the writer.

1.2.2.2 Offline Vs. Online Handwriting

In terms of data acquisition, identifying the author of text can be categorized into two categories named, online and offline handwriting. The concept of identification of an

individual based on offline handwriting is more robust and encouraging as compared to online handwriting due to the fact that offline handwriting accommodates many hidden features and variations (Geetha *et al*., 2021; Zhang *et al.,* 2017; Tan *et al.,* 2009). Figure 1.4 shows a view of offline and online handwriting.

On-line handwriting means a collection of handwriting features at the same time when it is generated. The writer usually creates handwriting via a mouse or an electronic pen, and the output contains both sequential and spatial information (Kaur and Kumar, 2021; Jindal *et al.,* 2021; Singh *et al*., 2019). Therefore, the user must know the operational knowledge of the device. Online data supply the dynamic and trajectory information to the system for the writer's identification. As it is dynamic in nature, it can be represented as a function of time. Also, in the case of overwriting, the presence of strobes helps in the unique identification.

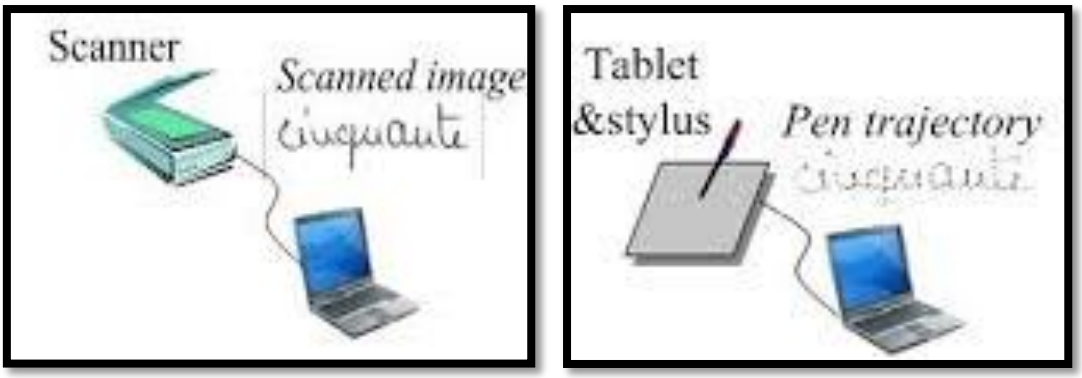
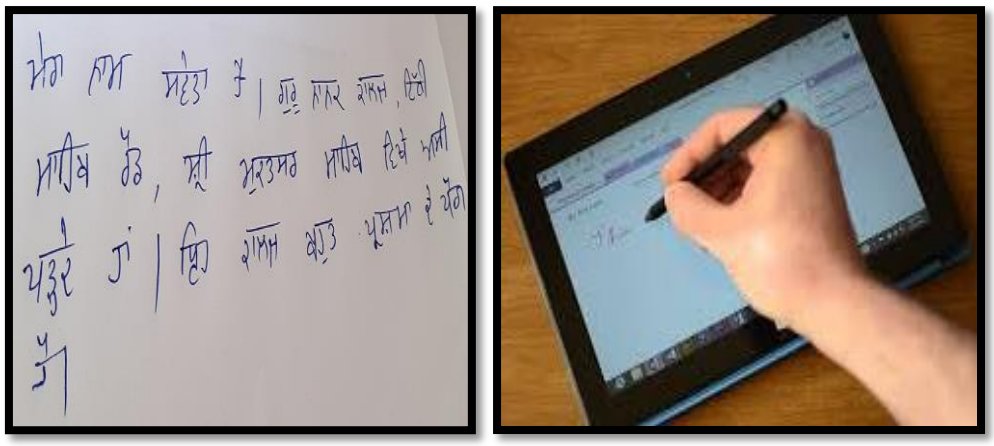
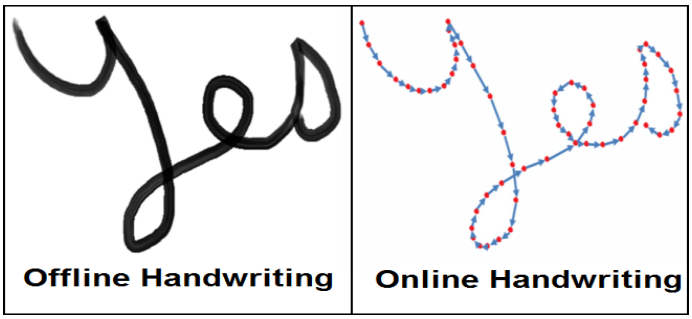


Figure 1.4. Hardware requirement for Offline vs Online Handwriting

Offline data refer to a static image of a handwritten document. Offline handwriting is captured by a scanner or camera and is stored as an image. Due to the lack of sequential and dynamic information at the time of writing, offline writer identification is difficult in comparison with online identification. Also, offline data do not supply timely and trajectory information to the system and hence identification is a difficult process (Priya *et al.*, 2016). Online and offline writings are described below in Figure 1.5 (a), (b) and (c).



(a) (b)



(c)

Figure 1.5. Sample of (a) Offline handwriting (b) Online handwriting (c) Offline and Online handwriting

1.2.2.3 Masculine vs Feminine Handwriting

There has always been a close connection between the brain and handwriting. The handwriting of females is generally considered as more legible and better than male handwriting because of social stereotypes, physical features, and prenatal hormones (Yang *et al.*, 2020; Beech *et al.*, 2005). As women are more patient and more prone to compromise, therefore they have round, loopier, and curvy handwriting. Females are largely dependent on the decisions and approval, and hence handwriting includes large and clean letters.

