People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Larbi Tebessi University - Tebessa

Faculty of Exact Sciences and Sciencesof Nature and Life

Department of Mathematics and Computer Sciences

Thesis submitted in fulfilment of the requirements for the degree of :

MASTER

In

Domain: Mathematics and Computer Sciences

Sector: Computer Science

Option: Networks and IT security

Theme

Contribution to the Analysis of Handwriting in Medieval Documents

Presented by:

Mouhammed Ibrahim Jadla

In front of the jury composed by :

Mr. Lakhdar Laimech	*Associate Professor*	*Larbi Tebessi University*	*President*
Mr. IsmailHadjaj	*Assistant Professor*	*Larbi Tebessi University*	*Examiner*
Mr. Chawki Djeddi	*Associate Professor*	*Larbi Tebessi University*	*Advisor*

On the date of: September 29, 2020

Dedication

"To my mother , to my father and everybody who loves and supports me "

Appreciation

*“Foremost, I would like to express my sincere **gratitude** to my advisor Pr. Chawki Djeddi for the continuous support of my Master study and research, for his patience, motivation, enthusiasm, and immense knowledge.”*

Résumé

Dans ce manuscrit, un système automatique pour l'analyse et la classification demanuscrits anciens est introduit. Le système proposé comprend deux étapes principales: l'extraction de caractéristiques et la classification (Identification de scripteurs). Dans la première étape, les distributions des longueurs de segments, les distributions des directions des contours, les distributions des charnières des contours sont extraites à partir de documents anciens manuscrits. Dans la deuxième étape, nous avons utilisé les séparateurs à vaste marge (SVM) avec la stratégie un contre tous pour la classification. Les expérimentations sont menées sur un ensemble de données qui comprend 735 documents manuscrits anciens. Le taux d'identification le plus élevé est obtenu par l'utilisation d'un classifieur SVM avec les distributions des charnières des contours.

Keywords: Classification,documents anciens, distribution des longueurs de segments, distributions des directions des contours, distributions des charnières des contours, séparateurs à vaste marge.

Abstract

In this manuscript, an automatic system for the analysis andclassification of ancient manuscripts is introduced. The proposed system comprises two main steps: feature extraction and classification (writer identification). In the first step, distributions run-lengths, distributions of contour directions, distributions of contour hinges are extracted from ancient handwritten documents. In the second step, we have used support vector machines (SVM) with one against all strategy for classification. The experiments are carried out on a dataset which includes 735 ancient handwritten documents. The highest writer identification rate is achieved by the use of SVM classifier with edge-hinge distributions.

Keywords: Classification, ancient documents, run-lengths distributions, edge-hinge distributions, edge-direction distributions, support vector machines.

الملخص

في هذا العمل، تم تقديم نظام لتحليل و تصنيف المخطوطات القديمة .يتكون النظام المقترح من مرحلتين أساسيتين :استخلاص المميزات وتصنيفها (التعرف على الخطاط).في الخطوة الأولى ، يتم استخلاص توزيع قياسات أطوال المسارات، توزيع مفصلات الحافة ، توزيع ا تجاه الحافة و كذلك معاملات الانحدار الذاتي من الوثائق القديمة المكتوبة بخط اليد. خلالالمرحلةالثانية استخدمنا أجهزة المتجهات الاعتماديةمع إستراتيجية واحد ضد الكل للتصنيف . أجريت التجارب على قاعدة بيانات تحتوي على 735 وثيقة قديمة مكتوبة بخط اليد .تم تحقيق أعلى دقة من خلال استخدام مصنف أجهزة المتجهات الاعتماديةمع ميزات توزيع مفصلات الحافة.

الكلمات المفتاحية :تصنيف، مخطوطات قديمة، أطوال المسارات، توزيعات مفصلات الحافة، توزيعات اتجاه الحافة، معاملات الانحدار الذاتي، جداول الدعم الآلية

.

TABLE OF CONTENTS

Figures Table :

Tables Table :

GENERAL INTRODUCTION

1. Research Context

Handwriting is one of the most ancient inventions in the history of humanity, it was invented back in 3200 BC in the Middle East and 3150 BC in North Africa as the latest researches claims. Many historians consider the invention of handwriting as the begging of our civilization. Not only one of the most important communication toolsbut handwriting has been the only tool for conserving and transferring: information, discovered knowledge, sciences, and the creation of literature for thousands of years. In addition, even with the invention of primers, it still playing a major role nowadays. but not only that a handwriting manuscript could be decisive in the field of science and kinds of literature but it also tells a lot of its origin and its writer handwriting styles can highly give a right prediction about the manuscript motherland, the figure below shows different styles of handwriting from a country to another.

Textura:
A style devollped in northen france in 1000 to become after the style of occidental europe

Cursiva :

The cursive form 'Anglicana' developed in England from Textura to become the most widely-used book hand of the later Middle Ages in Britain and northern France. It first appears in documents in around 1260

Court hand:
Devollped and used in law courts in england in middle ages around 1100

Figure 1: *Different handwriting styles in middle ages [1]*

So ancient manuscripts especially (Figure 2) help historians and palaeographers studying handwriting styles and the evolution of handwriting they can find out when a manuscript is written and cultural and social movement of a certain nation not only that they can even find out the author of the manuscript and this is thing that we interested into in this research.

Figure2: *An Arabicmedicine ancient manuscript [2].*

So, many hidden information in manuscript can tell a lot about its author (Figure 3): identity, sex, age, nationality … it is like a fingerprint for every single person. so that's why there is major difference that is absolutely noticeable from a person's handwriting to another and that's where sciences generally and computer sciences specifically want invest getting to know the identity of a person from his handwriting will be ideal for security and biometry and it may be revolutionary for the fields of history and paleography, that's why so many universities all around the world are leading programs curios about middle ages manuscripts recognitionand many companies are sponsoring projects for the same benefits.

Not only that, but neurologist claims that handwriting has a relation with our emotions and psychological state this why they can describe a person's state throw his handwriting. Handwriting analysis can serve a variety of scientifically benefits since it is like a mirror to persons and even societies it can solve historical mysteries, improve security, and evaluate academic scientifically researches.

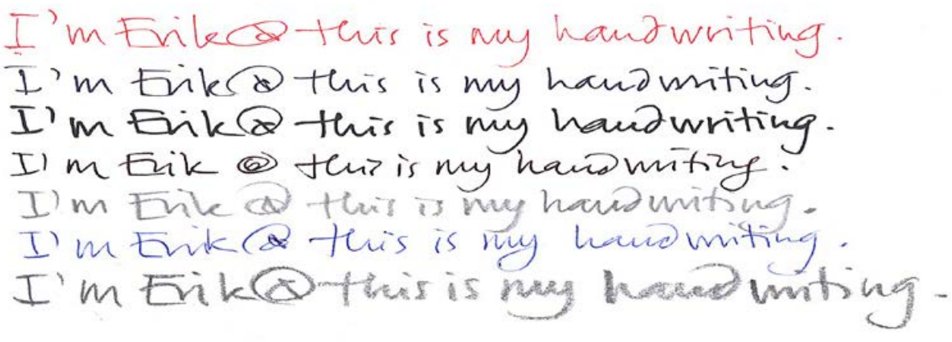

*Figure 3:*Handwriting samples from different peoples [3].

