الجمـهورية الجزائرية الديمقراطية الشعبيـة
وزارة التعليم العالي والبحث العلمي
جـامعـة سعـيـدة د. مـولاي الطـاهـر
كلية الرياضيات و الإعلام الآلي و الاتصالات السلكية و اللاسلكية
قسم: الإعلام الآلي

Mémoire de Master en informatique

Spécialité : Intelligence Artificielle – Princes et Applications

T h è m e

Prédiction du genre d'un individu à partir de son écriture manuscrite hors-ligne

▪ **Présenté par :**

Ahmed Abdelmouneim MOUFFOK

▪ **Dirigé par :**

Dr. Tayeb BAHRAM

Année universitaire 2025-2026

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Introduction

This thesis addresses the problem of automatic gender prediction using scanned images of handwriting. Predicting the gender of the author of a handwritten sample using automatic image-based methods is an interesting pattern recognition problem with direct applicability in the forensic and historic document analysis fields [2]. Approaching this challenging problem raises a number of important research themes in computer vision:

- How can male and female handwriting styles be characterized using computer algorithms?
- What representations or features are most appropriate and how can they be combined?
- What performance can be achieved using automatic gender classification methods?

The current study describes a new and very effective technique that we have developed for automatic prediction of gender in offline handwriting. The goal of our research was to design state-of-the-art automatic methods involving only a reduced number of adjustable parameters and to create a robust gender recognition system capable of managing hundreds to thousands of writers [45]. There are two distinguishing characteristics of our approach: human intervention is minimized in the prediction process and we encode individual handwriting style using features designed to be independent of the textual content of the handwritten sample. Writer gender individuality is encoded using probability distribution functions extracted from handwritten text blocks and, in our methods, the computer is completely unaware of what has been written in the samples.

This thesis introduces a gender prediction technique that could have a significant impact on the field of forensic science [5]. Our method has been statistically evaluated using benchmarking datasets to differentiate between male and female writers.

1.2 Motivation and Context

Biometric systems have become an integral part of modern security and identification applications [7]. Traditional authentication methods, such as passwords and identification cards, are vulnerable to risks like theft, duplication, and unauthorized access. Consequently, biometric technologies have emerged as reliable alternatives that provide

enhanced security, user convenience, and fraud prevention. Handwriting occupies a unique position among biometric modalities because it reflects an individual's physiological and behavioral characteristics. Influenced by hand movement, writing habits, muscular coordination, and cognitive behavior, handwriting is a highly individualized trait. Offline automatic gender prediction systems analyze scanned images of handwritten documents. These systems do not require dynamic information, such as pen pressure or writing speed. Instead, they rely solely on the visual appearance and structural characteristics of handwriting.

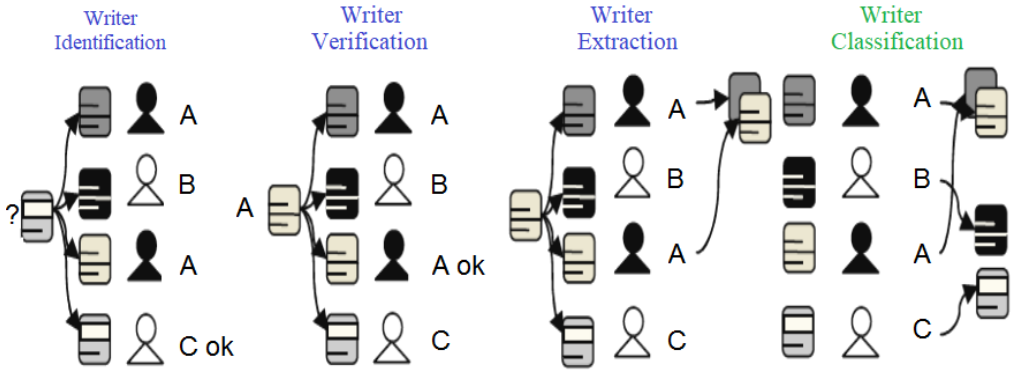


Figure 1.1: Illustration of different classification and recognition scenarios of writers.

The field of writer identification and verification in monolingual and multi-script contexts has advanced considerably, with many contributions addressing major world scripts [5]. However, the field of gender classification remains largely unexplored. Few existing studies address how features behave across languages or how to maintain performance when writers use different scripts. This research addresses these challenges by developing a robust, scalable gender classification system that can handle the following:

- A text-independent analysis is performed when the content of the writing is either unknown or irrelevant;
- Short and variable-length writing samples;
- High inter-class similarity and intra-class variability between male and female handwriting.

This work was also motivated by the limitations of current recognition methods (*i.e., automatic prediction of age, gender, and nationality in offline handwriting*). Texture- and contour-based methods often fail to capture the subtle geometric nuances of handwriting. Therefore, developing robust feature representation techniques for gender prediction is an important research challenge.

1.3 Objectives of the Research

The primary aim of this dissertation is to advance the research of automatic prediction of gender using offline handwritten texts within a fully text-independent framework. The proposed technique aims to demonstrate high accuracy and reliability in gender classification across different scripts. The specific objectives of this project are as follows:

- Propose handwriting features that are both discriminative and descriptive;
- Improve gender prediction performance by designing and evaluating complex feature representations that include combinations of texture descriptors and shape metrics;
- Build a complete offline gender classification system utilizing supervised classifiers such as the k-Nearest Neighbors (k-NN) algorithm;
- Use benchmark datasets to evaluate the system.

1.4 Dissertation Organization

This document is organized into the following six chapters:

- **Chapter 1 (General Introduction)** : This section (Section 1) introduces the fundamental concepts of biometrics and handwriting analysis.
- **Chapter 2 (Individual Characterization)** : Section 2 surveys the main research contributions to offline gender prediction, covering handcrafted and learning-based approaches.
- **Chapter 3 (Gender Prediction: State-of-the-Art)** : Section 3 provides a concise overview of feature extraction methods for offline writer recognition, encompassing classification (i.e; automatic prediction of gender).
- **Chapter 4 (Proposed Approach)** : Section 4 elaborates on the three main steps of the proposed system: preprocessing, feature extraction/combination, and classification.
- **Chapter 5 (Results and analysis)** : Section **??** outlines the well-known handwritten databases utilized in this research.
- **Chapter 6 (Conclusion)** : Section **??** concludes the thesis and outlines ideas and recommendations for future research.

1.5 Methodology and Technologies

These objectives will be achieved using the following methodology and tools:

- **Programming Languages:** Microsoft Visual C++ (MSVC) is used for system development, data preprocessing, and model training.
- **Machine Learning Frameworks:** The standard C and C++ libraries are used to implement classifiers such as k-NN, SVM and k-means.
- **Documentation:** TeXStudio and MiKTeX for LaTeX writing and thesis formatting.

Chapter 2

Individual Characterization

Every individual possesses unique characteristics. These characteristics form the basis of verification and authentication systems, which are becoming increasingly important for highly secure applications. In recent years, there has been a considerable increase in demand for automatic systems that can identify or verify identity, whether for controlling access to personal data or assisting forensic experts in analyzing handwritten documents. This chapter introduces the key concepts of writer recognition, i.e. verification (or authentication).

2.1 Introduction

Biometrics refers to the technologies used to measure and analyze human characteristics, whether physiological or behavioural, for identification or authentication purposes. As each person has unique traits, biometrics is a reliable method of personal identification. Unlike traditional identifiers, such as names, passwords or ID cards, which can be lost, stolen or forgotten, biometric traits are intrinsic to the individual (see 2.1).

Biometric systems operate in two main modes: verification (1:1 comparison) and identification (1:N comparison). In verification mode, the system confirms a claimed identity by comparing the input biometric data with a stored template (i.e. producing an accept/reject decision). In identification mode, the system determines an individual's identity by comparing the input data with multiple stored templates to find the best match (i.e. the most likely match in closed-set identification or declaring no match in open-set identification) [5].

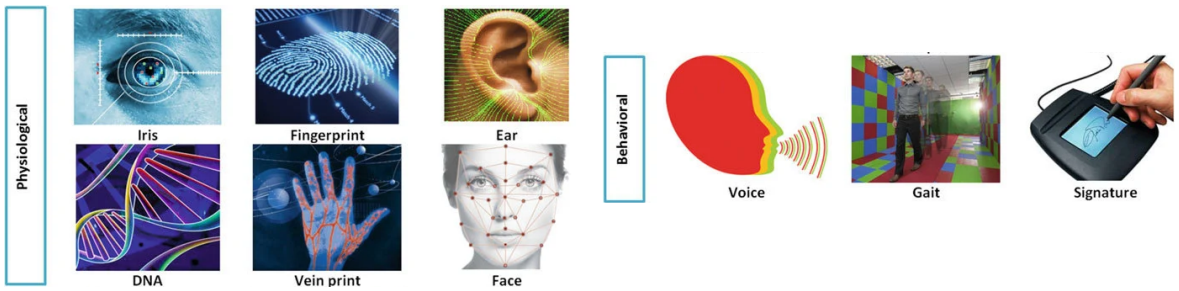


Figure 2.1: Biometric traits that can be used for identification or verification.

2.2 Handwriting as a Behavioral Biometric Trait

Offline handwriting, in particular, has attracted significant interest among various behavioural biometric modalities due to its natural use in many real-world scenarios, such as legal documents, forms, and historical archives. Handwriting contains many personalized features that can reflect an individual's motor habits, cognitive style and neuromuscular characteristics.

Writer recognition, also known as biometric recognition based on handwriting samples, plays a crucial role in the field of pattern recognition. It is a classical problem in forensic science with applications in several significant research domains. These include forensic examination and criminal justice systems, paleography (i.e., the study of historical and medieval handwriting), biometric recognition, security, and handwriting identification and verification. Forensic writer recognition typically involves the task of automatically determining the authorship of a given handwriting sample (e.g., a suspicious or fraudulent letter, ransom note, etc.). This task is commonly divided into two main modes: identification and verification. In the identification mode, which is a one-to-many comparison or multi-class classification problem, the objective is to determine the author (or writer) of an unknown handwritten sample from a set of known writers. In contrast, the verification mode focuses on determining whether two handwriting samples were produced by the same individual. This involves one-to-one comparisons and is framed as a binary classification problem [6].

Writer recognition process is challenging due to intra-writer variability (i.e. variations in an individual's writing over time or under different conditions) and inter-writer similarity (i.e. similarities between different individuals' writing styles) (see 2.2). Nevertheless, handwriting has one key advantage: it can be acquired non-intrusively using simple devices such as scanners or cameras, making it ideal for forensic and commercial applications [13].

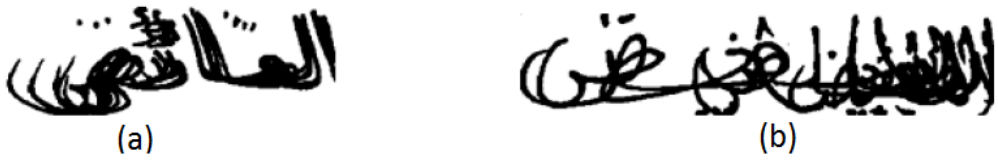


Figure 2.2: Illustration of intra- and inter-writer variability by superimposition: (a) a word written 6 times by the same writer and (b) the same word written by 6 different writers.