In general, women are more observant, keen, dedicated, attentive, hard-working, assiduous, therefore letters are clearer [w6]. On the other hand, the handwriting of men may depend on their nature of apathy and suppressive behavior, they are temperament, less patient, hurry, and less prone to tell their actions to others. Table 1.1 presents the differences between male and female handwriting.

Table 1.1. Characteristics of Male and Female handwriting

Male Handwriting	Female Handwriting
Messier	More Rounded
Small in size	Tidier and Neater
More Slanted	Circular
More angular	Curvier
Unorganized	Organized
Rushed	Bigger
More Straight Lines	Letters big and spaced apart
More Scribbling	Bubbly
More Italic	Regular

1.2.2.4 Handwriting Identification Vs Handwriting Verification

Handwriting Identification and verification systems deal with determining the special nature of writing styles of an author and handwriting interpretation are the processes to filter out the changes to determine the message [w5]. Handwriting identification is the process of identifying the authorship of handwritten text and handwriting verification is to evaluate the confidence that the given handwriting sample has been written by the same writer or a different writer. Bal and Saha (2016) presented a methodology for handwriting recognition strategies using writing pressure detection, baseline recognition, and segmentation through skew recognition, writing pressure, and segmentation for cursive handwriting. Daraee *et al.* (2021) proposed a model to build handwriting keyword spotting using the deep model and achieved great success in the novel application. Rosyda and Purboyo (2018) systematically presented in detail, a state-of-the-art work on the handwriting recognition methods. Sueiras *et al.* (2018) experienced a novel idea of sequence-to-sequence NN for offline handwriting recognition and achieved promising results.

1.2.2.5 Significance of Handwriting in proposed Study

Handwriting is defined as something written by hand and is broadly a distinct perspective from the print version (Altwaijry and Turaiki, 2021; Lincy and Gayathri, 2021). It encompasses many versatile characteristics of an individual and is unique to one another. Based on such incredible property, it can be used as a major tool for

developing many challenging and stipulating applications with both Indic and non-Indic scripts. Pressure on pen and on paper, nature of strokes, speed of handwriting are the key parameters used for the identification and verification process. Yang *et al.* (2019) worked on the handwriting posture determination using convolution kernel models and achieved 93.3% accuracy results. Mekhaznia *et al.* (2021) presented the effect of handwriting on personnel identification.

Applications of Handwriting

With handwriting as a rigid and secure biometric trait, many stipulating and triggering applications such as writer, gender, left or right-handedness, age, personality, stress, handwriting posture can be identified which is very well used in the following applications such as signature identification, forensic investigations, crime detection, forgery detection, font personalization, movement simulation, detecting alterations, analyzing indented handwritings, improving personal relationships, gender classification, nationality, age writer identification, autopsy determination, career guidance and personnel selection, human-computer interaction and stress prediction.

1.3 GENDER CLASSIFICATION SYSTEM

Gender Classification System (GCS) is also known as a two-class system or a binary problem. In today's digital scenario, it reflects novel insights into the evolving researches of machine learning and pattern recognition. Gender is a criticality and its classification is a two-class problem which classifies male and female based on the physiological and behavioral biometric traits such as handwriting, facial expressions, speech, etc. Gender classification is necessary because dividing the population into subcategories is interesting and important for the development of many applications based on the features extracted from the human body, such as facial expressions, behavior, etc., various successful approaches have already been in practice for gender classification (Cordasco *et al.*, 2020; Gattal *et al.*, 2020; Mekhaznia *et al.*, 2021; Rahmanian *et al.*, 2021; Bi *et al*., 2021). But on the basis of handwriting, it is a motivating, alluring, and pragmatic move to implement. Based on the facial expressions and other physiological traits such as gait information, the development of the gender classification system has been successfully recognized, but on the basis of the handwriting trait, it is still a novel, motivating, alluring, and pragmatic move to

implement. Figure 1.6 shows handwriting samples collected from six males and six female writers.