2. Research objectives

The features used in the field of writer recognition (identification and verification) are divided into two main categories: texturals and allographics. Among the textural features, we can cite: contoursdirections [4], edge hinges [4], autocorrelation [5], angle co-occurrence [6], Gabor filters [7], co-occurrences matrixes [8] and Freeman codes [9]. Allographic features are extracted by the segmentation of text into lines, words, characters, graphemes, and sometimes into small traits etc.... An important methodology is to group the graphemes and generate a codebook that will be used in order to characterize the writer. Data sets containing data from about twenty to one thousand of writers have been considered in the litterature.

The main objective of this work is to introduce a new text-independent method for writer identification using historical documents. The proposed method exploits the distributions of run-lengths, the distributions of contour- directions, the autoregressive coefficients as well as the edge-hinge distributions as features. Classification is performed using multi-class support vector machines (SVM). The experiments will be carried out on a part of the data set provided by the organizers of the “ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI)” [10].

3. Thesis organization:

This document is structured in two chapters. The first is devoted to the presentation of the main concepts and tools relating to the study undertaken.In the second chapter, we discuss in detail our conceptual choices, the implementation as well as the results obtained by the proposed system for writer identification using historical documents.

Chapter 1:Writer recognition: Fundamental Concepts and Tools

This chapter is devoted to the presentation of the fundamental concepts and tools in the field of writer recognition. It presents the mechanisms of handwriting production as well as the main sources of variability of handwriting. We adopt a categorization of the different writer classification systems according to the task envisaged (identification or verification), the dependence on the text (text-dependent or text-independent) as well as the mode of acquisition of the writing (online or offline). In the end of this chapter, we will present an overview of the main databases used in the field.

Chapter 2:An automatic system for writer identification based on historical handwriting

This chapter breaks away from the theoretical aspects discussed in the first chapter and is oriented towards the presentation of our contribution which consists of a method for writer identification from historical documents based on the characterization of the different images of texts by distributions of run-lengths, edge-direction distributions, autoregressive coefficients and edge-hinge distributions. We describe the data set used before focusing on the detailed presentation of the proposed feature extraction method.Multi-class support vector machines (SVM) are used for classification. The experiments carried out will also be presented. At the end of this chapter, the results are presented and discussed.

At the end of this document, we present our conclusions on the work we have undertaken in the field of writer identification. We also present the prospects for future extensions of the work that we have presented in this document.

CHAPTRE 1

Writer recognition: Fundamental Concepts and Tools

This chapter is devoted to the presentation of the fundamental concepts and tools in the field of writer recognition. It presents the mechanisms of handwriting production as well as the main sources of variability of handwriting. We adopt a categorization of the different writer classification systems according to the task envisaged (identification or verification), the dependence on the text (text-dependent or text-independent) as well as the mode of acquisition of the writing (online or offline). In the end of this chapter, we will present an overview of the main databases used in the field.

1.1. Introduction

The development of the fields of artificial intelligence and pattern recognition is due in large part to one of the very difficult problems of handwriting identification. The classification of modern and ancient handwriting is becoming very important these days because of the importance in scientific research. Identifying a writer from his or her handwriting applies to many areas, such as history and paleography. It is evident that the task of writer identification has become more important these days. Obviously, the number of researchers involved in this complex problem is increasing because of these opportunities. This problem is based on handwriting comes up against a number of sub-problems, such as the design of algorithms to analyze the handwriting of different individuals, to identify the most relevant characteristics allowing to characterize the different writing styles and finally choose the automatic classification techniques best suited to this problem.

1.2. Definition of handwriting

Handwriting is the particular way in which someone forms letters with a pen or pencil[11], or it is a style or manner of writing by hand, especially that which characterizes a particular person penmanship [12], but scientifically it may refer to something else for example : Handwriting is a suggests that of expressing language, similar to speech, and it conjointly leaves a lasting trace.Some say it is a physical way of expressing thoughts and concepts and a method of communicating with others.

Handwriting is a terribly advanced ability to master, one that involves linguistic, cognitive, perceptual, and motor elements, all of that ought to be coordinated into associate degree integrated fashion. though we have a tendency to take it as a right, some folks, young and old, realize handwriting terribly tough to perform and feel they have facilitate to perfect the ability. Support from those like an expert and skill is almost always appreciated[13]

1.3. Variability of handwriting

Although the individuals of a given country receive a common education, and undergo, in particular, learning to write together, the graphics they produce are extremely variable.There are three types of variability in handwriting [4]: inter-writer variability (Figure 1.1) which denotes the variation in writing style between different people, the intra-writer variability (Figure 1.2) which represents the variations in the writing of the same person over time and which depends on their physical and psychological state (variations intrinsic to each writer) as well as the inter-class variability (Figure 1.3) which denotes the variation of writing style between the two categories of writers (male and female).

The recognition of writers is only possible insofar as the inter-writer variability exceeds the intra-writer variability, in other words, the capacity of the writer recognition systems to recognize a person is essentially based on the ability to discriminate against people thanks to variability in writing.

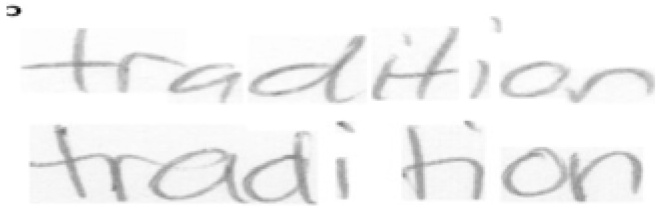

Figure 1.1: Sample of variability inter writer[14].

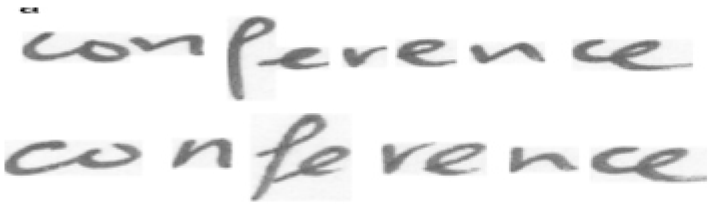

Figure 1.2: Sample of variability intra writer [14].

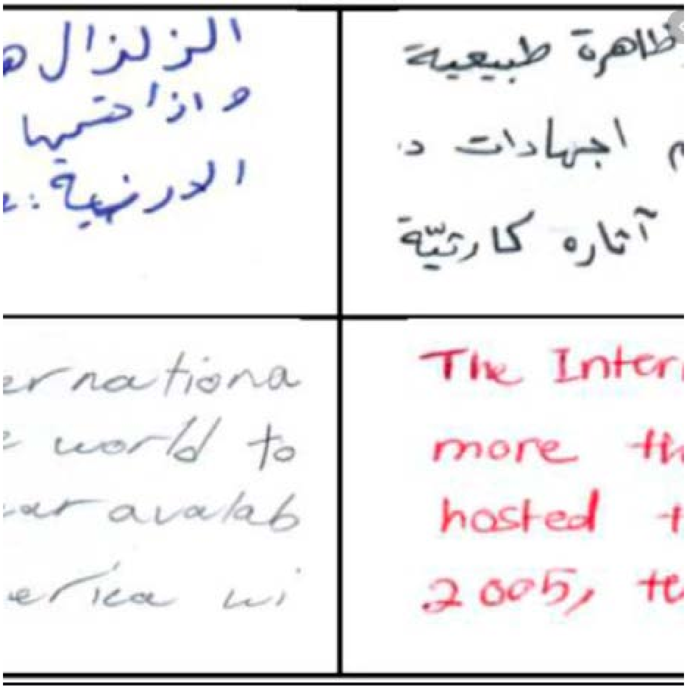

Figure 1.3: inter-class variability male on the left female on the right [15]