2.3 Within-writer variance vs. Between-writer variation

Writer verification are only possible to the extent that the variation in handwriting style between different writers exceeds the variations intrinsic to every single writer considered in isolation. The results reported in this thesis ultimately represent statistical analyses, on our datasets, of the relationship opposing the between-writer variation and the within-writer variability in feature space. The present study assumes that the handwriting was produced using a natural writing attitude. It is important to observe

that forged or disguised handwriting is not addressed in our approach. The forger tries to change the handwriting style usually by changing the slant and/or the chosen allographs. Using detailed manual analysis, forensic experts are sometimes able to correctly identify a forged handwritten sample. On the other hand, our proposed algorithms operate on the scanned handwriting faithfully considering all the graphical shapes encountered in the image under the premise that they are created by the habitual and natural script style of the writer.

2.4 Factors causing variability in handwriting

Figure 1.1 shows four factors causing variability in handwriting [13]:

1. *Affine transforms*: Affine transforms are under voluntary control. However, writing slant constitutes a habitual parameter which may be exploited in writer recognition;

2. *Neuro-biomechanical variability*: Neuro-biomechanical variability refers to the amount of effort which is spent on overcoming the low-pass characteristics of the biomechanical limb by conscious cognitive motor control;

3. *Sequencing variability*: Sequencing variability becomes evident from stochastic variations in the production of the strokes in a capital E or of strokes in Chinese characters, as well as stroke variations due to slips of the pen;

4. *Allographic variation*: Allographic variation refers to individual use of character shapes.

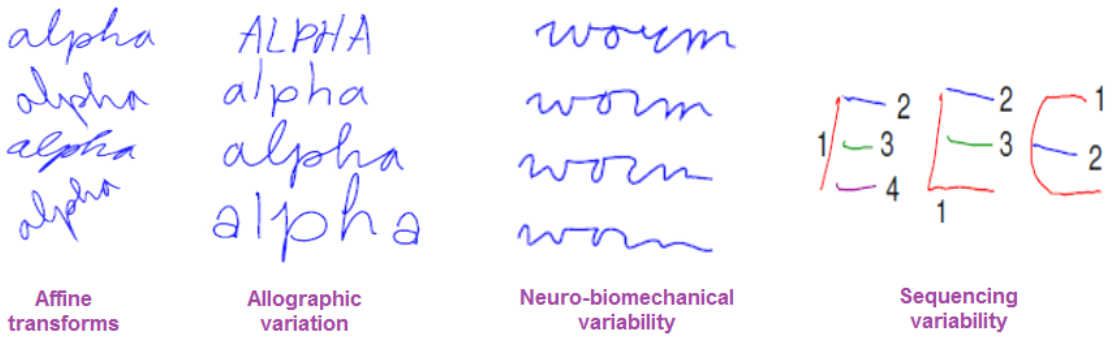


Figure 2.3: Factors causing handwriting variability [13].

2.5 Writer recognition

Handwriting, as a form of behavioral biometric trait, encapsulates the neuromuscular patterns and cognitive processes unique to each individual. The act of writing is influenced by a combination of motor coordination, learned habits, psychological state, and physiological conditions, making it a rich source of biometric data. These characteristics are particularly useful for tasks such as writer identification and writer verification, which are two core problems in the domain of handwriting biometrics.

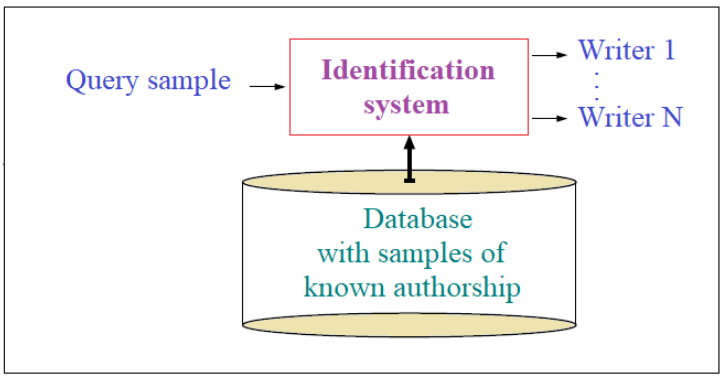


Figure 2.4: A writer identification system retrieves, from a database containing handwritings of known authorship, those samples that are most similar to the query. The hit list is then analyzed in detail by a human expert [13].

After image preprocessing, asserting writer identity based on handwriting images requires three main operational phases:

1. Feature extraction;
2. Feature matching / feature combination
3. Writer recognition (identification and verification).

Figure 2.5: A writer verification system compares two handwriting samples and takes an automatic decision whether or not the input samples were written by the same person [13].

A *writer identification* system performs a one-to-many search in a large database with handwriting samples of known authorship and returns a likely list of candidates (see Fig. 2.4). This list is further scrutinized by the forensic expert who takes the final decision regarding the identity of the author of the questioned sample [26]. *Writer verification* involves a one-to-one comparison with a decision whether or not the two samples are written by the same person (see Fig. 2.5). The decidability of this problem gives insight into the nature of handwriting individuality [7]. In writer identification searches, all the samples in the dataset are ordered with increasing dissimilarity (or distance) from the query sample. This represents a special case of image retrieval, where the retrieval process is based on features capturing handwriting individuality [6]. In writer verification trials, if the distance between two chosen samples is smaller than a predefined threshold, the samples are deemed to have been written by the same person. Otherwise, the samples are considered to have been written by different writers. Writer verification has potential applicability in a scenario in which a specific writer must be automatically detected in a stream of handwritten documents. In forensic practice,

both identification and verification play a central role [13]. Gender (age or nationality, i.e., classification) detection of the writer form a handwriting has been a challenging problem for documents analysis community [23] (see Fig. 2.6). From the handwriting analysis perspective, gender detection is a task between handwriting recognition and writer identification. In handwriting recognition, the variability of handwriting should be blurred to raise a common feature among the writers to enable more accurate handwritten document recognition. In contrast, the variability in the writer identification task should be highlighted to increase differences among the handwritings of different writers in order to achieve more accurate gender/age classification results. In gender detection, however, we look for common features between a particular group (male or female) of writers [3]. The gender detection problem is a binary classification, while age detection is a multi-class classification problem. In general, to address the problems of gender and age detection, signal and image processing followed by pattern classification techniques are applied to handwriting signals or images.

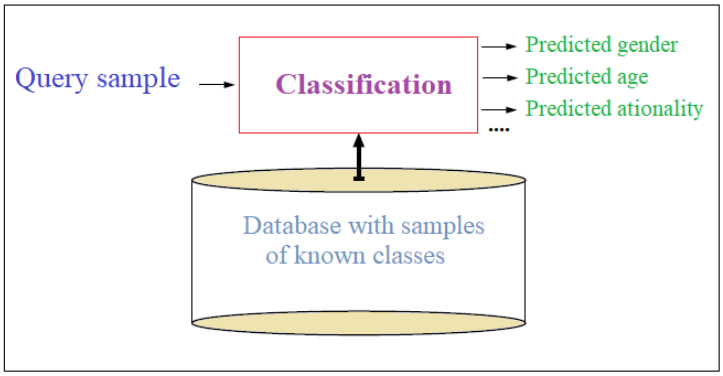


Figure 2.6: A writer classification (Gender, age and nationality) system[23].

2.6 Online and Offline Writer Recognition and prediction in offline handwriting

Traditionally, writer recognition (identification and verification) or prediction (gender, age, or nationality) systems are divided into two kinds: online and offline [43]. The online handwritten data contain more information about the writing style of a person, such as pentip position, pressure, and/or angles (as a function of time), that is not available in the offline data. However, offline systems work on digitized images of writing. Consequently, compared to online systems, it is more difficult for offline systems to achieve significantly higher recognition rates, and offline systems are widely considered more challenging than their online counterpart.

2.7 Text-dependent vs. Text-independent methods

The techniques for writer recognition and classification are traditionally categorized into two broad classes: *text-dependent* and *text-independent* methods. Text-dependent methods require the writing samples to be compared to contain the same fixed text. Signature verification, for example, can be considered in this category. These methods

usually compare individual characters or words from a known transcription. Therefore, the text must first be recognized or segmented (manually or automatically) into characters or words before writer recognition (or prediction) can occur [8]. Text-independent methods, on the other hand, verify the author of a document regardless of its semantic content. These methods use features extracted from an entire text image or a region of interest. Only a minimal amount of handwriting is necessary to derive stable, text-insensitive features [14].

2.8 Conclusion

In this chapter, we learned that characterizing individuals is not a new concept. Based on the measurement of physical or behavioral characteristics, also known as biometrics, this concept dates back to the 14th century. Studies of biometrics have revealed four essential properties of characteristics: distinction, permanence, quantifiable, and universality. According to these properties, each characteristic has its own set of advantages and disadvantages. In the next chapter, we will review the various offline writer identification and verification approaches proposed in the literature.

Chapter 3

Writer Recognition and Prediction : State-of-the-Art

3.1 Introduction

Despite the development of electronic documents and predictions of a paperless world, the importance of handwritten documents has retained its place and the problems of identification and authentication of writers have been an active area of research over the past few years. A wide variety of systems that are based on the use of computer image processing and pattern recognition techniques have been proposed to solve the problems encountered in automatic analysis of handwriting and recognition of the writer of a questioned document. A comprehensive survey of the work in writer recognition until 2022 has been presented in [5]. Here, we will be more interested in surveying the approaches developed in the last several years, thanks to the renewed interest of the document analysis community for this domain.

This section offers a summary of research initiatives, highlighting progress in the filed of offline text-independent writer recognition (i.e., identification and verification). A key component of writer recognition systems is the essential process of feature extraction, which entails converting an input handwriting image into a feature vector representation (for instance, PDF or Probability Distribution Function as cited in [8]). The existing literature classifies feature extraction methods into two primary categories: *handcrafted features* and *learned features*.

3.2 Handcrafted features

Numerous effective techniques for well-established handcrafted features have been documented in the literature for the purpose of identifying and verifying the authorship of handwritten texts [6]. This initial category can be further divided into three primary classifications: 1. codebook-based techniques, 2. texture-based methods, and contour-based methods.