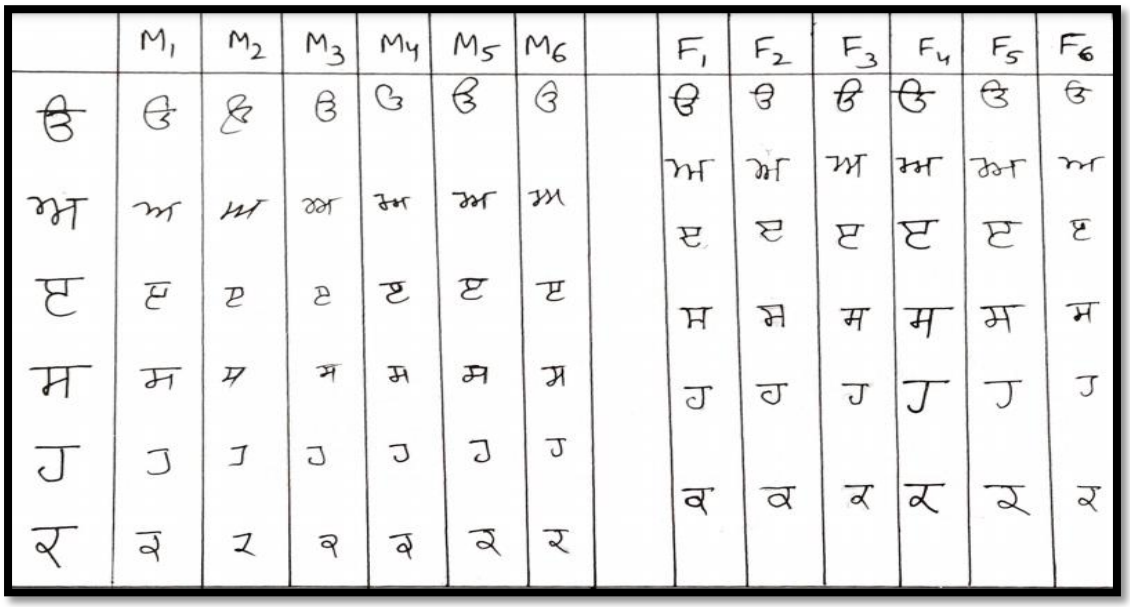
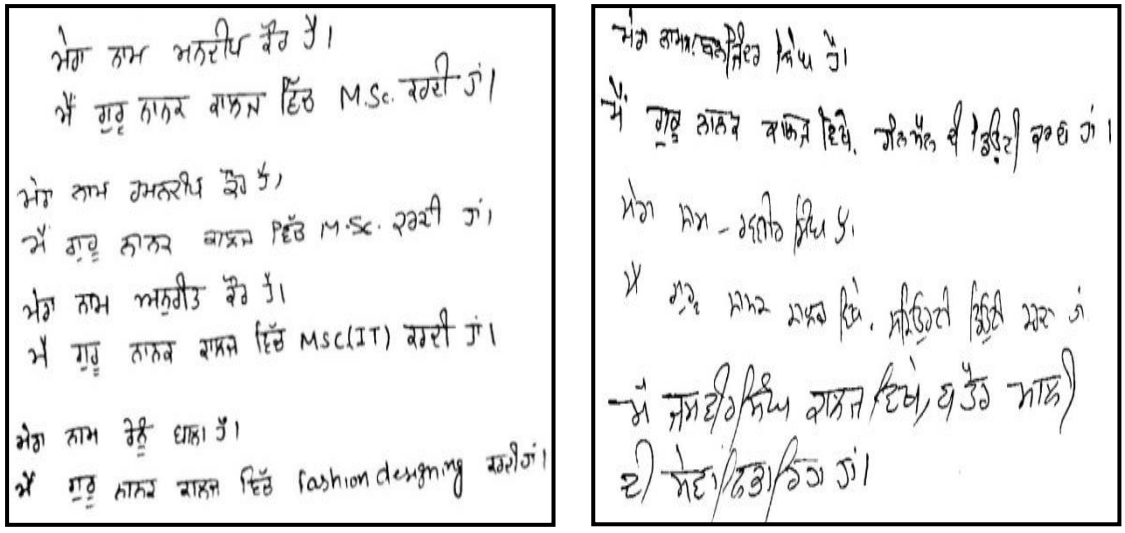


Figure 1.6. Male and Female handwriting

Individuals who exist outside these groups fall under the category called gender-queer, third gender, or Transgender. Table 1.1 shows the properties and the characteristics of male and female handwriting.

Gender is the range of characteristics that help in differentiating the masculine and feminine characteristics of an individual. Broadly gender is considered as a binary, i.e., two genders (male or female), and individuals who exist outside these groups fall under the umbrella called gender-queer or non-binary, often called the third gender or transgender. Gender classification is a two-class problem that classifies males and females based on physiological and behavioral biometric traits such as speech, facial expressions, gait information, retina and handwriting, and so on. The difference between male and female handwriting can be speculated from the visual appearances of the handwriting as the handwriting of male is generally messy, small, more slanted, more angular, unorganized, rushed, hurried, italic, and scribbling whereas female handwriting is more rounded, decorative, homogeneous, tidier, neater, curvier, organized, bigger, bubbly and regular. Figure 1.7 shows the samples of the handwriting of males and females in Gurumukhi script.



(a) (b)

Figure 1.7. Samples of (a) females (b) males handwriting in Gurumukhi script

1.4 WRITER IDENTIFICATION SYSTEM

Writer identification is an endeavor to identify the authorship of the given document or handwritten text. It is such a paradigm of machine learning which deals with the identification of an individual based upon distinct identifiers such as characters, words, lines, and signatures in both Indic and non-Indic scripts (Ahmed *et al.*, 2017; Kumar *et al.*, 2017; Bangarimath *et al.*, 2018; Adak *et al.*, 2019; Ahmed *et al.*, 2019; Durou *et al.*, 2019; Bennour*et al.*, 2019; Chahi *et al.*, 2019; Girdher and Sharma, 2020; Javidi and Jampour, 2020; Litifu *et al.*, 2021; Mohammed and Ahmed 2021; Sharma and Kaushik, 2020; Abbas *et al.*, 2021; Chen *et al.*, 2021; BabaAli, 2021; He and Schomaker, 2021). It is a big confrontation as the working of a brain is difficult to parameterize and predict.

Writer identification is an efficacious and serviceable move in the area of pattern recognition, document analysis, and machine learning which reflects the novel and upbringing perceptions in handwriting research. The process is also linked with the writing styles, brain, feelings, perception, and behavior of a person. It works on the principle of one-to-many searches in a dataset by taking a sample of a known author and matching it with a training data set. Figure 1.8 shows the writing style of male and female writers.