1.4. Categorization according to the data acquisition mode

A fundamental difference between writerrecognition systems relates to the means used for data acquisition. A distinction is made between online systems (Figure 1.4), where the trace of the writing is retrieved during its generation and offline systems (Figure 1.5) where the digitization of the trace is performed after its generation. If only an image of the writing is available, the task is performed offline; on the other hand, if the temporal and spatial data of the writing is available, the task is performed in real time (online).[16,17]

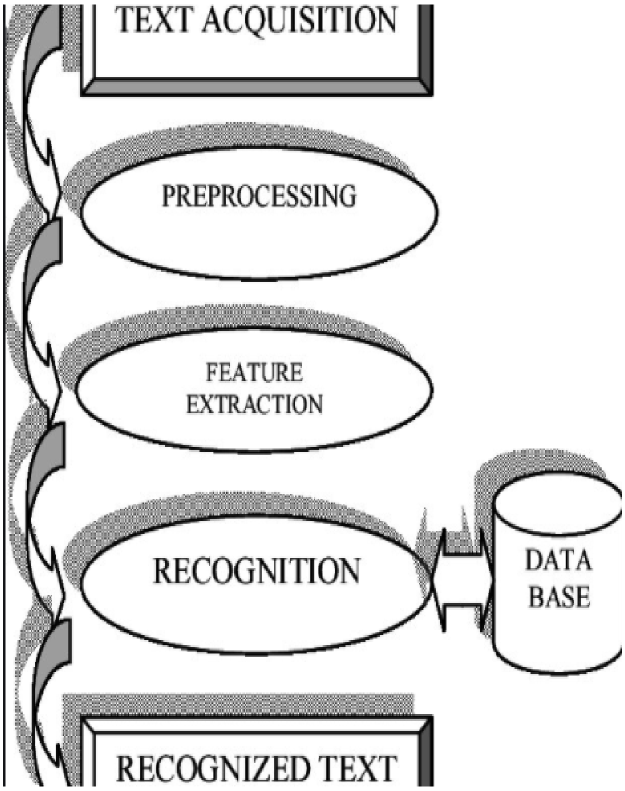

Figure 1.4 Phases of Online Handwriting Recognition System[16]

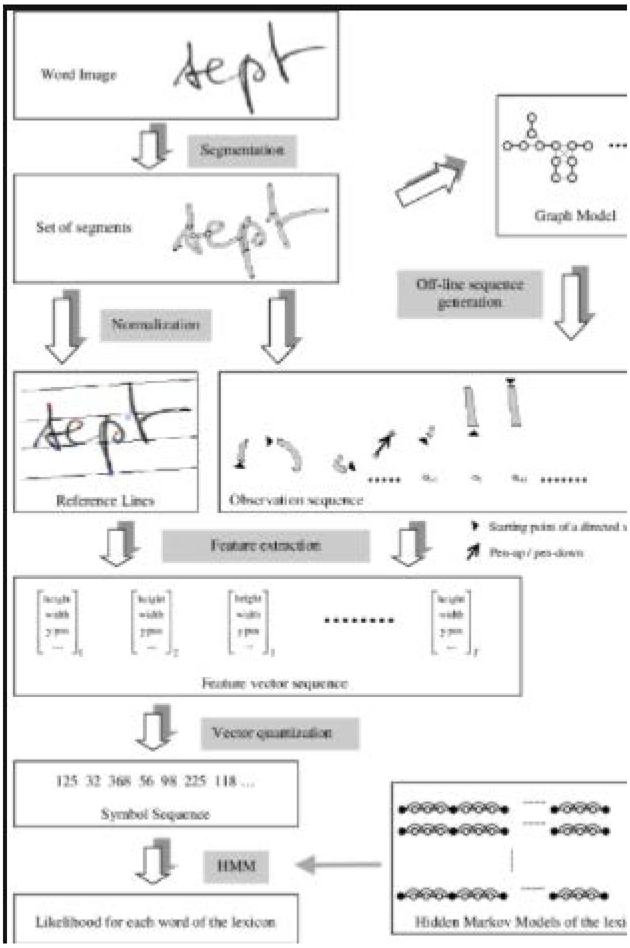

Figure 1.5. Offline handwriting recognition [17]

1.5. Categorization by text content

A second classification within the categories mentioned in the previous section is based on the level of text content. Writer recognition systems can be categorized into two broad families: text-independent systems and text-dependent systems.Text-dependent systems are more constrained and require a writer to write a particular predefined text that will be used to identify (verify) their identity.

Text-dependent systems are very similar to handwritten signature verification systems and employ the comparison between different characters or words of known semantic content. These methods therefore require the prior location and segmentation of appropriate information. This is usually performed interactively by a human user or automatically by a segmentation algorithm.

If any text can be used in order to establish (verify) the identity the writer, the task is said to be text-independent because no restriction on the subject of the written text can be made, the task is more general and, thus, more difficult to accomplish. From a user point of view, it is important to mention that a minimal amount of writing (for example, a paragraph containing a few lines of text) is necessary in order to derive stable characteristics, insensitive to the content of the text, to from the writing sample.[18]

1.6. Applications of the classification of writers

Writer recognition has been the focus of very researches as it plays very important roles nowadays, because it can be implicated in several fields including biometry and law courts, in the next subsections we describe several applications of writers recognition.

1.6.1. Adaptive handwriting recognition systems

Historically, the main areas of application for handwriting recognition were those where a large number of handwritten documents had to be processed quickly, namely mail sorting and check recognition. The performance of automatic recognition systems for these two applications is now superior to human performance because the texts are very constrained. From the 2000s, such systems have been used on more complex documents, for example to process incoming mail in large organizations "Grosicki, 2009, Brunessaux, 2014", for the indexing of handwritten personal notes "LabcomINKS, 2016" or for the indexing of historical archival documents "Bluche, 2017".

Today, thanks to extensive digitization programs, libraries and archives around the world have made available millions of manuscript documents. In addition, the adoption by a growing number of heritage institutions of the IIIF (International Image Interoperability Framework) protocol, has opened up the possibility of processing large collections of images on service platforms in the cloud thanks to standardized access to images from their collections on the web "Boros, 2019". The demand from users is now to be able to search their textual content, as is commonly done on web pages. However, some expertise is still needed to adapt and apply handwriting recognition systems.

Finally, while the performance of systems in the laboratory is now superior to the performance of a human on writings that are difficult to decipher, the challenge of handwriting recognition now lies in scaling up: knowing how to process millions of documents and very diverse writings, without having to intervene each time to retrain the systems. For this, it is necessary to develop techniques based on semi-supervised learning - from labeled and unlabelled data - for the automatic adaptation of recognition systems to new or evolving writings, without having need large annotated databases and the intervention of an engineer to control learning. True adaptive handwriting recognition technology[21]

1.6.2.Legal examination of documents

Also called Subject Matter Review [19]is the examination of potentially contentious documents in a very court of law. Its main objective is to produce proof of a suspicious or questionable document using scientific processes and techniques. proof may embody alterations, chain of custody, document harm, forgery, origin, quality or different issues that arise once a document is challenged in court. many QD exams involve a comparison of the document in question, or components of the document, against a gaggle of proverbial standards. the foremost common form of review involves handwriting within which the reviewer tries to handle issues concerning the potential author.

A document examiner is commonly asked to visualize whether or not or not degree item in question comes from a similar provide because the proverbial item (s), then gift their opinion on the matter in court as associate witness. different common tasks embody crucial what happened to a document, crucial once a document was created, or decrypting data concerning the document that was blanked, erased, or erased.

