l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

GuoXian TAN

Mémoire présenté en vue de l'obtention
du grade de Docteur de l'Université de Nantes
Sous le label de l'Université Nantes Angers Le Mans

Discipline : Informatique
Spécialité : Automatique et Informatique Appliquée
Laboratoire :IRCCyN (UMR CNRS 6597)

Soutenue le 6 juin 2013

École doctorale : 503 (STIM)

Thèse N° ED 503-194

Thèse en co-tutelle avec
Nanyang Technological University, Singapore

Writing Style Modelling Based on Grapheme Distributions Application to On-Line Writer Identification

Modélisation des styles d'écriture basée distributions de graphèmes Application à l'identification de scripteurs

JURY

Rapporteurs :	**M. CAMPISI Patrizio**, Professeur, Université de Rome III **M. PAQUET Thierry**, Professeur, Université de Rouen
Examinateur :	**M. SIYAL Mohammed Yakoob**, Professeur, Nanyang Technological University
Directeur de Thèse :	**M. VIARD-GAUDIN Christian**, Professeur, Université de Nantes
Co-directeur de Thèse :	**M. KOT Alex**, Professeur, Nanyang Technological University

Table of Contents

Preface **5**

Acknowledgements **7**

1. Introduction **15**

1.1 Motivation 16

1.2 Objectives 20

1.3 Major Contributions of the Thesis 20

1.4 Organization of the Thesis 24

2. Literature Review **27**

2.1 Introduction 28

2.2 Questioned document examination by human experts 28

2.3 Individuality of handwritten documents 29

2.4 Writer Identification in different languages 30

2.5 Writer identification and verification 38

2.6 Publicly available databases 47

2.7 Competitions 49

2.8 Summary 53

3. Writer Identification: Character level **55**

3.1 General Overview 56

3.2 Character Prototyping Model 59

3.2.1 Prototype training stage 60

3.2.2 Feature extraction 61

3.2.3 Document labeling stage 64

3.2.4 Fuzzy C-means Model 65
3.2.5 Classification stage 71
3.3 Character Prototype Model Evaluation 72
3.3.1 Data Acquisition 72
3.3.2 Database 74
3.3.3 Fuzzy C-means Model (FCM) Discussion 77
3.3.4 FCM Kernel Evaluation 78
3.3.5 FCM with Kullback-Leibler Divergence 84
3.3.6 Chi-square Distance 86
3.4 Sensitivity Analysis and Limitation of Proposed Models 87
3.4.1 Effect of Number of Character Prototypes on Accuracy 87
3.4.2 Effect of Length of Text 88
3.5 Summary 90
4. Writer Identification: Alphabet level 93
4.1 General Overview 94
4.2 Alphabet Knowledge Model 94
4.2.1 Discriminative Power of different letters of the alphabet 94
4.2.2 Proposed Methodology in using Alphabet Knowledge 100
4.2.3 Experiment 104
4.2.4 Alphabet Knowledge Model 108
4.3 Alphabet Information Coefficient 110
4.3.1 Effect of Alphabet Information Coefficient 110
4.3.2 Design of the Alphabet Information Coefficient 111
4.4 Discussion 115
4.4.1 Discriminative power of different languages in the Latin alphabet system 115
4.4.2 Discussion of Alphabet Information 119
4.5 Summary 120
5. Conclusions 123
5.1 Framework for writer identification on Alphabet-based writing systems 124
5.2 Establishment of a language-specific information criterion - Alphabet Information Coefficient 124
5.3 Future Directions for Writer Identification: Character level 126

5.4 Future Directions for Writer Identification: Alphabet level 128

Author's Publications 133

Bibliography 137

Preface

The increasingly pervasive spread of mobile digital devices such as mobile smartphones or digital tablets that use digital pens brought about the emergence of a new class of documents; online handwritten documents. The rapid increase in the number of online handwritten documents using such mobile devices leads to mounting pressure on finding innovative solutions towards faster processing, indexing and retrieval of these documents from databases. One such method to address this issue could be to extract writer information derived from the raw ink signal for indexing and retrieval of the documents. Hence, online writer identification is a topic of much renewed interest today because of its importance in applications such as writer adaptation, routing of documents and forensic document analysis.