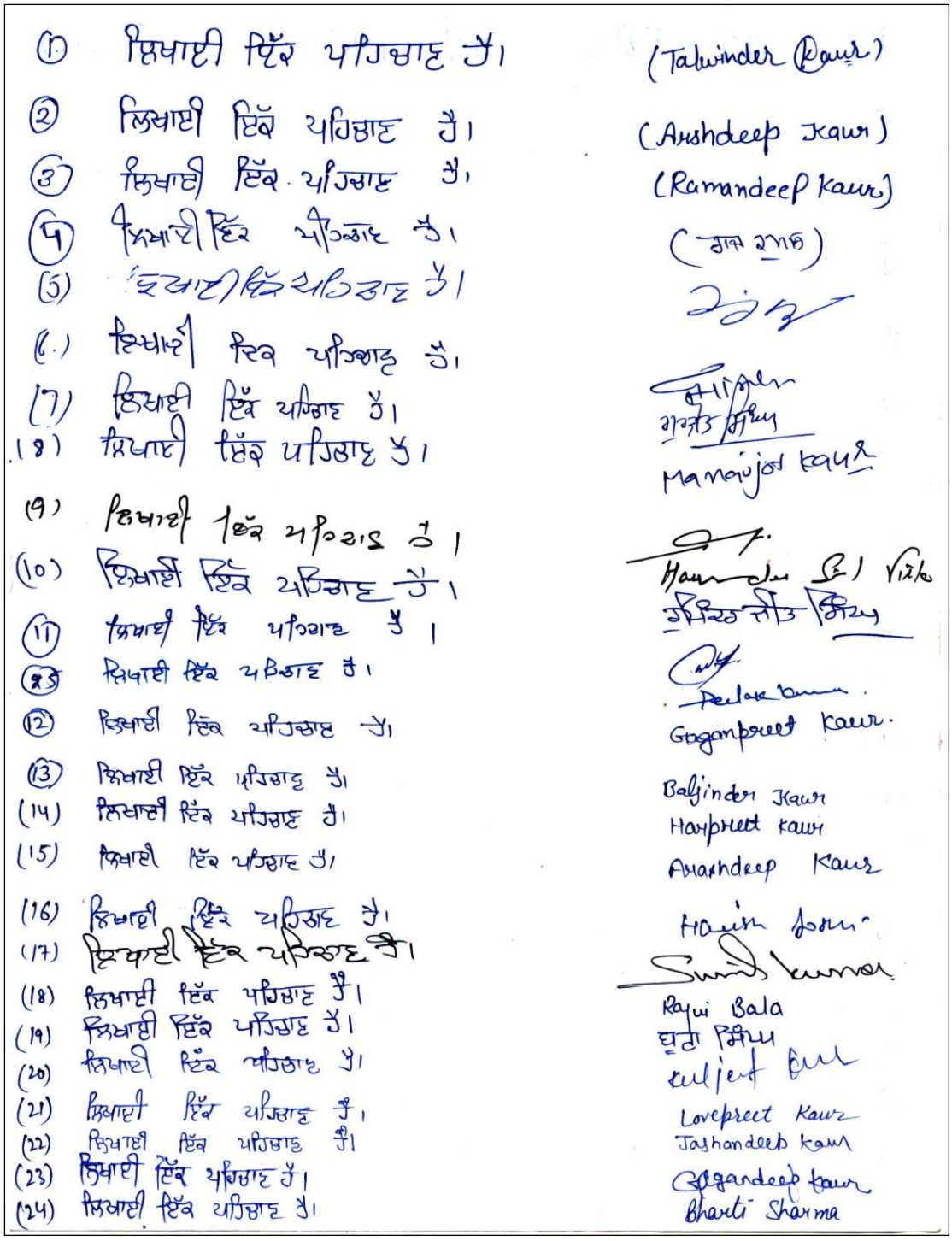


Figure 1.8: Writing pattern of male and female writers

Writing is usually done with two types of forces:

- **Conscious Force:** A conscious force during handwriting is a force which is managed by a person's own power and is restricted. It is a concept of deep writing in which the writer knows his full potential and awareness regarding the written text. It is a force that comes with self-realization and self-expression.

- **Unconscious Force**: A kind of force which is not in the hands of an individual, but surely influences the feelings, behavior, and the conclusion.

The writer identification system is a system of determining the author or writer of the given text after matching the sample with the training data set. Dargan and Kumar (2019) presented a deep and comprehensive state-of-the-art work on writer identification systems with Indic and non-Indic scripts. It is really a challenging task as the writing of an individual has intrinsic characteristics. Because of ample advancements in technology, writer identification is widely used for forensic investigations, document analysis of criminal or suspected cases, human-computer interaction, forgery detection, autopsy determination, signature identification, and verification in the bank, etc. This can only be achieved by using one-to-many mapping in a database using samples of the known authorship. Dargan *et al.* (2020) presented a writer identification system based on offline handwritten samples of Devanagari script and achieved a great promising rate.

Figure 1.9 describes the differences in writing the word "ਕੰਪਿਊਟਰ" from male and female writers. It shows the versatility of writing the same word with differences. These are written by male and female writers and their signed are shown in the figure.