3.2.1 Codebook-based methods

The codebook or bag-of-shapes can be used as a distinctive measure (i.e., feature vector) to characterize the writing style of each writer [8, 46, 18, 31]. The shape occurrence probability is a characteristic for a given writer and may be employed to distinguish

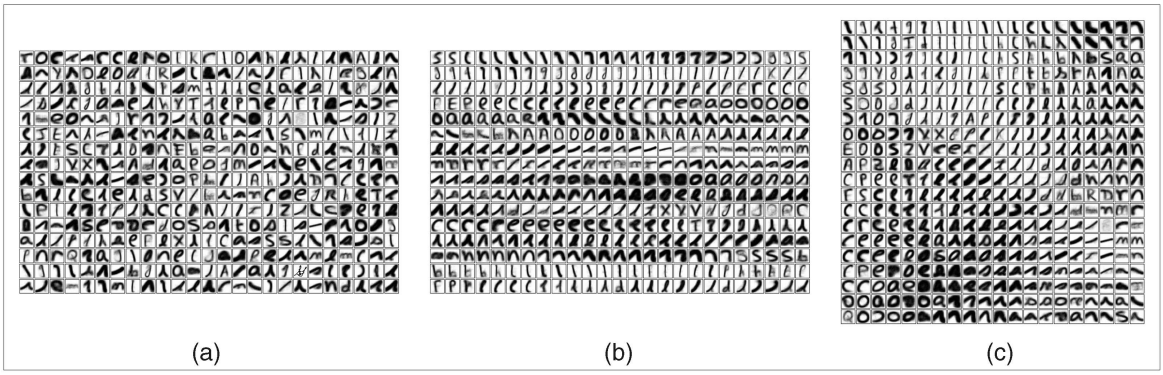


Figure 3.1: Examples of codebooks with 400 graphemes. For (a) kmeans and (b) ksom1D the graphemes have been placed 25 in a row, while, for (c) ksom2D, the 20×20 original SOM organization has been maintained.

between intra-cluster and inter-cluster similarity. Several methods are proposed in the literature to create the codebook, such as the Graphemes (writing fragments or small patterns) [13], Elliptic Graphemes [1], Junclets [28], Bagged Discrete Cosine Transform (BDCT) descriptors [33], and Implicit Shape (Patches around the key points) [9]. Regarding the allographic features, the writer is considered as a stochastic generator of graphemes and a set of emission shape probabilities is computed by building a normalized histogram. In 2007, Bulacu et al. [13] proposed the use of allographic features to characterize writer individuality thus making the approach quite similar to Bensefia et al. (2005) [10]. The probability distribution of each writer is computed using a common codebook of 20×20 graphemes generated with the ksom2D clustering algorithm for each document individually (see Fig. 3.2).

In 2015, Abdi and Khemakhem [1] used a beta-elliptic model (Elliptic Graphemes) to generate a synthetic codebook for Arabic writer recognition (see Fig. 3.1). A total of 60 feature vectors were extracted using template matching.

In the same year, He et al. [28] applied the detected junctions to writer identification using a compact representation, called Junclets. The junction detection in handwritten images is determined by analyzing the stroke-length distribution in every direction around a reference point inside the ink of texts. In 2017, Khan et al. [33] introduced a text independent writer identification method based on Bagged Discrete Cosine Transform (BDCT) descriptors. Universal codebooks are first employed to generate multiple

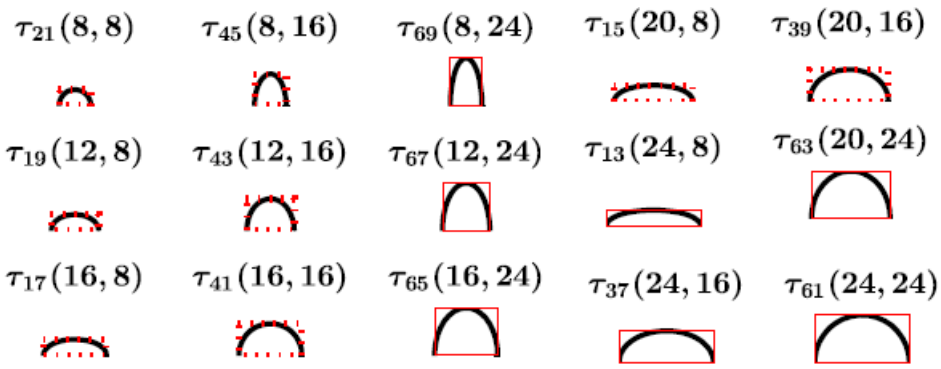
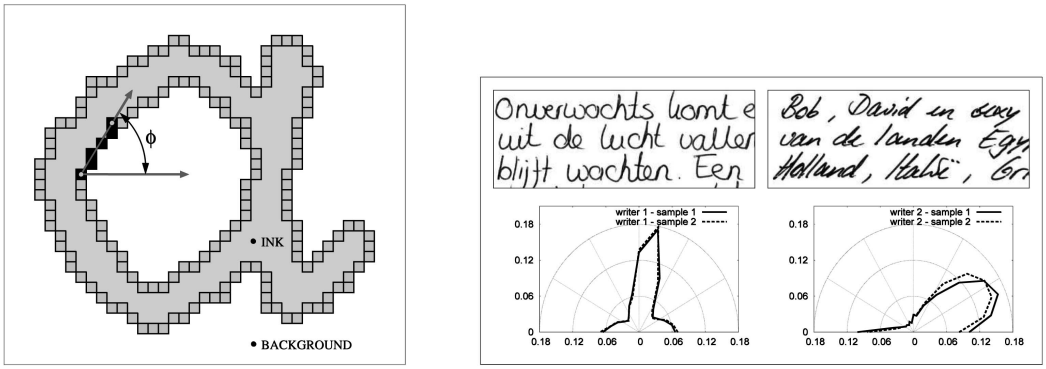


Figure 3.2: Examples of elliptic templates [1].



(a) (b)

Figure 3.3: (a) Schematic description for the extraction method of the contour-direction PDF (feature f1). The handwritten letter "a," provided as an example, would be roughly twice as large in reality. (b) Examples of lowercase handwriting from two different subjects [13].

predictor models and the final result on a handwritten sample is obtained using the majority voting rule from all predictor models. In 2019, Bennour et al. [9] exploited the key points in handwriting to generate the implicit shape codebook and identify the right writers. The problem with this type of methods is the complexity of the segmentation and feature extraction algorithms, which are highly expensive in terms of execution time.

3.2.2 Texture-based methods

Texture-based approaches to perform offline automatic writer identification take into account each grayscale handwriting image as a unique texture and rely on extracting features from the entire image (whole document) [13, 17, 25], blocks [11], connected-components [5, 6, 7, 8, 14, 32], or writing fragments [22] to produce texture descriptors (i.e. features, measures, or attributes). A Probability Distribution Function (PDF or vector of probabilities) in a given document is computed and employed to characterize the writer's style [13]. In this case, the appearance probability; i.e. histogram bin, is

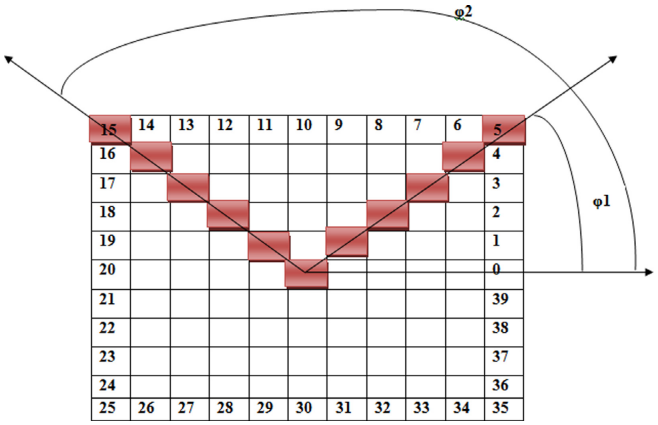


Figure 3.4: Extraction of edge-hinge distribution [30].

a characteristic for each writer and may be employed to distinguish between inter-class and intra-class similarities.

In 2007, Bulacu et al. [13] proposed to perform handwriting writer identification by using the textural features. Probability Distribution Functions (PDFs) are calculated using the entire handwriting image (document) and its contours for contour direction (see Fig. 3.3), contour-Hinge, direction co-occurrence and Run-Length on white distributions (RL). In 2013, Bertolini et al. (2013) published a texture-based descriptor for writer identification. The authors used Local Binary Patterns (LBP) in a comparative study with a Local Phase Quantization (LPQ) and concluded that LPQ performed better than their LBP variant.

One year later, Wu et al. [49] wished-for a method for offline text-independent writer identification based on the Scale Invariant Feature Transform (SIFT). In particular, the SIFT Descriptor Signature (SDS) and the Scale Orientation Histogram (SOH) are extracted from handwriting images to characterize different writers. In 2016, we offered a set of textural features to characterize writer individuality [8]. These features include Direction-Length, Angle-Length, Direction Co-Probability, and Angle Co-Probability of connected components. In 2018, Singh et al. [47] divided cursive handwriting into nine texture blocks to compute histograms of the Local Binary Pattern (LBP) and the Center Symmetric Local Binary Co-occurrence Pattern (CSLBCoP). For each writing sample, a set of 9 blocks were created and feature vectors were computed for writer identification. Kessentini et al. [30] combined Edge-Hinge with a fragment length of 6 and 7 pixels (see Fig. 3.4) and Run-length features in the Dempster-Shafer Theory (DST) model to improve the identification rates (in the same year). In 2019, Hannad et al. [21] presented an approach for the identification and characterization of different writers that combines two textural features: the Histogram of Oriented Gradients (HOG) and Gray Level Run Length (GLRL) Matrices. Before the end of 2019, Khan et al. [32] applied a combination of the Scale-Invariant Feature Transform (SIFT) and RootSIFT descriptors in a set of Gaussian mixture models (dissimilarity SGMM and DGMM) to compare and classify the handwritten documents. In 2020, Chahi et al. [14] developed a Local gradient full-Scale Transform Patterns (LSTP) based method for writer identification. The authors extracted a set of feature code maps from resized

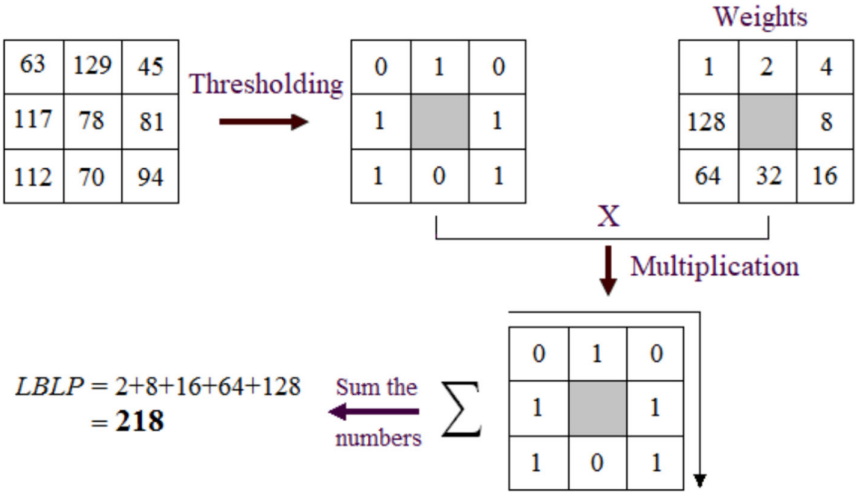


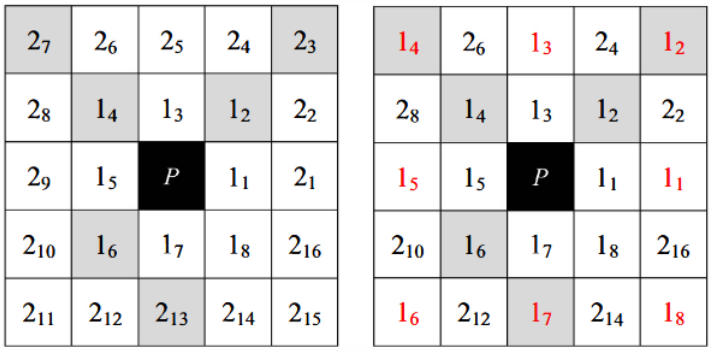
Figure 3.5: Example of LBLP code calculation [6].

small regions of interest of the writing (connected components sub-images) based on the distribution of local intensity gradients. Lastly, Bahram et al. [5] introduced a feature termed the Local Black Pattern Histogram (LBLP, see Fig. 3.5), which is grounded in texture connected-components and involves monitoring each black point for the purpose of text-independent writer identification (in 2025).