The discipline is thought by several names, including "forensic document review", "document review", "Diplomatic", "writing exam", or typically "writing analysis", though that the latter term is not typically used because it is also confused with graphology. Likewise, a document knowledgeable (FDE) is not to be confused with a specializer, and also the alternative manner around. several FDES receive intensive training altogether aspects of the discipline. As a result, they're competent to answer a good sort of queries relating to documentary proof. However, this “broad specialization” approach has not been adopted universally.

In some places a transparent distinction between the terms knowledgeable, "and" expert graphologist / examiner. during this case, the previous term refers to communication inationiners World Health Organization target non-writing exam kinds whereas the latter refers to folks that are trained only to undertake and do writing exams. Even in places where the extra general which implies is common, corresponding to North America or Australia, there are many of us World Health Organization only have specialised employment in comparatively restricted areas. As language varies from jurisdiction to jurisdiction, it is important to clarify the which implies of the title used by anyone affirmation to be a “document expert”.

1.6.3. Biometric recognition

Biometrics [22] is that the science of analyzing the physical or activity characteristics specific to each individual and allowing the authentication of their identity. In a very literal sense and in a very a lot of simplified manner, life science means that the “measurement of the human body”.There are 2 classes of biometric technologies: physiological measurements, and behavioural measurements.Physiological measurements can be morphological or biological, these square measure chiefly fingerprints, the shape of the hand, the finger, the eye (iris and retina), or the form of the face, for morphological analysis.

In terms of biological analyses, we tend to most frequently realize DNA, blood, saliva, or water employed in the medical field, for criminal investigations or maybe within the field of sport for doping controls.The commonest behavioural measures are speech recognition, signature dynamics (pen movement speed, accelerations, pressure exerted, tilt), input device writing dynamics, the way objects are used, the gait, the sound of footsteps, the gestures ...

1.6.4. Digital library

The Digital Library Federation [23] (DLF) proposes the subsequent definition: “Digital libraries are organizations that offer the resources, together with qualified personnel, to select, structure, offer intellectual access to, interpret, distribute and maintain the integrity of, and make sure the property of collections of digital works in order that they'll be simply and economically accessible to a defined community, or to a group of communities”. It is, therefore, associate degree innovation that is each technological and social since its primary goal is to concentrate on improving service to users.

The term "virtual library" has often been used in the same sense as "digital library", but it is now mainly used for libraries that have content natively in digital format.Although there are many different kinds, all digital libraries usually have three common characteristics: a collection of resources sharing the same types of encoding and delivery, the use of metadata as resource keys, and the textual expression of this metadata regardless of the format of the resource

1.6.5. Smart meeting rooms

A new generation of meeting rooms makes use of multimodal sensors to detect and capture the verbal and non-verbal behaviour of participants. This is called a smart meeting room.At a time when artificial intelligence (AI) is taking over all fields, what can it offer users of smart meeting rooms and what are the prospects for these “smart-meetings”? What are the benefits of AI during the three key moments of the meeting of connection, discussion and transcription?

Collaboration applications such as Teams, Hangouts Meet, or Zoom, for example, provide simplified solutions for organizing video conferencing meetings.With light and fast interfaces, videoconferencing applications make it possible to intelligently manage the agendas of the participants in order to find both the right time slot and the right meeting room while avoiding the many back and forth emails between all the participants, which are sometimes in different time zones.

Many video collaboration applications will offer the option of voice control to initiate a video conference. To save valuable time and reduce connection stress, imagine having to say, "Hey Assistant, start the meeting" to actually start your video conference, without even having to press a button.

A smart assistant could also provide impeccable customer service. Indeed, not all situations require enlisting the help of a human advisor. Today with the advancement of AI, machines are becoming more and more skilled at solving simple problems.

As a result, some situations with a simple framework will eventually be handled by AI-doped assistants who can, using simple questions, resolve connection barriers. And to anticipate the anxiety or reserve that some people feel when facing the machine, we can imagine these new “smart” assistants putting on a human face[24]

1.6.6. Ambient intelligence

Ambient intelligence [19] is the product of computer science which, by pushing technological limits in a disruptive manner, challenges the very concept of an information system or computer: of a processing activity exclusively cantered on the user until the end of the twentieth century, ambient intelligence aimed to govern interactions between communicating objects and humans.

Because technology makes it doable to manufacture small and omnipresent computers (nano-informatics), it opens up to most objects of daily life the power to trigger a spontaneous exchange of data, while not interaction with their user.

This thought looks to be ready to take the place of a non-literal translation of the ideas born in North America underneath the initial name of omnipresent computing, pervasive systems or perhaps impermanent computer.

1.7. Main databases

The handwritten text image databases are the focal point in the evaluation of handwriting analysis and recognition systems. They provide an effective means for the unification and comparison of work carried out within from different research teams around the world. In the next section,we present the main databases used for writer classification.

1.7.1. CEDAR database

The CEDAR [25] database was developed at the University of Buffalo, and it is considered as one of the first large databases developed for the classification of Latin handwritten scripts and more particularly, for the recognition of writers. It is composed of 4,701 images of handwritten texts written by 1,567 different writers who were selected to be representative of the population of the United States of America. Each writer copied three copies of the CEDAR letter, scanned with a300 dpi resolution. This letter is a document that contains 156 words, from a lexicon of 124, which includes all characters (letters and numbers). The document has been carefully designed to contain each letter of the alphabet in upper case in the initial position of a word and in lowercase in initial, intermediate and final position of a word

1.7.2. IAM database

The IAM database [25] consists of handwritten pages corresponding to English texts taken from the 'Lancaster-Oslo / Bergen' corpus (LOB). The corpus is a collection of texts which consists of approximately one million word instances. it is understood, in its first version, 556 images of texts produced by around 250 different writers, then, it was extended to contain 1539 images of texts produced by 657 different writers. In view of its availability to the public, its flexible structure, and the large number of writers it contains, the IAM database has been commonly used to the recognition of Latin writers as well as for the recognition of handwriting.

1.7.3. RIMES Database

RIMES [25] is a relatively new database, it includes 5600 mails manuscripts written in French, such as those sent by private individuals to companies or administrations. Each letter containing 2 to 3 pages, the database represents 12,600 original pages (in A4 format). 1300 volunteer scripters participated in the constitution of the RIMES database by drafting the letters with their own wording. This database can therefore be considered realistic because the formulation of the content documents was free. It should be noted that the RIMES database was used for the recognition of writers by Siddiqi and Vincent as well as for the recognition handwritten French writing during the competitions that took place as part of the ICDAR 2009 and ICDAR 2011 conferences.

1.7.4. BFL Database

The BFL [25] (Brazilian Forensic Letter) database is made up of 315 writers, with three samples per writer, for a total of 945 images. The samples were collected from undergraduate students in three different sessions over a period of a month. The texts were collected on a white sheet of A4 size, without squares or lines, then scanned in grayscale at 300 dpi (3760 × 2448 pixels). Each writer was allowed to use one's own pen, which means that many pen models have been used. The text is concise (131 words in Portuguese), and complete in the sense that it contains all characters (letters and numbers) and certain combinations of characters of interest. This makes it also suitable for the classification of writings in text independent mode. Note that this database has been the subject of work on the recognition of writers.