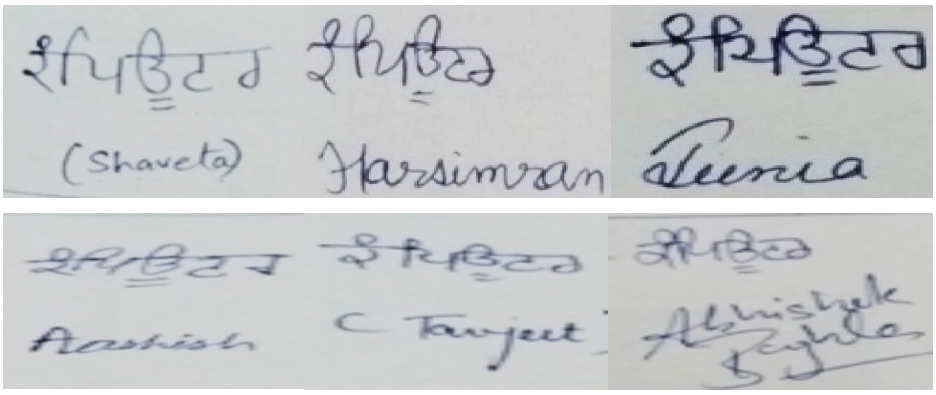


Figure 1.9. Word written by male and female writers in Gurumukhi script

1.5 CHALLENGES AND ISSUES

During the development of writer identification and gender classification, challenges and issues that may produce hindrances in the process of development are:

- Diverse style of handwriting makes the task difficult.
- The nature of the script also affects the accuracy rate.
- Selection and implementation of feature extraction methods can also affect the accuracy rates.
- Reduction of extraneous variables can help in increasing the performance.
- Pre-processing on the rough dataset plays an incredible role in boosting the accuracy.
- Degraded data, quality of paper, noisy data must be identified and discarded from the data sets.
- Proper maintenance of datasets also helps in getting successful results.
- Because writing is about emotions, feelings, age, mood and illness, it creates significant challenges during the matching and identification process.

1.6 STAGES OF GENDER CLASSIFICATION AND WRITER IDENTIFICATION SYSTEM

The development of gender classification and writer Identification system is based upon phased approach with an efficient framework in a constrained manner. The output of one phase is fed as an input to the next phase. Beginning with the data collection phase, then Pre-processing stage, segmentation part, feature extraction, and dimensionality reduction phase followed by the classification phase. Hybridization of feature values and hybridization of classification algorithms also boosts up the accuracy rates and strengthens the output of classifiers too. So here in the next subsections, we present details of phases to be carried out in the development process.

1.6.1 Data Collection

This stage is a primary stage in which samples of offline handwritten data are collected from the writers on a paper. It must be taken care that the quality of paper, categorization of individuals such as age, state, rural, urban, gender, quality of paper, the color of pen, nature of pen and number of samples, etc. must be considered at the commencing stage in case of offline handwritten data collection. For the qualitative collection of the data, some protocols must be set up initially and must be followed strictly.

1.6.2 Digitization and Pre-processing

After collecting the data on sheets, next comes, Digitization phase. It is the phase where data is scanned, at 300 dots per inch (dpi), which is called the standard value for digitization, which is easy to read and share and considered as the best dpi for scanning. Poor or awful quality samples i.e., degraded samples must be discarded as these can decrease the performance of the system. Pre-processing helps in transforming the data to a machine-compatible form for further processing. It covers cleaning i.e., removing the noise present in the data, and also covers normalization, and thinning of the data. Therefore, the main goal is to create a digitized image and finally to thinned image so that feature extraction and classification would be an easy and efficient task.

1.6.3 Feature Extraction

This phase deals with extracting features from the pre-segmented and preprocessed characters so that the classification phase can be accomplished. The goal is to retrieve informative and non-redundant values that will help in learning and generalization. It is believed by the practitioners that properly optimized feature extraction is the key to the accomplishment of framework building. There are many statistics and structure-based techniques for extracting features. Therefore, based on the curves, strokes, shape, diagonal, zones, edges, one can extract features from the offline handwritten character that will assist in the development of applications.

1.6.4 Dimensionality Reduction

In view of pattern recognition and machine learning-based applications, dimensionality reduction methods are best suited to boost up the accuracy rate by mapping from higher dimensions to lower-dimensional space. The goal is to reduce extraneous variables so that it leads to less effort, less cost, less complexity, and fast processing. There are a number of methods used for dimensionality reduction such as Principal Component Analysis (PCA), missing value ratio, low variance filter, Linear Discriminant Analysis (LDA), Auto-encoder (AE), and so on. The target is to maximize the variance of the data in low dimensional space. For the proposed work, the PCA technique has been implemented in which eigenvectors with maximum eigenvalues are calculated to build up the feature space. PCA is quite similar to factor

analysis. Examples of non-linear dimension reduction methods are PCA, Kohonen features, and Multidimensional Scaling (MDS).

1.6.5 Classification

In general, classification is the process of assigning labels to the data samples based on the extraction of features. This is the phase that deals with the classification process, i.e. to classify the class, i.e. male or female, two classes for gender classification, and output the name of a specific writer among 200 classes for writer identification. There are many parametric and non-parametric classification techniques that have been efficiently working on handwriting traits. To overcome the limitations of these classifiers, there is another simulating approach called hybridization of classification techniques. Examples of these classification methods are a nearest neighbor, decision tree, support vector machine (SVM), random forest, multilayer perceptron (MLP), artificial neural network (ANN), convolution neural network (CNN) and recurrent neural network (RNN), etc.

1.6.6 Hybridization of Feature Extraction and Classification Techniques

This phase is concerned with the hybridization of feature extraction techniques so that a huge feature vector of values can be generated for exploring, improving and enhancing the classification. Similarly, the goal of hybridization of classification techniques is to hybridize the outputs of classification techniques i.e., to strengthen the result, based on efficient methodology such as majority voting scheme. Elbarawy and Ghonaim (2021) presented hybridized model using a convolution neural network and multiclass support vector machine for Arabic and English scripts using the KHATT dataset and received 99.6% promising accuracy. Chaudhuri *et al.* (2009) presented a multiscale version of hybridization of parametric and non-parametric classifiers based on the kernel and model function.