3.2.3 Contour-based methods

Contour-based techniques constitute a unique category within the field of offline writer identification and verification [30]. In these methodologies, Probability Distribution Functions (PDFs) of local attributes or features are derived from the inner and outer contour pixels, opposed to the conventional image pixels. These attributes are meticulously designed to capture characteristics unique to individual writer, such as angle, direction, curvature, slant, ink-trace widths, and letter shapes [5].

In 2015, Xiong et al. [50] approached writer identification from a texture contour perspective, utilizing the SIFT descriptor along with contour directional features (CDF) (see Fig. 3.6).



(a) Grid microstructure feature (b) Contour-directional feature

Figure 3.6: Comparison of GMF and CDF [50].

One year later, Bahram et al. [8] introduced four features: direction-length (DL), direction co-probability (DCP), angle-length (AL), and angle co-probability (ACP). These features are based on texture contours and involve monitoring each contour point to identify writers, regardless of the text. Expanding upon the co-occurrence feature for writer recognition, contour-based features can be formulated, including CoHinge and QuadHinge. Importantly, Bahram [5] released a notable writer identification system in 2022 that employs texture contour-based features. This system harnesses the joint distribution of the Modified Local Binary Pattern (MLBP) and the Ink-trace Width and Shape Letters (IWSL) measurements (MLBP-IWSL) to capture complex handwriting details, encompassing direction, curvature, slant, and shapes, thereby characterizing individual writing styles. Finally, in the same year, Bahram et al. [6] investigated the relationship between ink direction and ink width (width patterns or width ink traces) to aid in writer recognition through the Local Contour Pattern with Ink-trace Width and Shape Letters (LCP-IWSL) feature.

3.3 Learned features

In this section, features that have traditionally been crafted by experts, including those based on texture, contour, and codebook methodologies, can now be automatically derived through the implementation of Convolutional Neural Networks (CNNs) [48]. This approach enables the generation of features that are learned directly from the data, specifically aimed at offline writer identification and verification [51]. A review of existing literature reveals that most methodologies utilized for feature extraction are trained on handwriting samples comprising fragments, words, or lines. Notably, in 2015, Fiel and Sablatnig [19], along with Christlein et al. [16], emerged as pioneers in the use of CNNs within the field of writer recognition, employing scanned handwriting images. Following this, Javidi and Jampour [29] presented a hybrid deep architecture that merges handcrafted and deep features to define writer individuality, offering a streamlined extended version of FragNet that leverages auxiliary information from the Handwriting Thickness Descriptor (HTD). Kumar and Sharma [35] proposed a segmentation-free model that employs a CNN to identify query writers, referred to as SEG-WI. Moreover, in 2021, reference [27] introduced a technique for extracting global-context information from handwritten word images using CNNs, incorporating a recurrent neural network (RNN) to model the spatial relationships among sequences of fragments. Another CNN-based methodology was elaborated in [44], concentrating on deep feature extraction around key points within small handwriting patches. Key points were detected using the FAST and Harris corner detectors, while the feature encoding process utilized V LAD and triangulation embedding (TE). More recently, in 2023, Chahi et al. [15] unveiled a multipath deep CNN-based technique named WriterINet, designed to learn and directly identify writers. Kumar [37] introduced a writer identification modified Histogram of Oriented Gradients (HOG) to create writer descriptors, which can be used to determine the authorship of a specific handwritten word image. Furthermore, in 2024, Briber and Chibani [12] developed a closed system consisting of multiple convolutional layers. frequently trained on entire documents, to achieve superior performance, albeit at a considerable computational expense. Finally, in 2025, Zhao et al. [52] introduced a writer identification method based on an integration strategy combining complementary learning models. This approach involves combining multiple probabilistic mass functions, i.e. the properties of the top-k values.

3.4 Comparative Analysis and Challenges

It is crucial to emphasize that offline writer recognition employing textural features (such as contours or images) within a single (or multi)-script context has shown high recognition rates, has proven efficient regarding processing time, and is generally preferred when at least three lines of text are accessible for training and testing purposes [7]. These approaches are typically not intended for the automatic identification or verification of authors from handwriting samples that incorporate multiple scripts, such as Arabic and French [44]. Moreover, they require a significant amount of handwritten text samples from each individual to attain effective classification and are sensitive to structural variations [5, 44]. It is important to highlight that techniques based on learned features exhibit robustness and enhanced recognition performance, especially at the word level [36]. However, these methods necessitate a substantial number of samples

and entail considerable time investment concerning parameter tuning and calculations during the training phase, with a tendency towards over fitting.

In light of the aforementioned observations, this paper introduces a texture-based technique for *offline automatic prediction of gende*r, focusing on the image analysis. The handwriting features utilized in this study are derived from images of short paragraphs, each comprising approximately three lines. These features have been employed to characterize distinct handwriting styles across various scripts and languages. A comprehensive description of our methodology can be found in Chapter 4, while Chapter **??** discusses the efficacy of the proposed features in prediction of gender, utilizing established benchmark database (HHD).

3.5 Conclusion

Predicting an individual's gender from their offline handwriting remains challenging, primarily due to variability in scripts and differences in text length. While HHD benchmark has primarily focused on prediction (classification) scenarios, this dataset remain valuable for evaluating new approaches.

This study proposes two new text-independent features for automatic prediction and evaluates their effectiveness in detection of gender. Using the HHD dataset, we demonstrated that these features can capture writer-specific traits with high discriminative power across varying conditions. The experimental results, detailed in the following chapters, confirm the potential of our features to generalize limited text lengths. These findings suggest promising avenues for developing robust, language-independent prediction of gender in offline handwriting systems.

Chapter 4

Proposed approach

4.1 Introduction

In this section, binary connected components are used to represent the scanned handwriting images during the feature extraction step. This is a very effective vectorial representation that will allow a fast computation of the proposed features. Furthermore, connected components are crucial for describing writing shapes. For writer characterization, we propose two complementary features based on the Local Gray-level Value Pattern (LGP) framework to characterize writers.

1. LGP_B (black pixels)

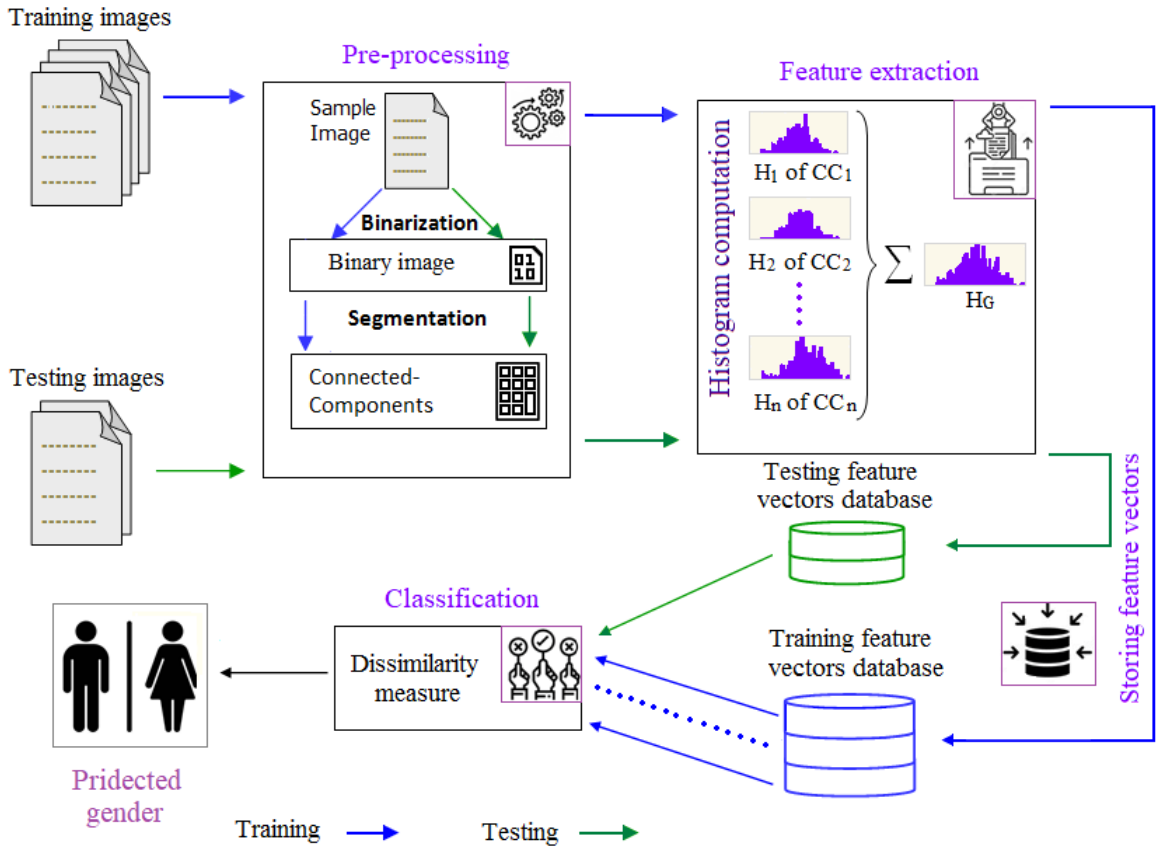


Figure 4.1: The framework of our writer recognition (verification) system.

2. LGP_W (white pixels)

As illustrated in Fig. 4.1, our proposed approach consists of three main processing steps:

1. *Preprocessing*: Normalizes and enhances scanned handwriting images to ensure consistent feature extraction across different datasets;

2. *Feature Extraction*: Computes LGP_B (black pixels) and LGP_W (white pixels) from binary connected components. Then, it combines them to encode characteristics specific to the writer (i.e., $LGP_{BW} = LGP_B + LGP_W$);

3. Classification (Prediction of gender): It uses distance metrics, such as chi-squared (χ^2) and Manhattan (Mh) to match query samples with a database of known classes (Male/Female).

4.2 Pre-processing

Like most of the pattern recognition problems, the preprocessing step plays an important role in the performance of the writer recognition and prediction (gender, age or nationality) systems. For our experiments, the input data (i.e., feature extraction step) are in black connected-components, so the pre-processing step is necessary and it consists in binarizing the document images (foreground/background separation), extracting the connected-components, and removing the non-significant connected components (i.e., small components, dots, etc.). In this section, a very effective vectorial representation is considered which will subsequently allow a fast calculation of the proposed feature.