1.7.5. CVL Database

The CVL [25] dataset is a public dataset created in 2013, it is used for writer retrieval, writer recognition yet as word-spotting. The info consists of 1609 texts from 311 totally different writers, twenty seven writers contributed by five documents every so the remaining 284 writers contributed seven documents every. for every text, a RGB color image (300 dpi) as well as written text and a written sample of an equivalent text is offered. The CVL info consists of pictures with texts German and English cursive manuscripts that are chosen from literary works. This info was used for the popularity of writers by Fiel and Sablatnig .

1.7.6. KHATT Database

KHATT [25] dataset is also new dataset that contains images of Arabic texts, it is used for the identification of writers, it had been made public in Sept 2012. The KHATT dataset contains 4000 images of paragraphs, these images contain scanned texts at totally different resolutions (200, 300 and 600 dpi). A thousand of writers of variousages and origins from eighteen different countries participated in the creation of this dataset.

1.7.7IAM-HistDB

IAM Historical Handwriting Database (IAM-HistDB) contains medieval manuscripts of the epic poem Parzival by Wolfram von Eschenbach, one of the most important epics of the European Middle Ages. There are many manuscripts of the poem, written withink on parchment or paper, which vary in style of writing and language dialect. There are approximately 600 manuscript pages in the IAM-HistDB dataset for which a transcription is available. The transcriptions were obtained by the University of Bern 's Department of German Language and Literature using the TUSTEP 3 method to handle Latin and non-Latin transcripts. [26]

1.8. Existing systems

Since there are many researches on the field that results many classification systems, in this section we will describe and define several one such as wanda and cedar , and their uses, advantages, brief history, and surely how much they are applicable and practicable.

1.8.1. WANDA system

WANDA [4,25] is a software tool used for the comparison of handwriting based on well-defined characteristics and on the knowledge and experience of an expert in writing .This system allows the digitization of writing samples thanks to a optical device, such as a scanner or digital camera or using a pen or electronic tablet. WANDA includes a set of tools for processing images from documents, extracting characteristics manually, semi-automatically or automatic, database management, and finally writer identification.

WANDA has an interactive measurement module, called WAM Throughthis module, manual characteristics can be extracted, they include height, the thickness, the inclination, the mean line spacing in addition to measures characterizing certain aspects on the shape of the curls. It is also possible to extract characteristics in a way completely automatic, we can cite for example: autocorrelation functions which are used to detect the presence of regularities in the writing, the length distributions of horizontal and vertical segments, the distributions of the ink density present at the end of the lines, the distributions of the directions of the contours as well as the distributions of the hinges of contours.

1.8.2. CEDAR-FOX system

CEDAR-FOX [4,25] is a computer system for analyzing entries written for legal applications. this technique was created by a team of analysis at the Center of Excellence for Document Analysis and Recognition (CEDAR) of the University of Buffalo (New House of York, USA) unitedly with the National Institute of Justice (NIJ). CEDAR-FOX has numerous functions that are utilized by writing consultants, among these functions, we are able to cite the popularity of handwriting, the popularity of scripters, signature verification, image processing, script segmentation and a number of other analysis modalities. As a document management system for legal and judicial analysis, CEDAR-FOX offers users 3 major functionalities. It are often used as a system document analysis or to create a digital library of written documents legal, and eventually, as a database management system for the rummage around for documents and therefore the recognition of scriptwriters. because it is AN interactive system for analysing documents, a graphical interface is provided, it permits you to scan or load a image of a handwritten document. The system can initial mechanically extract characteristics based on image process and recognition techniques of written documents. The user will then use the tools provided to perform reviewing documents and extracting metrics from them. These tools embody options resembling image selection, image sweetening, and description show. once verification of the writer, once a better-known document is compared to the thought of document, CEDAR-FOX analyzes the latter and calculates a similarity score that is employed to make your mind up whether the 2 documents are from identical author or not. additionally, the system has the flexibility to find out from better-known documents for the task of corroborative the writer. The system needs a minimum of 4 samples of identical author thus that he are often trained on writing the author. once coaching the system, the documents in question will then be compared to better-known samples of the author for verification. CEDAR-FOX uses a process practicality by batches for the identification of the author from a group of better-known documents, the system will realize the documents most almost like this document. CEDAR-FOX has 2 modes of signature verification. within the initial mode, a better-known signature and another unknown (questioned signature) are compared; the system then generates a score indicating whether or not the unknown signature thought of is authentic or not. The second mode of signature verification corresponds to the flexibility to find out from better-known signature samples. The system recommends that a minimum of 4 better-known samples ought to be used. once coaching the system on language the better-known author, unknown signatures can be compared and a score is generated, it indicates whether or not the unknown signatures thought of are possible to be authentic or not. CEDAR-FOX has been tested by the North American country Border Services Agency, United States intelligence agency u. s., the United States Bureau of Investigation (FBI) and is presently being evaluated by the urban center local department. additionally, the system has been authorised by the

Netherlands rhetorical Institute (NFI) and by the North American country Border Services Agency. an endeavor version of CEDAR-FOX is obtainable for transfer [CEDAR]. An Arabic version of CEDAR-FOX, referred to as CEDARABIC, is additionally obtainable. could be a totally machine-controlled and functional system that uses handwriting as a biometric symbol. FLASH ID is software utilized by the Federal analysis workplace (FBI) of the u. s. of America (USA). FLASH ID extracts, in an exceedingly utterly automatic, graphic knowledge from handwritten documents, then analyzes these knowledge exploitation better-known applied mathematics strategies and then classifies the documents into

according to the alikeness of the written documents. The system extracts the characteristics graphics from the characters severally and these characteristics are often enthusiastic about the writer. FLASH ID is ready to act on any sample unknown written document and returns the nearest worth in its info of written documents, it therefore provides the nearest match to the sample of the written document thought of. FLASH ID technology is language freelance. The FLASH ID system digital computer interface at the laboratory of the government department of analysis (FBI)

1.8.3. Script system

The Script system [4,25] is the result of an important collaboration between the NIFO (Netherlands Institute for Forensic Examinations and Research) and the University of Delft and TNO (Dutch Organization for Applied Scientific Research). This system was intended to replace the manual analysis process of handwritings by a technique assisted by computer, save important information in a file and find quickly all samples similar to the sample entering the system.

The SCRIPT system uses completely independent attributes for the characterization of handwritten documents, this independence also exists with regard to disguise, the effect of alcohol or drugs as well as the education system. These attributes are based mainly on measurements of angles and distances. It is important to note that the SCRIPT system has not been specially developed for use by experts in handwritten writings.

At the end of this section, it is more than necessary to note that the performance of the different systems just presented are not fully published; the risk of error exists, an error which can become crucial in some cases. The developed systems are not therefore not intended for an ordinary user but for an expert in writing, so that the handwriting analysis results can ultimately be validated by the expert. It is also important to note that no attempt has been made to develop a system for the recognition of the genre of a writer, this is due to the youth of this field which remains far from the concerns of computer scientists for unknown reasons.

1.9. Conclusion

In this chapter, we first presented the writing which represents the basic entity on which the work of this thesis is based, we then described the different types of variations necessary for a writer recognition system can identify an individual. After having opted for a categorization of the classification systems of writers according to the task to be performed (identification or verification), the dependence on the textual content of the writing samples to be used (dependent or independent) as well as the mode adopted for the acquisition of writing. We ended this chapter with a conclusion preceded by the presentation of the main data sets used in the literature.