1.7 SALIENT TERMINOLOGIES

1.7.1 Writer Identification Vs Writer Verification

Writer Identification is the process of identifying the writer of the handwritten text from the writers stored in the data sets. The developments on writer identification systems based on the Indic and non-Indic scripts have been in progress successfully

by the handwriting-based researchers. For the Indic scripts, there have been many new directions for researchers to build up novel applications. Writer identification system has been accepted in many Indic and non-Indic scripts, but to the best of our observations, the development of a writer identification system for Gurumukhi script with a sufficiently large dataset has not been satisfactorily and successfully recognized as compared to the state-of-the-art work whereas writer verification is the verification of the writer based on handwriting by matching from the training dataset. Nakamura and Kidode (2006) presented online writer verification based on the frequency distribution of deviations and yielded successful results for online handwriting.

Bensefia and Paquet (2016) proposed a writer verification system with an 87.0% accuracy rate using a single word based on the Fisher Wagner algorithm and Levenshtein distance, segmentation module, and graphemes, etc. that was implemented on 100 writers of the IAM dataset. Aubin *et al.* (2018) proposed a novel system of writer verification from the strokes analysis, pressure applied, the width of stroke using Discrete Cosine Transform (DCT), Principal Component Analysis (PCA), Support Vector Machine (SVM) and yielded successful rates. Dargan *et al.* (2020) presented a novel experiment for writer identification system for the Devanagari script and yielded successful and promising results.

1.7.2 Text Dependent vs Text Independent

In text-dependent approach, text in the questioned document should always be similar to the text in the training process. For implementing such systems, statistical techniques are required. Whereas in text independent systems, text in the questioned documents should not necessary be same with the text in the training process (Wang, 2019). Here, structural approaches are required to extract the features for identification. Dhandra and Vijaylaksmi (2014) proposed text-independent approach for writer identification in Kannada script and achieved great success. Dhandra and Vijaylaksmi (2015) developed a text-dependent system for writer identification based on multi-resolution features and structural features and nearest neighbor classifier, by achieving 93.25%. Gharahbagh and Yaghmaee (2018) presented a study on the Persian script for writer identification.

1.7.3 Single script vs Multi scripts

If the development of a system is based on one script, both for the training and testing process then the system is said to be a single script system on the other hand if the development of the system is based on multiple scripts or dataset for training and testing process, then the system is said to be multiscript system (Vijay and Indumathi, 2021). The idea behind the concept is that the art or style of handwriting remains the same as a writer independent of the script used. Rahmanian *et al.* (2021) presented a multiscript system based on IAM and KHATT scripts and executing CNN based model achieved 84.0% for gender classification and 99.15% for handedness. Gattal *et al.* (2020) presented a single script-based gender classification system using Cloud of Line Distribution (COLD) and Hinge features and yielded 64.40%. Bertolini *et al.* (2016) presented writer identification based on Arabic and English script using local binary pattern and local phase quantization and achieved successful results.

1.8 APPROACHES USED FOR GENDER CLASSIFICATION AND WRITER IDENTIFICATION SYSTEM

Statistical and structural approaches have been used for the pattern recognition-based gender classification and writer identification applications. These approaches are based on the formalism and mathematical tools, out of which some techniques are more powerful in terms of computation and others, are having limited results. So, these are complementary to each other, but the strengths and benefits of these methods are helpful in the development of a novel and unbelievable applications. Parvez and Mahmoud (2013) worked on both syntactic and statistical methods for handwriting recognition on Arabic script.

- ✓ **Statistical Approach**: The statistical approach deals with the development of the proposed application based on the features extracted from the samples (Bunke and Riesen, 2012). Based on the probabilistic nature, patterns are extracted forming a unique feature space and hence the formation of the decision boundary [w7]. In the statistical identification of the writer, based on the decision and probability theory, numerous simple descriptive statistics and complex transformations are utilized in the statistical approach for the identification process. Also based on the probability density distribution, numerous parametric and non-parametric methods of classification can be

applied. E.g., of statistical-based applications are health sciences, demography, operation research, and of course, machine learning [w8]. Transforming methods, filtering methods are examples for performing the statistical evaluation.

- ✓ **Structural Approach**: To eliminate the ambiguity in the results because of the presence of inherent relationships in the patterns, it is preferred to use a structural approach. Therefore, in complex pattern recognition scenarios, the structural approach is best suited that will use hierarchical patterns and identify morphological relationships existing inside the character or word, e.g., relational description, formal grammar, a decision tree is built based on the absence or presence of a character (Bunke and Riesen, 2012). On the basis of pattern primitives and description language, the development of the system and classification can be achieved [w8]. Based on the structural approach, successful numerous applications are social representations, knowledge sharing, database maintenance, etc.

1.9 APPLICATIONS OF GENDER CLASSIFICATION AND WRITER IDENTIFICATION SYSTEM

The development of gender classification and writer identification systems, will surely impart something novel directions in forensic investigations, security, and surveillance. The writer identification and gender classification system have multiple applications which are described here in this section.