4.2.1 Binarization

The vast majority of offline writer recognition and prediction techniques use binary input images rather than color or grayscale ones [7]. While no comprehensive studies appear to have been conducted, the available results suggest that this format is the most useful despite the loss of information. While grayscale may be appropriate for certain features (e.g. attempting to reconstruct pen pressure), it generally produces images that are too complex for further processing.

Binarization is a key step in preparing images for analysis. It takes a grayscale or color image and simplifies it by converting it to black and white. In this simplified image, only two colors exist: black for important details such as text or handwriting and white for the background, such as the paper. This simplification makes tasks such as character recognition, writer identification and verification, and shape and outline detection much easier. Depending on how the threshold is determined, binarization methods are commonly divided into two main categories.

1. *Global Binarisation* : Global binarisation applies a single threshold value to the entire image. Each pixel is classified as foreground or background based on this one global threshold.

   —Advantages: Simple, fast, and effective for images with uniform lighting and contrast.

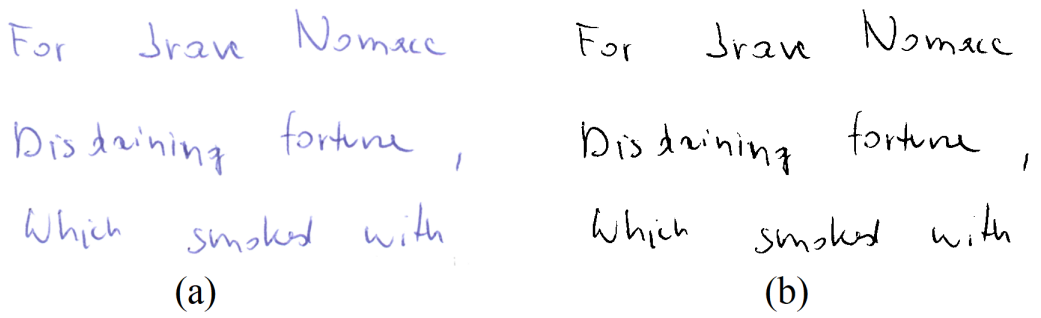


Figure 4.2: Global Image Thresholding using Otsu's Method [42].

—Common method: Otsu's method, which automatically selects an optimal threshold by maximizing inter-class variance.

2. *Local Binarisation* : Local binarisation, also known as adaptive binarisation, computes a different threshold for each pixel based on the local neighborhood (a small window around each pixel). This makes it more suitable for documents with irregular lighting, shadows, and low quality.

   —Advantages: Durable to variations in lighting and background noise.

   —Common methods: Niblack, Sauvola, and Wolf algorithms, which dynamically adjust thresholds using local mean and standard deviation.

We decided to use Otsu's method [42] to binarize the handwriting images for our work. Our images are clean and well-scanned with good lighting and a clear distinction between text and background, so this method is ideal for our needs (see Fig. 4.2). Otsu's method automatically identifies the optimal threshold to distinguish black text from a white background, making it ideal for this situation. It's also fast and simple to use, making it a practical choice. Using this method allowed us to obtain clean binary images, making subsequent steps, such as writer analysis, easier and more reliable.

4.2.2 Connected Components Extraction

Extracting connected components is a critical step in processing binary images. The goal is to identify and separate individual objects made up of connected pixels. In a black-and-white image, for instance, the black pixels typically represent meaningful content, such as handwritten text, symbols, or shapes, while the white pixels represent the background. The extraction process examines the image to find groups of black pixels

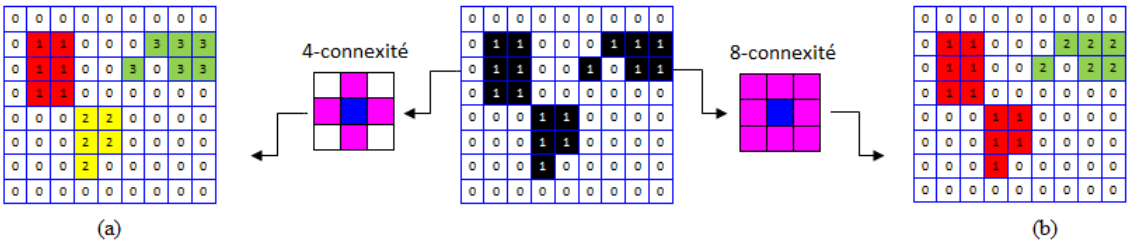


Figure 4.3: Connected components extraction, (a) 4-connectivity and (b) 8-connectivity.

connected to each other based on a connectivity rule. This rule typically considers 4-connectivity, which includes pixels connected horizontally or vertically, or 8-connectivity, which includes diagonal neighbors (see Fig. 4.3). Pixels that meet this condition are grouped together as one connected component. Once these groups have been identified, each connected component is assigned a unique label. This labeling allows us to treat each component as an individual object for further analysis. In handwriting recognition or writer recognition, for example, connected components often correspond to letters, strokes, or words requiring separate analysis.

This step is crucial because it simplifies complex images by breaking them down into smaller, meaningful parts. After extracting connected components, other important tasks such as feature extraction or classification become more manageable and accurate.

4.2.3 Removal of small connected components

As stated in Bahram et al. [8], a connected component can be defined as an ink blob or complete region of ink trace, determined by a set of pixels, i.e., all connected to each other, drawn without lifting the pen from paper. In binary images of handwriting, the foreground (black) pixels correspond to the black ink trace and the white pixels correspond to the background. In order to capture more details of the directional strokes, the slant and skew corrections of the handwritten document are not used in this work. Subsequently, some non-significant connected components like diacritics small-components with a width or height smaller than one stroke-width, scatter noise, periods, commas, and dots, are discarded in this step (see Fig. 4.4).

As illustrated in Fig. 4.1 and Algorithm 1, each handwritten document, comprising various words or lines of text, is segmented into images of connected components $\{CC_i\}$. The pages from the CVL database [34], initially available in RGB format [7].

Algorithm 1 : Preprocessing step

Input: Gray-scale image I_G
Output: $\bigcup_{i=1}^{N} CC_i$

1: $I_B \leftarrow$ ***Otsu***(I_G)
2: **for** $i \leftarrow 1$ **to** N **do**
3: $\quad CC_i =$ ***CCExtract***$(I_B, X_{min}(i), Y_{min}(i), width(i), hight(i))$
4: $\quad$ **if** $(\psi(CC_i) > \xi)$ **then**
5: $\quad\quad$ **Add**(CC_i)
6: $\quad$ **else**
7: $\quad\quad$ **Skip**(CC_i)
8: $\quad$ **end if**
9: **end for**
10: **return** $(\bigcup_{i=1}^{N} CC_i)$

4.3 Feature Extraction

Feature extraction is the process of converting an input handwriting image into feature vectors with the aim of increasing the performance of the model. It is the heart of handwriting-based biometrics [1]. In the case of statistical features, the feature vector

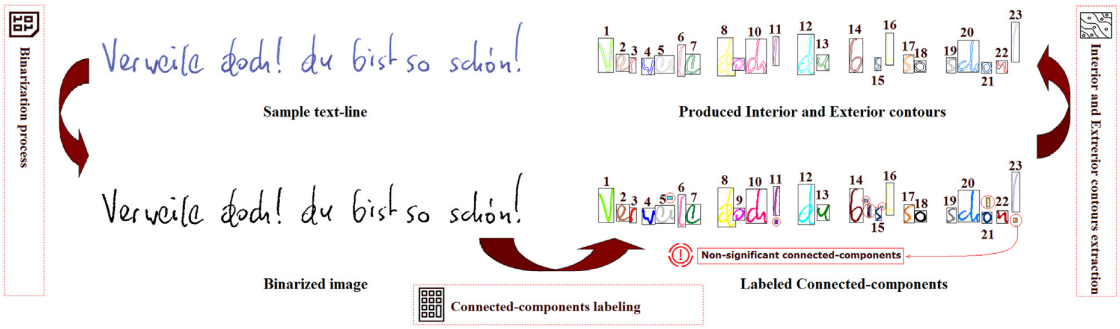


Figure 4.4: Preprocessing of an example handwritten text line taken from CVL [34] database (from [5]).

is calculated by carrying out a large number of measurements, accumulating the results in a histogram, normalizing and interpreting it into a probability distribution [13]. In this section, we characterize the individual writing style of a given writer by computing two texture-based features: LGP_B and LGP_W.

4.3.1 LBP operator

The traditional Local Binary Pattern (LBP) operator proposed by Ojala et al. in [40] and improved in [41] for texture analysis and description is a fast and computationally efficient texture descriptor for grayscale images. The LBP descriptor has been used to identify the authorship of handwritten documents in most of the texture techniques in the literature [11, 25], which divides the entire image into regions to extract unique and useful texture features. This powerful and most popular operator labels each center (or referenced) pixel p_c of a grayscale image I_G by thresholding its eight surrounding pixels $(p_0, \ldots, p_7)$, and considering the result as a binary string (i.e., obtained by concatenating all these binary values) or a decimal number. Fig. 4.5 shows an example of the original grayscale LBP code, which is computed by comparing the center pixel (x_c, y_c) intensity with eight neighbors (x_k, y_k) in a 3×3 neighborhood.

A limitation of the basic LBP operator itself is its small 3×3 neighborhood around the central pixel: this may leave it unable to capture the dominant features with large-

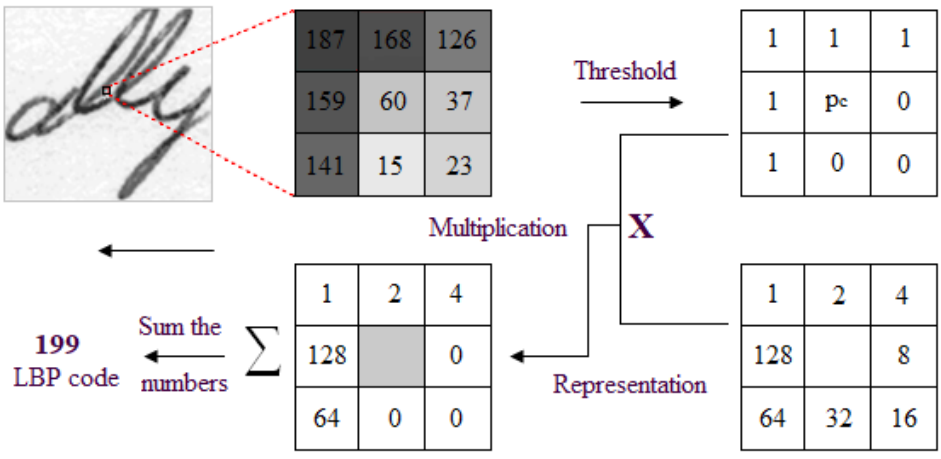


Figure 4.5: Example of the traditional LBP operator [40, 41].