CHAPTER 2

AN AUTOMATIC SYSTEM FOR WRITER IDENTIFICATION BASED ON HISTORICAL HANDWRITING

This chapter presents our main contribution, which consists of a study on writer identification from historical documents. The proposed method is based on extracting a set of textural features from samples of historical documents and training a classifier so that they can identify the writer of a questioned document. Attributes like orientation, curvature, and texture are estimated by calculating run-length distributions, edge-direction distributions, edge-hinge distributions and autoregressive coefficients. Classification is performed using support vector machines (SVM) with the one against all strategy. The proposed method was evaluated using a dataset containing more than 700 historical documents where interesting results were recorded.

2.1. Introduction

The last few years have seen a significant increase in research in different fields of biometrics. Notable advances in this area have resulted in numerous biometric modalities including recognition of the palmar vein [27], facial recognition [28], palm recognition [29], fingerprint recognition [30], DNA recognition [31]. These recognition methods can be classified into physiological and behavioural biometrics [32]. Physical or physiological biometrics use certain physical characteristics of individuals for their authentication, while behavioural biometrics is based on behavioural traits learned and acquired over time that are exploited for authentication purposes.

Despite the considerable development in different biometric modalities, writer identification from handwritten documents has remained among the most popular and widely accepted authentication mechanisms in courts of justice. Automatic writer identification has remained an interesting pattern recognition problem for several decades [33] and recent advances on this problem have been summarized in a number of studies [34].

Writer identification is the process of identifying a writer from a set of several writers. The system must then have previously learned handwriting samples from each writer. While identifying the writer is part of the same issue as handwriting recognition, it does not seem to pose the same type of difficulty. Indeed, the task of identification must take advantage of the variability of the writings in order to discriminate them, while the task of recognition must, on the contrary, manage to overcome this variability between the writers in order to identify the textual message written. The potential applications of writer identification are numerous.

This chapter aims to develop a fully automatic writer identification system that will potentially assist forensic experts and paleo graphs. The proposed system is designed to work on offline historical documents. The developed system can potentially be deployed in forensic laboratories, police stations and other judicial institutions.

This chapter is organized as follows. In the next section, we describe the database. We then present the description of the classification techniques used in section 2.3, followed by the description of the characteristics proposed in section 2.4. The experimental results and their analysis are presented in section 2.5, while the last section concludes this chapter.

2.2. Database

The proposed method is evaluated on the dataset provided by the organizers of the international competition on writer identification from historical documents (Historical-WI) [35] which took place in conjunction with the ICDAR 2017 conference. This set includes 1182 images of documents written by 394 different writers at the rate of three documents per writer. These documents were written between the 13th and the 20th century and mainly contain correspondence in German, Latin and French. The dataset used belongs to the Basel University Electronic Library [19] which includes a large database of 140,000 images. This database contains not only images of documents, but also drawings, musical scores, photographs, blank pages, envelopes, small pieces of handwritten pages and technical drawings. Document images mostly consist of matches, but some notes are included as well. The documents are written in different languages. Most of the time, German and French are used, but the Arabic script is also found in the database. All images used in this brief are scanned with a resolution of 300 dpi and are stored in .JPG format.

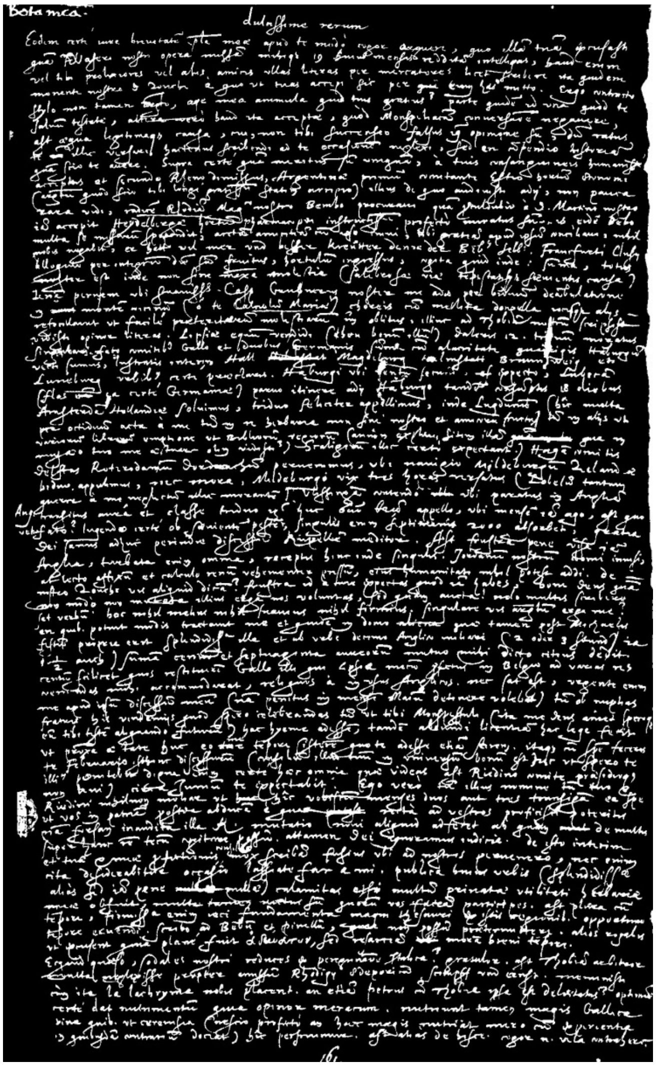

Figure 2.1: Sample of an historical document extracted from the data set used.

2.3. Classification

Once the images of documents to be compared are represented by their features, we proceed to use these features vectors for training or testing. Training and classification is performed using Multi-class support vector machines (SVM). A brief introduction to this classifier is given in the following subsection.

SVM is a supervised machine learning method, this linear method is based on the concept of “hyperlan”, which is linear classifier that that allows us to distinguish between every possible two classes , the goal of this function is to predict the class that the new object belong to ,the most optimal hyperlan is the one that realises the maximum marge between all points of two classes . For multi-class SVM we decided to choose one against all algorithm over one vs one because of the number of hyperlans generated , number of hyperlan generated in one vs one algorithm equals n(n-1) considering n is the number of classes while it will equal to n with the one vs all algorithm. We followed a method of building matrixes from tables that we got from the previous extracting methods we build a big matrix called Xapp from 3 samples from each class an the other 2 we build from them a big matrix called Xtest which will be used on testing what the svm learned from the learning phase .A one vs all SVM multi-class builds a hyperlan every single time we pass a new class , the hyoerlan separates the class from the rest of elements and so on until it builds all hyperlan. For every writer of the 147 we used 3 samples for learning and 2 for the test, we tested this SVM using « SVM and Kernel Methods Matlab Toolbox

2.4. Features extraction

The proposed writer identification system is based on a set of textural features extracted from images of historical documents. These features are described in the following subsections.

2.4.1 Run length distributions

We characterize the different documents by calculating the run-lengths distributions. These features are determined from a binary image of the handwritten document, where the black pixels correspond to the ink trace and the white pixels correspond to the background. Run-lengths are calculated directly from the full document image. In order to calculate the run-lengths, the image is scanned in the four main directions: horizontal, vertical, left diagonal and right diagonal. The normalized histogram of these segment lengths is interpreted as a distribution probability in order to characterize the document.

We start by defining a "run" as a set of consecutive pixels with similar gray levels, in a given direction. If AiAj is a segment composed of pixels Ai, Ai+1, ..., Aj-1, Aj of Identical color, pixel Ai-1 must be of a different color from that of pixel Ai, the color of the pixel Aj must also be different from that of pixel Aj+1.

We then define a statistical method of characterizing the texture which consists in counting the number of the segments of the same intensity in a given direction and representing the results in a matrix called the run-length matrix P.