- Forensic Investigations: It is the collection and analysis of all crime-related physical evidence for the solution of the suspected cases. By analyzing handwriting, Morris (2021) questioned documents and texts can be resolved and analyzed thoroughly which is a subsection of forensic analysis. Forensic science is also called criminalities and handwriting is a motor skill that involves neurological, physiological, and sensory perspectives (Harrison *et al.*, 2009).
- Autopsy determination: In autopsy determination, handwriting plays an important role, i.e., to determine the cause of death or in the postmortem examination process (Jeblee *et al.*, 2019) by having availability of some

handwritten material. A handwritten text in itself becomes the main evidence to find the cause of death or suicide (Joshi and Garg, 2015; Chaudhuri *et al.*, 2020).

- Forgery detection: During forgery detection, handwriting samples are taken from the real authors and forgery authors. Forged handwriting samples often wrinkled trace as compared to the original one. Cha and Tappert (2002) presented automatic detection of forgery and achieved 88.0% results. Gideon *et al.* (2018) proposed handwriting signature-based forgery detection using a convolution neural network and achieved promising results. Roy and Bag (2019) presented handwriting based forgery detection in the documents.
- Authentication with financial transactions and property security: Machine recognition of handwriting plays an incredible role in case of any fraud or investigations in financial or security concerns. With samples of handwriting, the machine can be trained and then tested for the questioned document (Chambers *et al.*, 2015).
- Library archival**:** Archives of the library can be identified using handwriting samples of the readers and also to verify the authorship of the archives with the help of a writer identification system.
- Indexing, analysis, retrieving historical documents can be easily traced if one can have samples of handwriting. The author of the historical document can easily be identified with promising traits using handwriting biometric traits.

1.10 OVERVIEW OF GURUMUKHI SCRIPT

Gurumukhi is the name of the script which is used to write mainly Punjabi and secondarily Sindhi language. It is the standardized script used by the second Guru of Sikhs, Guru Angad Dev Ji. Gurumukhi means the mouth of the Guru [w9]. Gurumukhi was developed to be a very precise phonetic script. It is the official script of Punjab. The primary scripture of Sikhism, “Sri Guru Granth Sahib Ji”, is written in Gurumukhi script. Gurumukhi is a relatively new script. It is less than 500 years old.

In the west, it is called the Shahmukhi script (used by Punjabi Muslims) and in the east, it is called the Gurumukhi script (used by Sikhs in Punjab). It is a script that originated from the Brahmi script named Aryan script which came between the 8th

and 6th century BC [W10]. The word Gurumukhi consists of two words, Guru and mukh, which means the script used to record the sayings from the mouth of the Gurus, (Aggarwal and Singh, 2015). By understanding and learning Gurumukhi, the holy utterances of Guru Granth Sahib ji can only be understood as shown in Figure 1.10 [w11]. Gurumukhi script was discovered to open the entry of the Shabad Guru and superior consciousness to all persons and begins to generate changes in the physiology [w12].

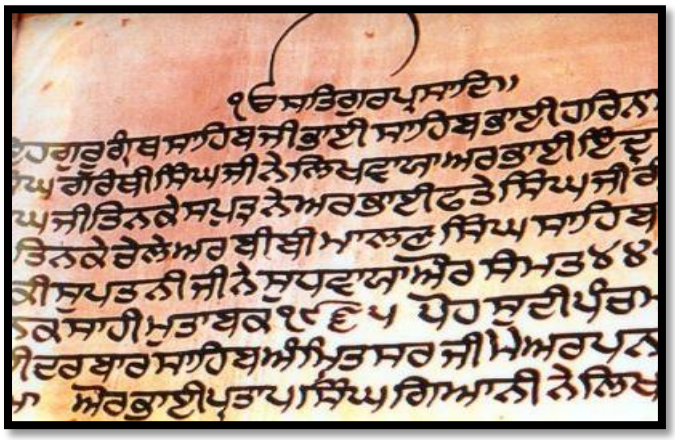


Figure 1.10. Sample of Utterances in the Gurumukhi script

It is the first language of Punjab and the world's 14th widely spoken language. Punjabi is spoken by around 115 million people in the west Punjab of Pakistan and the East Punjab in India. Mainly Gurumukhi is written by the people of Punjab. Gurumukhi script is cursive in nature. The character set consists of a total of fifty-six characters, of which thirty-five are primary characters named "painti", six additional consonants, nine vowel modifiers, three auxiliary signs, and three half subscript characters [w13]. Table 1.2 shows the deep representation of the Gurumukhi character set. Gurumukhi characters and words can be written in three lines, horizontal line at the upper part. These characters are joined by a line called headline and in letters with no vertical inter-character gap. The upper zone holds the area above the headline, the middle zone covers the area below the headline that consists of the consonants and the subparts of vowels are present there. In the lowest zone, there are vowels and half characters that are present in the foot of consonants.

The character set of the Gurumukhi script includes primary and secondary characters [w14]. Primary characters are the 35 characters called "paintee", which comes in the category of consonants and vowel bearers as shown in Table 1.2.

Secondary characters which are twenty-one consisting of additional consonants, vowel modifiers, auxiliary signs, and half characters as shown in Table 1.2.