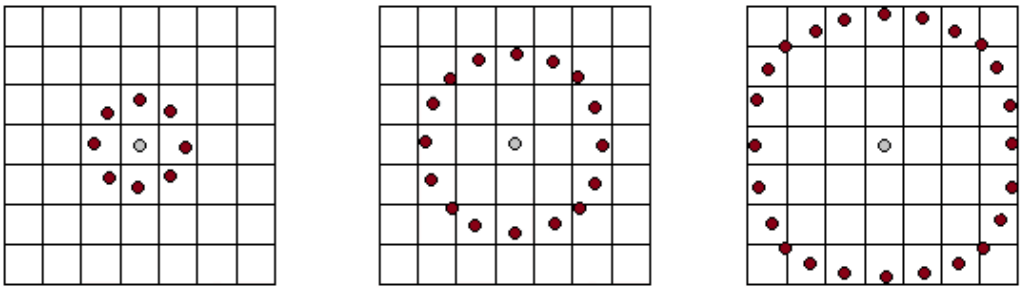


Figure 4.6: Examples of the extended LBP operator [41] (three different values of P and R).

scale structures. Subsequently, the original LBP operator was generalized to use different neighborhood sizes, i.e. to capture discriminative features at different scales [41]. For a given pixel p_c , its P neighbors $\{p_k\}_{,k=0...P-1}$ are obtained using a bilinear interpolation. Formally, the LBP operator of the center pixel p_c at position (x_c, y_c), can be expressed as (cf. Eq. 4.1)

$$LBP_{P,R}(x_c, y_c) = \sum_{k=0}^{P-1} s(i_k - i_c)\ 2^k \tag{4.1}$$

where i_c represents the gray value of the center pixel p_c, i_k represents the gray values of P surrounding pixels on a circle of radius $R(R > 0)$, P is the number of neighbors, 2^k is the binomial factor, and s defines the sign function as follows (cf. Eq. 4.2):

$$s(i_k - i_c) = \begin{cases} 1, & if \quad i_k \geq i_c \\ 0, & if \quad i_k < i_c \end{cases} \tag{4.2}$$

Fig. 4.6 illustrates an example of the extended LBP operator using different neighborhood sizes (i.e. different values of P and R).

4.3.2 *LGP* features extraction (texture-based)

We analyze the distinctiveness of an individual's handwriting style by computing the normalized histogram of the Local Gray-level Value Pattern (LGP) from scanned images of their handwritten documents. This LGP feature is instrumental in minimizing misclassification rates and enhancing the overall performance and robustness of our texture-based writer recognition system.

LGP operator The Local Gray-level Value Pattern (LGP) texture descriptor assigns labels to the pixels of a grayscale image I_G by assessing the neighborhood surrounding each pixel $p_c(x, y)$, with the resulting LGP represented as a decimal number. The LGP value for a specific pixel can be mathematically expressed as

$$LGP_{E,R}(x, y) = \sum_{k=1}^{8} g(i_k - th)\ 2^{k-1} \tag{4.3}$$

The function $g(i_k - th)$ is defined as

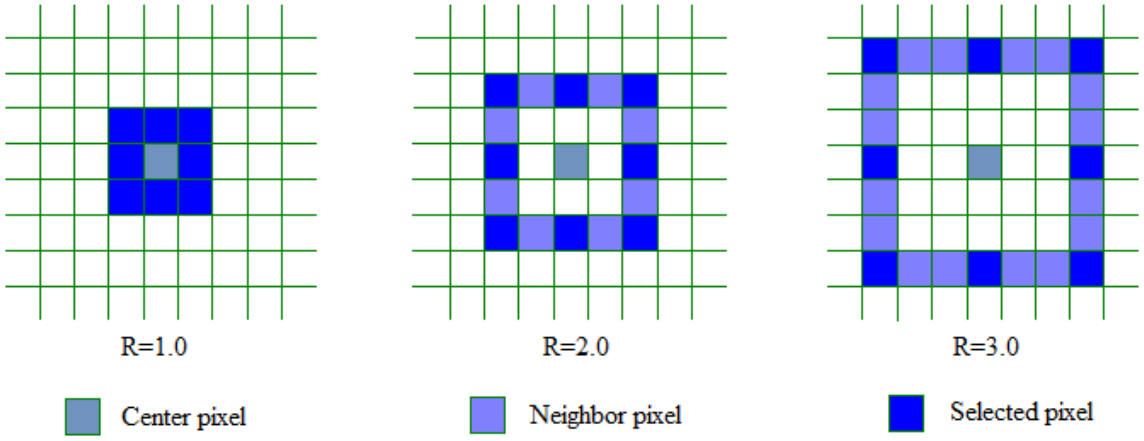


Figure 4.7: Local information used for LGP calculation: the circular $(8, 1.0)$, $(8, 2.0)$ and $(8, 3.0)$ neighborhoods.

Algorithm 2 : The eight neighborhood pixel selections

Input: Gray-scale image I_G, Center pixel $p_c(x, y)$, Radius R
Output: Set of the eight neighboring pixels that are selected (E)

```
1: E ← ∅
2: for i ← −R to R step R do
3:     for j ← −R to R step R do
4:         if (i ≠ 0 or j ≠ 0) then
5:             Add(p(x + i, y + j), E)     // Insert a new point, p, in E
6:         end if
7:     end for
8: end for
```

$$g(i_k - th) = \begin{cases} 1, & if \quad i_k \geq th \quad \& \quad (x_k, y_k) \in E \\ 0, & otherwise \end{cases} \tag{4.4}$$

Fig. 4.8 illustrates the local information utilized for the calculation of the LGP, specifically highlighting the circular neighborhoods defined by parameters $(8, 1.0)$, $(8, 2.0)$, and $(8, 3.0)$. In this context, i_k and th are the gray-level values of the selected pixel, $p_k \in E$, and the chosen threshold, respectively, while R signifies the radius of the neighborhood. The variable E encompasses the eight neighboring pixels selected through a pixel selection algorithm proposed in this study, which can be applied to any central pixel (x, y) within the image I_G. An overview of this selection process is depicted in Fig. 4.8, and the algorithm is detailed in pseudocode format in Algorithm 2. Furthermore, an example of the proposed LGP operator is presented in Fig. 4.9, specifically for the case where $R = 1.0$.

LGP on binarized images (LGP_B, LGP_W) In binary images, the pixels representing the foreground correspond to the black ink traces, while the white pixels signify the background, indicating areas without ink. To simplify the analysis, we propose two representations of the Local Gray-level Pattern (LGP) operator by extracting two distinct LGP patterns based on the black and white pixels, referred to as $GLGP_B$ and LGP_W, respectively. These two features significantly enhance the robustness of our

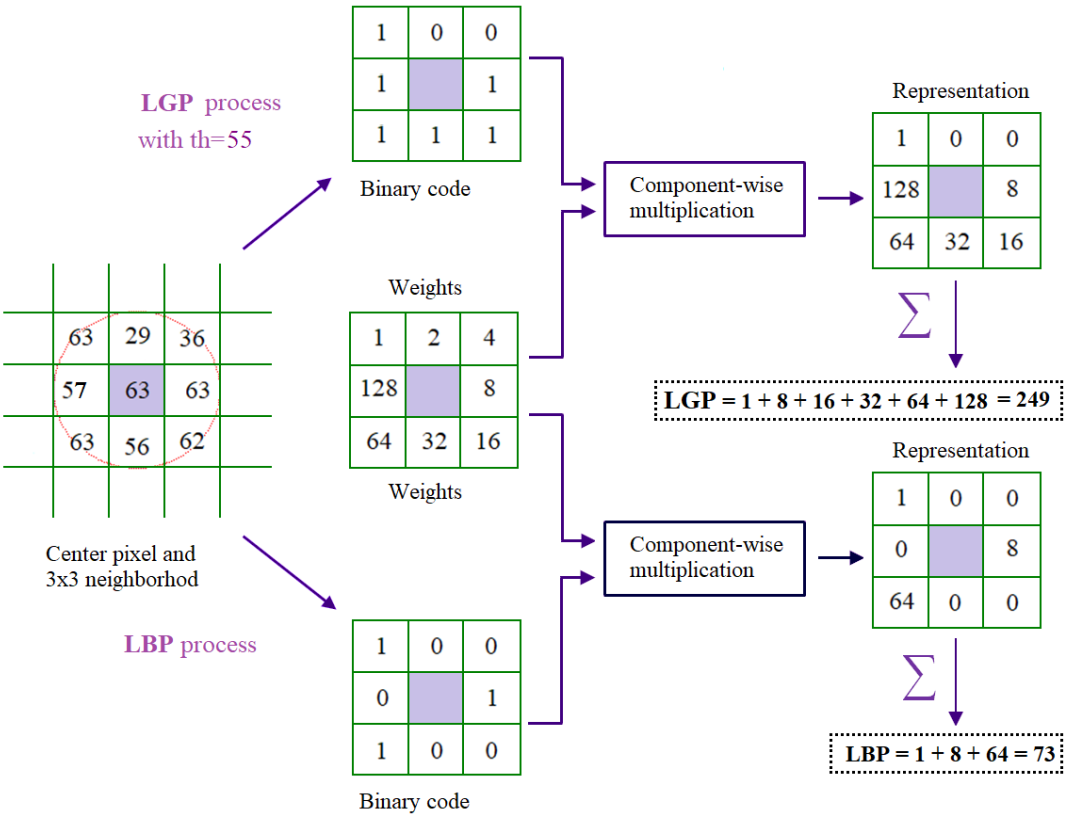


Figure 4.8: An example of the proposed LGP and the well-known LBP operators with 8 neighbors: the circular ($P = 8$, $R = 1.0$) neighborhoods.

writer recognition system, effectively addressing the $GLGP$'s susceptibility to noise and distortion. The computation of these features can be articulated as follows:

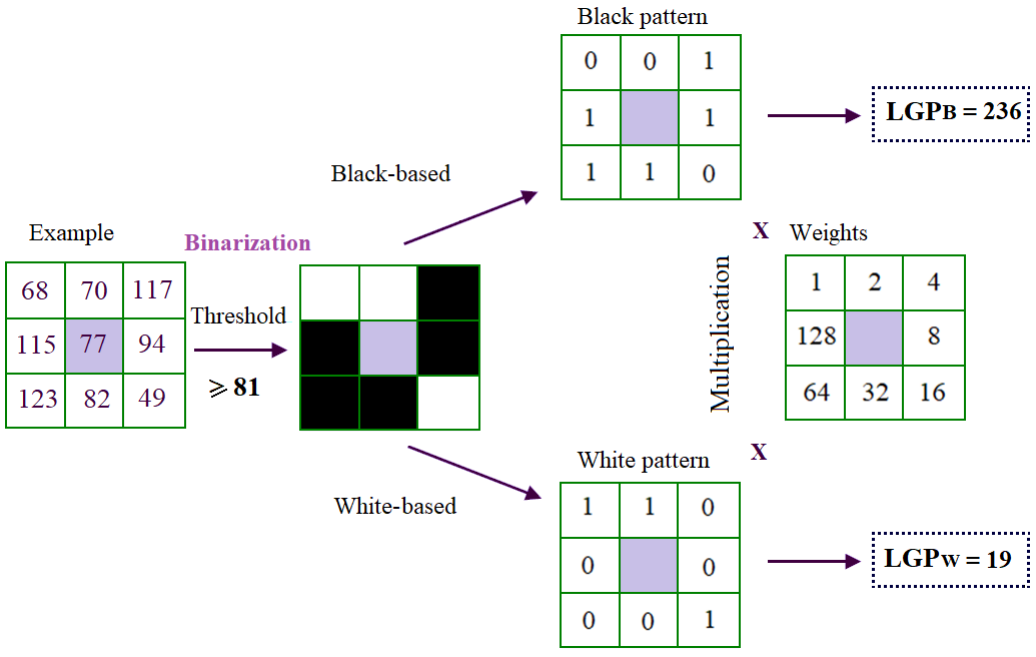


Figure 4.9: Example of LGP_B and LGP_W codes calculation.