The proposed feature extraction method is illustrated by an example in Figure 2.2. We consider an 8 × 6 image with two colors C = {0, 1}. This image represents the number '2'.

0	1	1	1	1	0
1	0	0	0	0	1
1	0	0	0	0	1
0	0	0	0	1	0
0	0	0	1	0	0
0	0	1	0	0	0
0	1	0	0	0	1
1	1	1	1	1	1

2	1	2	2	1	1
9	2	0	0	0	1

3	2	3	2	1	0	0	0
11	4	0	0	0	0	0	0

7	3	2	1	0	1
15	2	0	0	0	0

4	2	3	3	0	0
9	0	0	1	0	1

Figure 2.2: Graphical description for run-length method. [4]

Each element of the matrix indicates the number of times the image contains a segment of length {1, 2, 3, 4, 5, 6} in the 45 °, 135 ° and 180 ° directions and segments of a length of {1, 2, 3, 4, 5, 6, 7, 8} in the 90 ° direction. The first element of the first row of the matrix corresponds to the number of times the color 0 appears in isolation, the second item is the number of times color 0 appears in two-pixel segments and so on ... The following line captures the same information from the image for color 1.[4]

2.4.2 Edge direction distributions

This contour distribution is calculated very quickly using the outline representation, with the added advantage that the influence of ink trace width is eliminated. The distribution of the contour directions is extracted by considering the orientation of the local fragments of the contour. A fragment is determined by two edge pixels (xk, yk) and (xk + ω, yk + ω) taken at a certain distance ω. The angle that the fragment makes with the horizontal is calculated using:

$$\varphi = arctan(\frac{y_{k+\omega} - y_k}{x_{k+\omega} - x_k})$$

When the algorithm runs on the contour, the histogram of the angles is constructed. This histogram is then normalized to a probability distribution which gives the probability of finding in the image a contour fragment oriented at each φ. The φ angle lies in the first two quadrants because, without online information, we don't know with what inclination the user signed. The histogram is spread over the interval 0 ° -180 ° and divided into n sections. Therefore, each section weighs 15 °, which makes the proposed features sufficiently robust. Figure 2.3 presents a graphical description of the edge direction distributions.[4,36]

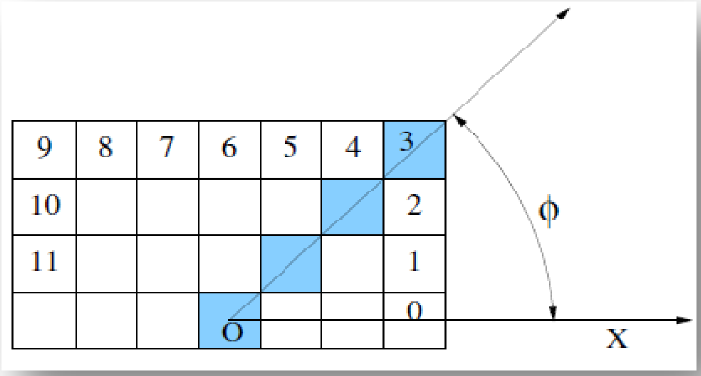

Figure 2.4: Graphical description of the edge direction distributions.[4]

2.4.3 Edge-hinge distributions

In order to capture the curvature of the contour.[4,37], as well as its orientation, the edge-hinge distributions are used. This feature extraction method is similar to the one described above, but it is a bit more complex. The central idea is to consider in the neighborhood, not one, but two edge fragments emerging from the central pixel and then to calculate the joint probability distribution of the orientations of the two fragments. Figure 2.5 shows a graphical description of the edge-hinge features.

Table 2.1 summarizes, for each of the used features, the corresponding number, description and dimension.

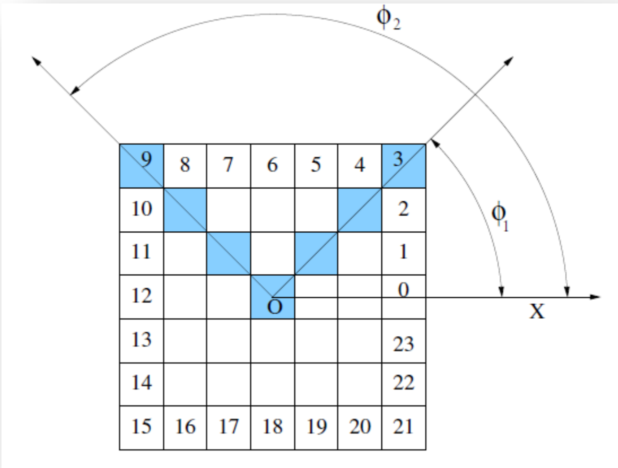

Figure 2.5: Graphical description for edge-hinge distributions.[4]

Feature	Description	Dimension
f1	Edge-hinge features	3136
f2	Edge-directions features	40
F3	Run-length on Black pixels	400
f4	Run-lenght on White pixels	400

Table 2.1: An overview of the features implemented and their dimensionality.

2.5. Experimental results

In this section, we present and analyse the performance of the proposed features in writer identification using historical handwritten documents.

In this experiment, the two first images from each writer are used for training while the other is used in the test. Table 2.2 shows the overall identification rates recorded using Multi-class support vector machines (SVM) based on the results of the conducted experiment.

Features	**SVM (%)**					
	Top 1	**Top 2**	**Top 3**	**Top 4**	**Top 5**	**Top 10**
f1	50.5051%	56.5657%	60.101%	62.8788%	63.6364%	68.9394%
f2	34.8485%	41.9192%	47.2222%	50.5051%	52.7778%	59.8485%
f3	30.303%	34.596%	38.1313%	41.4141%	43.9394%	51.7677%
f4	33.0808%	40.9091%	47.9798%	51.0101%	55.5556%	62.6263%

Table 2.2. Identification rates achieved with individual features.

The analyses of the results presented in the previous table can lead to the observations below:

- The edge-hinge features (f1) gives the best result 68,94% and it is most efficient method.
- The white run-lengths (f4) are more informative than the black run-lengths (f3).
- The black run-lengths (f3) alone are not very discriminative.
- The edge-directions distributions (f2) are the least efficient features.

2.7 Conclusion

This work aimed to present a method for writer identification from historical documents. We used a set of textural features that showed promising results on a database of historical documents. The evaluations were carried out on a database containing more than 700 samples of historical handwritten texts. The results obtained for writer identification are encouraging. They reflect the effectiveness of textural features on historical documents. Another interesting aspect of this study is the evaluation and combination of a number of textural methods on this database. In all cases, some features have been shown to perform better than other features.

GENERAL CONCLUSION

This work addressed the problem of writer identification from historical documents using run-length distributions, edge-direction distributions, edge-hinge distributions as well as autoregressive coefficients as features. Support vector machines (SVMs) with the one-on-all strategy were used for classification.Samples of handwritten documents extracted from a large historical database were used for the performance evaluation of the proposed system. The proposed system has achieved very encouraging performance especially when using Edge-hinge distributions.

Further studies on this subject will be intended to introduce additional features and then apply a feature selection mechanism to find out which features are the most discriminating for this problem and for similar problems. It is necessary to remember that the performance of the proposed system does not depend only on the used classification technique used only, but also on the chosen features.

In this context, it would be very interesting to exploit the combination of the features proposed in this document with those of the state-of-the-art in order to improve the performance of the proposed system. For classification, we believe that it would be interesting to consider the use of other classification techniques than those we have adopted in this document. It would also be very interesting also to consider and experiment with the possibilities of combining classification techniques and to consider larger databases than that used in the context of this work.