Table 1.2. Character set of the Gurumukhi script

Sr. No.	Category	Characters	Total
1.	Consonants	ਸ ਹ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਝ ਞ ਟ ਠ ਡ ਢ ਣ ਤ ਥ ਦ ਧ ਨ ਪ ਫ ਬ ਭ ਮ ਯ ਰ ਲ ਵ ੜ	32
2.	Vowel Bearers	ੳ ਅ ੲ	03
3.	Additional Consonants	ਸ਼ ਖ਼ ਗ਼ ਜ਼ ੜ੍ਹ ਫ਼ਲ	06
4.	Vowel Modifier	ੇ ੈ ੋ ੌ ਿ ੀ ੰ ੁ ੂ	09
5.	Auxiliary Signs	ੱ ੵ ◌	03
6.	Half Characters	◌੍ਚ ◌੍ਛ ◌੍ਵ	03
			Total=56

1.11 OBJECTIVES OF THE PROPOSED WORK

1. To propose a gender classification and writer identification framework for offline handwritten Gurumukhi text.
2. To generate a corpus of both male and female writers of the Gurumukhi script by collecting offline handwritten samples.
3. To implement available feature extraction techniques like zoning, diagonal, transition, intersection, curve fitting, and peak extent based features for the developed corpus.
4. To propose a new feature extraction method that will help in gender classification and writer identification.

5. To propose a hybrid classifier based on the exploration of various classifiers like ANN, MLP, K-NN, Decision tree, SVM, and Random forest for gender classification & writer identification.

1.12 MAJOR ASSUMPTIONS

For successfully completing the proposed developments, some basic assumptions have been considered are:

- Working on pre-segmented characters.
- Working on two genders i.e., male and female sample collection.
- Scanning done at 300 dpi (dots per inch) resolution, which is a standard value.
- Characters considered in this work are free of noise.
- Skew detection and correction are not considered here.
- Data considered in this work does not contain any non-text items such as images, figures, etc.
- 80% of data is considered as a training set and the remaining 20% as a testing dataset.

1.13 MAJOR CONTRIBUTIONS AND ACHIEVEMENTS

- The proposed research work helps the handwriting-based research communities by proposing a framework that will classify the gender and identify the writer based on the offline handwritten text in Gurumukhi script.
- The development of gender classification and writer identification systems has been successfully achieved with 200 writers with an accuracy of 94.27% and 91.23% respectively. This has been accomplished with hybridization of feature extraction techniques such as Zoning, Diagonal, Transition and Peak Extent and hybridization of classification techniques such as ANN, MLP, Decision Trees, Random Forest, and Adaptive Boosting.
- Creation of a dataset with 200 writers having 100 females and 100 male writers has been accomplished for gender classification and writer identification systems. Dataset for writer identification system contains total 70,000 Gurumukhi characters (200×35×100) and dataset for gender classification system contains 35,000 Female Gurumukhi characters and 35,000 male Gurumukhi characters.

- Proposal of new feature extraction algorithm based on the hybridization of existing feature extraction techniques has been attained with promising accuracy rate.
- Performance parameters have also evaluated to characterize the quality of the experiment [w14].
- Implemented PCA based dimensionality reduction method for dimensionality reduction with a target to increase the accuracy and reducing CPU elapsed time has been a remarkable achievement.
- Hybridization of classification algorithms using a majority voting scheme has also been exploited to strengthen the classification results for writer identification and gender classification system.
- The hybridization of feature extraction techniques along with the hybridization of classification techniques has been experienced with successful accuracy rate.

1.14 CHAPTER SUMMARY

This chapter covers the introductory part of the gender classification and writer identification system, nature, technology, needs, and application areas of the proposed system. The approaches used for the system and some basic terminologies are also discussed along with the framework and phases. As the development of the systems is based on the behavioral biometric traits, discussion on physiological and behavioral traits, characteristics of handwriting traits, online and offline handwriting, masculine vs feminine handwriting, samples of male and female handwritings are also presented with illustrations in this chapter. The nature of the Gurumukhi script along with the detailed discussion on the character set with various categories is also represented. Research objectives that are to be fixed on the commencing stage are also stated in the first chapter. Major assumptions, major contributions, and achievements during the experimental work are also comprehensively presented. The chapter is concluded with the organization of the thesis.

1.15 ORGANIZATION OF THE THESIS

The main objective of the thesis is the development of gender classification and writer identification systems for offline handwritten Gurumukhi text. Therefore in order to

achieve satisfactory, promising high accuracy results, numerous experiments based on efficient and challenging approaches have been carried out.

The thesis has been organized into eight chapters in which the first chapter deals with the introductory part, application areas, and approach used for the gender classification and writer identification. Types of biometric traits, handwriting modality, characteristics of handwriting are also discussed in detail. The major assumptions, contribution, framework, and research objectives are also part of this chapter. Chapter 2 presents the state-of-the-art work on the writer identification system and the gender classification system. The goal is to cover all the necessary parameters such as author name, year, journal, feature extraction approach, classification methods, results achieved, etc.

In Chapter 3, the authors are presenting in detail the generation of corpus along with the data collection and maintenance. The framework and discussion on the phases with all necessary images are presented in this chapter. Chapter 4 presents the experimental part in developing the system based on the implementation of the feature extraction techniques and focusing on the hybridization of feature extraction techniques. To find the importance of hybridization, results are also compared. In Chapter 5, Principal component analysis-based developments of both the systems have been presented comprehensively so as to reduce the dimensionality and computational complexity. Chapter 6 presents another experimental work for achieving high accuracy results by implementing curve fitting and intersection and open endpoint-based feature extraction techniques along with the hybridization of classification, techniques have also been experienced to achieve the target high accuracy results. In Chapter 7, the combination of hybridization of feature extraction techniques and hybridization of classification techniques, have been experienced to attain promising results. Chapter 8 concludes by reporting the maximum accuracy achieved for both the gender classification and writer identification systems and also presented a table showing the summarized view of the results attained from the experiments. Finally, future perspectives of the proposed systems have been discussed that will surely open many novel ideas and concepts in the mind of novel researchers of handwriting based communities, followed by the reference section.