$$LGP_{B\ E,R}(x,y) = \sum_{k=1}^{8} b(i_k)\ 2^{k-1} \tag{4.5}$$

The function $b(i_k)$ is defined as

$$b(i_k) = \begin{cases} 1, & if \quad p_c \in BL \ \ \& \ \ p_k \in BL \cap E \\ 0, & otherwise \end{cases} \tag{4.6}$$

In this context, BL and E represent the collection of black pixels found within the monochrome image I_B and the eight designated neighboring pixels for which a pixel selection algorithm is proposed in this study (refer to Algorithm 2 and Fig. 4.8), respectively. Each pixel within the subset $BL \subseteq I_B$ is assigned a value of 1 and is commonly identified as a black pixel, indicative of a black ink trace.

$$LGP_{W\ E,R}(x,y) = \sum_{k=1}^{8} w(i_k)\ 2^{k-1} \tag{4.7}$$

The function $w(i_k)$ is defined as

$$w(i_k) = \begin{cases} 1, & if \quad p_c \in WH \ \ \& \ \ p_k \in WH \cap E \\ 0, & otherwise \end{cases} \tag{4.8}$$

In this context, WH is defined as $(I_B \backslash BL)$, where WH is a subset of I_B, representing the collection of white pixels within I_B, with a value of 0 assigned to each pixel in this collection. The set E consists of the eight neighboring pixels selected for analysis (refer to Algorithm 2 and Fig. 4.8). The operations of LGP_B and LGP_W are depicted in Fig. 4.9. For each instance in the training or testing dataset, the feature vectors generated by the proposed LGP_B and LGP_W methods correspond to the probability distributions obtained through the normalization of the histograms H_{LGP_B} and H_{LGP_W}, respectively. These normalized histograms are calculated independently for each binary image, represented as the connected components of I_B, expressed mathematically as $\left(I_B = \bigcup_{k=1}^{N} CC_k\right)$, where N denotes the total number of black connected components in I_B. It is important to note that the final feature vectors for LGP_B (black pixels) and LGP_W (white pixels) are both of size 256. Further details regarding the computation of the normalized histograms can be found in reference [5].

Table 4.1 summarizes the texture-based feature vectors employed in our investigation.

4.4 Classification (Gender detection)

We derive texture-based features LGP_B (black pixels) and LGP_W (white pixels) from the black linked components prior to proceeding with classification, specifically gender detection, which constitutes the final stage of the framework. Consequently, the dissimilarity measure between two distinct handwritten samples, as indicated by their feature histogram vectors, was employed to assess intraclass and interclass variations in the biometric writing traces.

Table 4.1: Summary of texture-based feature vectors used in our study, N is the number of connected-components.

Feature	Description	Dim.	Computed from	Observation
LGP_B	Local Gray-level Value Pattern (LGP) (Black pixels)	256	$I_B = \bigcup_{k=1}^{N} CC_k$	Texture-based
LGP_W	Local Gray-level Value Pattern (LGP) (White pixels)	256	$I_B = \bigcup_{k=1}^{N} CC_k$	Texture-based
LGP_{BW}	Local Gray-level Value Pattern (LGP) (B/W pixels)	256	$I_{BW} = \bigcup_{k=1}^{N} CC_k$	Texture-based

4.4.1 Dissimilarity measure

In order to classify the gender of writers, it is essential to calculate the dissimilarity of feature histograms through a distance metric. We investigated and evaluated various established distance metric alternatives, including Euclidean, Chi-square, Manhattan, Bhattacharyya, and Minkowski (up to order 5). Following tests conducted on the HHD widely used database, we determined that the Chi-square (χ^2 for individual features) and Manhattan (Mh for feature combinations) distance proved to be the most effective and is utilized in this research.

In general, the χ^2 distance function is highly efficient and straightforward, making it ideal for probability distribution characteristics with tiny values [5]. Assume S_1 and S_2 are two handwritten samples, and $X = (x_1, x_2, \ldots, x_{size})$ and $Y = (y_1, y_2, \ldots, y_{size})$ denote the feature vectors (histograms or distributions) that will be compared. The Chi-square distance is used to compare the dissimilarity of two feature vectors (see Eq. 4.9).

$$d_{\chi^2}(X, Y) = \sum_{i=1}^{size} D(x_i, y_i) \tag{4.9}$$

The function $D(x_i, y_i)$ is defined as

$$D(x_i, y_i) = \begin{cases} 0, & if \quad x_i = y_i = 0, \quad i = 1 \ldots size \\ \dfrac{(x_i - y_i)^2}{(x_i + y_i)}, & otherwise \end{cases} \tag{4.10}$$

The dissimilarity of each feature vector (i.e., LGP_B and LGP_W) is represented by $size$, and i is an index to the elements of X and Y.

4.4.2 Gender classification using k-NN

Classification is carried out using k-Nearest Neighbors (k-NN) and Chi-Square (χ^2) and Manhattan (Mh) distances [7]. Unlike verification frameworks that rely on acceptance thresholds, the k-NN classifier performs gender classification by comparing an unknown query sample against a training dataset of known labels (Male and Female classes). Using the χ^2 (or Mh) metric, we compute the distance between the normalized histograms X and Y retrieved from the questioned sample and the reference dataset. The algorithm

maps the sample to either the Male or Female class based on the majority vote of its k most similar matches in the feature space.

4.4.2.1 Fusion techniques

The representation of handwritten document images utilizing LGP_B and LGP_W features necessitates an investigation into their amalgamation to improve performance. A multitude of experimental studies on automatic offline writer recognition using handwriting samples has distinctly demonstrated the advantages of a combined approach at both the feature and decision (or score) levels [21]. This research specifically examines the techniques for fusing score levels to enhance recognition rates. The fusion procedure is based on the generation of a scalar score which represents the resulting dissimilarity computed from dis-similarities of different texture-based descriptors using a fusion rule. As a result, various fusion rules, such as *sum*, *min*, *max*, *product*, and *median*, have been documented in the literature to evaluate probabilities [21]. Among these, the *max* fusion rule, which involves maximizing the confidence scores of base classifiers to obtain the final score for a given class. In this study, the rule combines LGP_B and LGP_W features (i.e., $LGP_{BW} = \max\{LGP_B, LGP_W\}$), produced the most promising outcomes in this study. Fig. 4.10 illustrates the score-level fusion procedure for dissimilarities of two different textural descriptors.

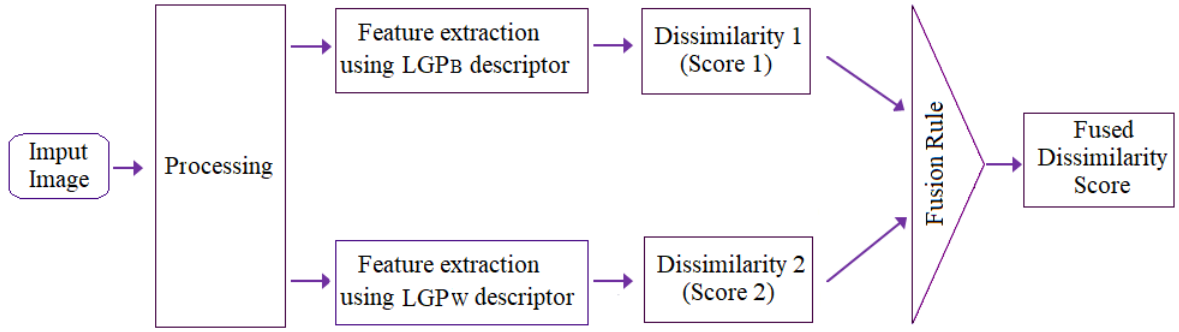


Figure 4.10: Score-level fusion procedure for dis-similarities of two different textural descriptors (LGP_B and LGP_W).

4.5 Conclusion

This chapter presented the steps necessary to extract discriminative features from handwritten document images. We described the preprocessing pipeline, which includes binarization using Otsu's algorithm, connected component extraction, and the elimination of small components. Two feature extraction methods were proposed: LGP_B (black pixels) and LGP_W (white pixels). Both methods extract features from connected components. The effectiveness of the system hinges on the selection of descriptors, distance measures, and score fusion strategies. Using multiple descriptors with the appropriate fusion rules increases the system's robustness and improves its verification accuracy. The next chapter will present the experimental results obtained using these features.

Bibliography

[1] M.N. Abdi and M. Khemakhem. A model-based approach to offline text-independent arabic writer identification and verification. *Pattern Recognition*, 48(5):1890–1903, 2015.

[2] S. Al-Máadeed and A. Hassaïne. Automatic prediction of age, gender, and nationality in offline handwriting. *EURASIP J. Image Video Process.*, 2014:10, 2014.

[3] F. Alaei and A. Alaei. Review of age and gender detection methods based on handwriting analysis. *Neural Computing and Applications*, 35:23909—-23925, 2023.

[4] B. Arazi. Handwriting identification by means of run-length measurements. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(12):878–881, 1977.

[5] T. Bahram. A texture-based approach for offline writer identification. *Journal of King Saud University - Computer and Information Sciences*, 34(8, Part A):5204–5222, 2022.

[6] T. Bahram. Offline writer identification using lcp-iwsl and lblp descriptors for detecting texture information. *Multimedia Tools and Applications*, 84:40091–40120, 2025.

[7] T. Bahram. An improved text-independent writer recognition framework using handcrafted descriptors. *Journal of King Saud University - Computer and Information Sciences*, 38(6):390, 2026.

[8] T. Bahram, A. Benyettou, and D. Ziadi. A set of features for text-independent writer identification. *International Review on Computers and Software (I.RE.CO.S.)*, 56(2):898–906, 2016.

[9] A. Bennour, C. Djeddi, A. Gattal, I. Siddiqi, and T. Mekhaznia. Handwriting based writer recognition using implicit shape codebook. *Forensic Science International*, 301:91–100, 2019.

[10] A. Bensefia, T. Paquet, and L. Heutte. A writer identification and verification system. *Pattern Recognition Letters*, 26(13):2080–2092, 2005.

[11] D. Bertolini, L.S. Oliveira, E. Justino, and R. Sabourin. Texture-based descriptors for writer identification and verification. *Expert Systems with Applications*, 40(6):2069–2080, 2013.

[12] A. Briber and Y. Chibani. Open writer identification from handwritten text fragments using lite convolutional neural network. *International Journal on Document Analysis and Recognition (IJDAR)*, 27(4):529–551, 2024.

[13] M. Bulacu and L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):701–717, 2007.

[14] A. Chahi, Y. El merabet, Y. Ruichek, and R. Touahn. Local gradient full-scale transform patterns based off-line text-independent writer identification. *Applied Soft Computing*, 92:106277, 2020.