Bibliography :

[1] The university of Nottingham, Manuscripts and special collections

[2] Islamic medical manuscripts by Abū Bakr Ḥāmid Ibn Samajūn the national library of medicine, Maryland US

[3] http://www.themaninblue.com/articles/handwritten_typographers/

[4] Chawki Djeddi , Labiba Souici-Meslati, Abdel Ennaji Writer Recognition on Arabic Handwritten Documents

[5] Shannon, Benjamin & Paliwal, Kuldip. (2006). Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition. Speech Communication. 48. 1458-1485. 10.1016/j.specom.2006.08.003.

[6] Alasadi, Tawfiq & Baiee, Wadhah. (2014). Analysis of GLCM Feature Extraction for Choosing Appropriate Angle Relative to BP Classifier. IOSR Journal of Computer Engineering (IOSR-JCE). 16. 65-69. 10.9790/0661-16176569.

[7] W. Li, K. Mao, H. Zhang and T. Chai, "Selection of Gabor filters for improved texture feature extraction," 2010 IEEE International Conference on Image Processing, Hong Kong, 2010, pp. 361-364, doi: 10.1109/ICIP.2010.5653278.

[8] Minarno, Agus & Munarko, Yuda & Kurniawardhani, Arrie & Bimantoro, Fitri & Suciati, Nanik. (2014). Texture feature extraction using co-occurrence matrices of sub-band image for batik image classification. 2014 2nd International Conference on Information and Communication Technology, ICoICT 2014. 249-254. 10.1109/ICoICT.2014.6914074.

[9] Azmi, Aini Najwa & Nasien, Dewi. (2014). Feature vector of binary image using Freeman Chain Code (FCC) representation based on structural classifier. International Journal of Advances in Soft Computing and its Applications. 6.

[10] ISánchez, Joan-Andreu & Romero, Veronica & Toselli, Alejandro & Villegas, Mauricio & Vidal, Enrique. (2017). ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset. 1383-1388. 10.1109/ICDAR.2017.226.

[11] *Merriam-Webster's Collegiate Dictionary*dictionary.cambridge.org/handwriting

[12] Allen Walker ReadEmeritus Professor of English, Columbia University. Author of "Approaches to Lexicography and Semantics" in Current Trends in Linguisticsbritannica.com/topic/dictionary/ handwriting

[13] Articles collected by British library , National British handwriting association nha-handwriting.org.uk

[14] Bensefia, Ameur & Paquet, Thierry. (2016). Writer verification based on a single handwriting word samples. EURASIP Journal on Image and Video Processing. 2016. 10.1186/s13640-016-0139-0.

[15] Gattal, Abdeljalil & Djeddi, Chawki & Siddiqi, Imran & Chibani, Youcef. (2018). Gender Classification from Offline Multi-Script Handwriting Images using Oriented Basic Image Features (oBIFs). Expert Systems with Applications. 99. 10.1016/j.eswa.2018.01.038.

[16] Abuzaraida, Mustafa & Zeki, Akram & Zeki, Ahmed. (2014). Online database of Quranic handwritten words. Journal of Theoretical and Applied Information Technology. 62. 485-492.

[17] Garris, M.D. & Blue, J.L. & Candela, G.T. & Dimmick, Darrin & Geist, Jon & Grother, P.J. & Janet, Stanley & Wilson, C.L.. (1995). Off-line handwriting recognition from forms. 2783 - 2788 vol.3. 10.1109/ICSMC.1995.538203.

[18] Djeddi, Chawki, Abdeljalil Gattal, Labiba Souici-Meslati, Imran Siddiqi, Youcef Chibani, and Haikal El Abed. "LAMIS-MSHD: A Multi-script Offline Handwriting Database", 2014 14th International Conference on Frontiers in Handwriting Recognition, 2014

[19] Wikipedia.com

[20] Feng, S.L.& Manmatha, R.. (2005). Classification Models for Historical Manuscript Recognition. Document Analysis and Recognition, International Conference on. 528-532. 10.1109/ICDAR.2005.73.

[21] americansecuritytoday.com

[22] Zineb Youbi, Larbi Boubchir, Abdelhani Boukrouche. "Human ear recognition based on local multi-scale LBP features with city-block distance", Multimedia Tools and Applications, 2018

[23] www.sims.berkeley.edu

[24] www.tdx.cat

[25] Djeddi, Chawki, Abdeljalil Gattal, Labiba Souici-Meslati, Imran Siddiqi, Youcef Chibani, and Haikal El Abed. "LAMIS-MSHD: A Multi-script Offline Handwriting Database", 2014 14th International Conference on Frontiers in Handwriting Recognition, 2014.

[26] Fischer, Andreas & Indermühle, Emanuel & Bunke, Horst & Viehhauser, Gabriel & Stolz, Michael. (2010). Ground truth creation for handwriting recognition in historical documents. ACM International Conference Proceeding Series. 3-10. 10.1145/1815330.1815331.

[27] C. Kauba, B. Prommegger and A. Uhl, "The Two Sides of the Finger - An Evaluation on the Recognition Performance of Dorsal vs. Palmar Finger-Veins," 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, 2018, pp. 1-5, doi: 10.23919/BIOSIG.2018.8553277.

[28] X. Tian, "Face Recognition System and It's Application," 2009 First International Conference on Information Science and Engineering, Nanjing, 2009, pp. 1244-1245, doi: 10.1109/ICISE.2009.583.

[29] A. Harb, M. Abbas, A. Cherry, H. Jaber and M. Ayache, "Palm print recognition," 2015 International Conference on Advances in Biomedical Engineering (ICABME), Beirut, 2015, pp. 13-16, doi: 10.1109/ICABME.2015.7323239.

[30] G. Aguilar, G. Sanchez, K. Toscano, M. Salinas, M. Nakano and H. Perez, "Fingerprint Recognition," Second International Conference on Internet

Monitoring and Protection (ICIMP 2007), San Jose, CA, 2007, pp. 32-32, doi: 10.1109/ICIMP.2007.18.

[31] S. P. Arokiaraj and L. Robert, "Pattern Recognition based DNA Sequence Compressor," 2012 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2012, pp. 1-5, doi: 10.1109/ICCIC.2012.6510211.

[32] J. Handa, S. Singh and S. Saraswat, "Approaches of Behavioural Biometric Traits," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 516-521, doi: 10.1109/CONFLUENCE.2019.8776905.

[33] M.s, Sreeraj & Idicula, Sumam. (2011). A Survey on Writer Identification Schemes. International Journal of Computer Applications. 26. 10.5120/3075-4205.

[34] Bulacu, Marius & Schomaker, Lambert. (2007). Automatic Handwriting Identification on Medieval Documents. Proceedings - 14th International conference on Image Analysis and Processing, ICIAP 2007. 279 - 284. 10.1109/ICIAP.2007.4362792.

[35] S. Fiel et al., "ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI)," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, 2017, pp. 1377-1382, doi: 10.1109/ICDAR.2017.225.

[36] Mourad Kedadouche, Zhaoheng Liu, Marc Thomas. "Chapter 32 Bearing Fault Feature Extraction Using Autoregressive Coefficients, Linear Discriminant Analysis and Support Vector Machine Under Variable Operating Conditions", Springer Science and Business Media LLC, 2018

[37] Maruf A. Dhali, Camilo Nathan Jansen, Jan Willem de Wit, Lambert Schomaker. "Feature-extraction methods for historical manuscript dating based on writing style development", Pattern Recognition Letters, 2020