[15] A. Chahi, Y. El Merabet, Y. Ruichek, and R. Touahni. Writerinet: a multi-path deep CNN for offline text-independent writer identification. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(2):89–107, 2023.

[16] V. Christlein, D. Bernecker, A.K. Maier, and E. Angelopoulou. Offline writer identification using convolutional neural network activation features. In J. Gall, P.V. Gehler, and B. Leibe, editors, *Pattern Recognition - 37th German Conference, GCPR 2015*, pages 540–552, Aachen, Germany, 2015. Springer.

[17] C. Djeddi, I. Siddiqi, L.S. Meslati, and A. Ennaji. Text-independent writer recognition using multi-script handwritten texts. *Pattern Recognition Letters*, 34(10):1196–1202, 2013.

[18] A. Durou, I. Aref, S. Al-Maadeed, A. Bouridane, and E. Benkhelifa. Writer identification approach based on bag of words with obi features. *Information Processing & Management*, 56(2):354–366, 2019.

[19] S. Fiel and R. Sablatnig. Writer identification and retrieval using a convolutional neural network. In G. Azzopardi and N. Petkov, editors, *Computer Analysis of Images and Patterns - 16th International Conference, CAIP 2015*, pages 26–37, Valletta, Malta, 2015. Springer.

[20] R.C. Gonzalez and R.E. Woods. *Digital image processing*. Prentice Hall (Addison-Wesley, 3rd edition), Upper Saddle River, N.J., 2008.

[21] Y. Hannad, I. Siddiqi, C. Djeddi, and M.E. El-Kettani. Improving arabic writer identification using score-level fusion of textural descriptors. *IET Biometrics*, 8(3):221–229, 2019.

[22] Y. Hannad, I. Siddiqi, and M.E. El-Kettani. Writer identification using texture descriptors of handwritten fragmen. *Expert Systems with Applications*, 47:14–22, 2016.

[23] A. Hassaïne, S. Al-Máadeed, J. Aljaam, and A. Jaoua. ICDAR 2013 competition on gender prediction from handwriting. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1417–1421, Washington, DC, USA, 2013. IEEE.

[24] S. He and L. Schomaker. Co-occurrence features for writer identification. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 78–83, Shenzhen, China, 2016. IEEE.

[25] S. He and L. Schomaker. Writer identification using curvature-free features. *Pattern Recognition*, 63:451–464, 2017.

[26] S. He and L. Schomaker. FragNet: Writer identification using deep fragment networks. *IEEE Transactions on Information Forensics and Security*, 15:3013–3022, 2020.

[27] S. He and L. Schomaker. GR-RNN: global-context residual recurrent neural networks for writer identification. *Pattern Recognition*, 117:107975, 2021.

[28] S. He, M. Wiering, and L. Schomaker. Junction detection in handwritten documents and its application to writer identification. *Pattern Recognition*, 48(12):4036–4048, 2015.

[29] M. Javidi and M. Jampour. A deep learning framework for text-independent writer identification. *Engineering Applications of Artificial Intelligence*, 95:103912, 2020.

[30] Y. Kessentini, S. BenAbderrahim, and C. Djeddi. Evidential combination of svm classifiers for writer recognition. *Neurocomputing*, 313:1–13, 2018.

[31] A. Khalifa, S. Al-Maadeed, M.A. Tahir, A. Bouridane, and A. Jamshed. Off-line writer identification using an ensemble of grapheme codebook features. *Pattern Recognition Letters*, 59:18–25, 2015.

[32] F.A. Khan, F. Khelifi, and M.A. Tahir. Dissimilarity gaussian mixture models for efficient offline handwritten text-independent identification using sift and rootsift descriptors. *IEEE Transactions on Information Forensics and Security*, 14(2):289–303, 2019.

[33] F.A. Khan, M.A Tahir, F. Khelifi, A. Bouridane, and R. Almotaeryi. Robust off-line text independent writer identification using bagged discrete cosine transform features. *Expert Systems with Applications*, 71:404–415, 2017.

[34] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig. CVL-Database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th International Conference on Document Analysis and Recognition*, pages 560–564, Washington, DC, USA, 2013. IEEE.

[35] P. Kumar and A. Sharma. Segmentation-free writer identification based on convolutional neural network. *Computers & Electrical Engineering*, 85:106707, 2020.

[36] V. Kumar and S. Sundaram. Siamese-based offline word level writer identification in a reduced subspace. *Engineering Applications of Artificial Intelligence*, 130:107720, 2024.

[37] V. Kumar and S. Sundaram. Utilization of information from CNN feature maps for offline word-level writer identification. *Expert Systems With Applications*, 238(Part A):121709, 2024.

[38] G. Louloudis, B. Gatos, N. Stamatopoulos, and A. Papandreou. ICDAR 2013 competition on writer identification. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1397–1401, Washington, DC, USA, 2013. IEEE.

[39] G. Louloudis, N. Stamatopoulos, and B. Gatos. ICDAR 2011 writer identification contest. In *2011 International Conference on Document Analysis and Recognition, ICDAR 2011*, pages 1475–1479, Beijing, China, 2011. IEEE.

[40] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[41] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[42] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[43] L. Schomaker. Advances in writer identification and verification. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, pages 1268–1273. IEEE Computer Society, 2007.

[44] A. Semma, Y. Hannad, I. Siddiqi, C. Djeddi, and M.E. El-Kettani. Writer identification using deep learning with fast keypoints and harris corner detector. *Expert Systems with Applications*, 184:115473, 2021.

[45] M. Sethi, M. Kumar, and M.K. Jindal. Gender prediction system through behavioral biometric handwriting: a comprehensive review. *Soft Computing*, 27:6307—-6327, 2023.

[46] I. Siddiqi and N. Vincent. Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. *Pattern Recognition*, 43(11):3853–3865, 2010.

[47] P. Singh, P. Pratim-Roy, and B. Raman. Writer identification using texture features: A comparative study. *Computers & Electrical Engineering*, 71:1–12, 2018.

[48] Y. Sun, D. Fan, H. Wu, Z. Wang, and J. Tian. Offline mongolian handwriting identification based on convolutional neural network. *Electronics*, 13(1):111, 2024.

[49] X. Wu, Y. Tang, and W. Bu. Offline text-independent writer identification based on scale invariant feature transform. *IEEE Transactions on Information Forensics and Security*, 9(3):526–536, 2014.

[50] Y.J. Xiong, Y. Wen, P.S.P. Wang, and Y. Lu. Text-independent writer identification using SIFT descriptor and contour-directional feature. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015*, pages 91–95. IEEE, 2015.

[51] J. Yang, M. Shokouhifar, P.L. Yee, A.A. Khan, M. Awais, and Z. Mousavi. DT2F-TLNet: A novel text-independent writer identification and verification model using a combination of deep type-2 fuzzy architecture and transfer learning networks based on handwriting data. *Expert Systems with Applications*, 242:122704, 2024.

[52] B. Zhao, X. Cao, W. Zhang, X. Liu, Q. Miao, and Y. Li. Compnet: Boosting image recognition and writer identification via complementary neural network post-processing. *Pattern Recognition*, 157:110880, 2025.

ملخص

ملخص باللغة العربية

يتناول العمل المعروض في هذه المخطوطة مشكلة التصنيف الآلي للكاتب من خلال خط يده. يعد استخراج الميزات خطوة مهمة في تحقيق أداء جيد لأنظمة التعرف على الكاتب (التصنيف) ,الهدف من هذه الخطوة هو اختيار المعلومات الأكثر صلة لمهمة تصنيف معينة في هذه الأطروحة يتم تمثيل عينات الخط اليدوي بسمات الكتابة مثل الاتجاه والإنحناء وأثر الحبر وحجم الحرف في البداية قمنا بتطوير نظام لتصنيف جنس الكاتب من خلال التحقق من صحة ميزتين قائمتين على النسيج وهما توصيف لعينات كبيرة أو صغيرة من الكتابات يتم إجراء التصنيف باستخدام خوارزمية الجيران الأقرب ومسافة كاي تربيع أظهرت النتائج التجريبية التي تم الحصول عليها على مجموعة بيانات الخط العبري مع ثلاثة سيناريوهات تضم 68 و 59 و 45 كاتبا أن المخطط المقترح يحقق أداء لافتا من أجل تحسين أداء نظام تصنيف الكاتب لدينا قمنا بدمج الميزات أظهرت النتائج التجريبية التي تم الحصول عليها على مجموعة بيانات الخط العبري أن الأداء المدمج يتجاوز أداء الميزتين الفرديتين.

الكلمات المفتاحية: تصنيف الكاتب، نمط القيمة الرمادية المحلية، تعلم الآلة، التعرف على الأنماط، الفحص الجنائي للمستندات

Abstract

Your summary in English

The work presented in this manuscript deals with the problem of automatic writer classification from their handwriting. Features extraction is an important step in achieving good performance of writer recognition (classification) systems. The purpose of this step is the selection of the most relevant information for a given classification task. In this thesis, the samples of handwriting are represented by writing attributes like orientation, curvature, ink-trace, and size of letters (LGP). Initially, we have developed a writer gender classification system by validating two texture-based features which a characterization of large or small samples of writings. Classification is carried out using k-Nearest Neighbors (k-NN) and Chi-Square ($\chi 2$) distance. The experimental results obtained Hebrew handwriting dataset with three scenarios of 68, 59, and 45 writers show that the proposed scheme achieves interesting performances. In order to enhance the performance of our writer classification system, we have combined the features (LGPBW = Max(LGPB,LGPW)). The experimental results obtained on Hebrew handwriting dataset, the combined performance exceeds the performances of all two individual features.

Keywords: Writer classification, Local Gray-level Value Pattern, Machine learning, Pattern recognition, Forensic document examination.

Résumé

Votre résumé en français

Les travaux présentés dans ce manuscrit abordent le problème de classification de scripteurs à partir de leur écriture manuscrite. La phase d'extraction de caractéristiques est une étape très importante pour un système de reconnaissance de scripteurs. L'objectif de cette étape est la sélection des informations les plus pertinentes pour la tache de classification. Dans ce mémoire les échantillons d'écritures manuscrites sont représentés par des attributs comme l'orientation, la cour bure, l'épaisseur des traits et la taille de l'écriture (LGP). Dans un premier temps nous avons développé un système de classification de genre de scripteurs en validant deux caractéristiques texturales qui permettent une caractérisation des échantillons d'écriture de grande et faible taille. La classification est réalisée en utilisant les k plus proches voisins (k-NN) et la distance de Chi-Square ($\chi 2$). La méthode proposée à été évaluée en utilisant base de données d'écriture manuscrites en hébreu de documents de 68, 59 et 45 scripteurs où des résultats intéressants ont été enregistrés. Afin d'améliorer les performances du système, une idée intéressante consiste à combiner les descripteurs (LGPBW =Max(LGPB,LGPW)). Les résultats expérimentaux obtenus sont nettement supérieures pour la combinaison que pour les caractéristiques individuelles.

Mots clés: Classification du scripteur, Modèle de niveau de gris Local, apprentissage automatique, reconnaissance de motifs, Examen légal de document.