This thesis proposes an automatic text-independent writer identification framework that integrates an industrial handwriting recognition system, which is used to perform an automatic segmentation of an online handwritten document at the character level. The proposed method is a text independent method that does not place any constraints on the content written or writing styles of the writers, to extract writer information at the character level from online handwritten documents and presents a novel approach to cluster and classify the resulting character prototypes for writer identification. This is a novel approach because prototypes are trained as characters using adapted Information Retrieval models, instead of the typical grapheme approach.

Subsequently, a fuzzy c-means approach is adopted to estimate statistical distributions of character prototypes on different letters of the alphabet. Character prototypes allow for a more intuitive prototype model compared to using grapheme prototypes which are often part

of a character and are not meaningful by themselves as prototypes. Furthermore, character prototypes allow for more robust and consistent prototypes to be built in the recognition process. These distributions model the unique handwriting styles of the writers. The proposed system attained an accuracy of 99.2% when retrieved from a database of 120 French writers.

In addition, the framework can be extended to any languages that use an alphabet writing system such as Latin, Greek or Cyrillic alphabet systems. In order to handle this, the framework is modified to examine the character prototypes at a deeper level. We hypothesize that the alphabet knowledge inherent in such character prototypes can provide additional writer information pertaining to their styles of writing and their identities. This thesis utilizes the character prototype approach previously mentioned to establish evidence that knowledge of the alphabet offer additional clues which help in the writer identification process. An Alphabet Information Coefficient is consequently introduced to better exploit such alphabet knowledge for writer identification. Our experiments showed an increase of writer identification accuracy from 66.0% to 87.0% on a database of 200 reference writers on a Reuters-21578 dataset of English writers when alphabet knowledge was used. Experiments related to the reduction in dimensionality of the writer identification system are also reported. Our results show that the discriminative power of different letters of the alphabet can be used to reduce the complexity while maintaining the same level of performance for the writer identification system.

Acknowledgements

"Give a man a fish and you'll feed him for a day

Teach a man to fish and he'll feed himself for a lifetime"

This is the philosophy of my supervisors, Professor Alex Kot Chichung and Professor Christian Viard-Gaudin who taught me the ways and means to acquire knowledge on my own instead of simply transferring them to me directly. I will like to take this opportunity to express my heartfelt gratitude to the both of them for their patience and understanding. I am truly grateful to them for their sharp perspectives and insightful criticisms. I also wish to take this chance to thank my previous final year project mentor, Dr. Kantisara Pita (Associate Professor) for his teachings that had equipped me to take on many of the challenges encountered over my PhD studies.

This self-discovery journey was also made possible by the following parties;
NTU and the French embassy for having this brilliant program and awarding me with the Merlion PhD scholarship and to Dr. Yang Huijuan, Dr. Cao Hong, for their patient guidance. I would also like to express my gratitude to Dr. Awal Montaser and his wife, Dr. Abdul Rahim Ahmad, Mr Steward Chu, Ms Evelyn Tee, Ms Pascaline Lebot, and many of the wonderful lab classmates and colleagues in Centre of Information Security, NTU, Image and Video Communication Lab, France and Vision Objects, France for their continual support in my daily course of work. I would also like to acknowledge the evolution of this thesis through previous works done by Ms Chan Siew Keng, Dr Tay Yong Haur and Dr Pierre-

Michel Lallican, without which, the works and contributions in this thesis would not have been possible.

Last but not least, I would simply like to take this opportunity to extend my heartfelt gratitude to all parties who has helped me in one way or another. They have contributed towards making this PhD project a rewarding and enriching research experience. Thank you, from the bottom of my heart.

List of Figures

Figure 2-1: A description of some writing systems commonly used [Namboodiri and Jain 2004; Omniglot 2012]. 32

Figure 2-2: Graph of percentage of users worldwide using a particular writing system. 33

Figure 2-3: A sample Greek handwritten text passage 34

Figure 2-4: Greek alphabet system, the first column shows the small Greek letters, and the second column shows the equivalent Greek letters in capital letters [Peter Allen Miller 2002]. 35

Figure 2-5: A sample Cyrillic handwritten text passage 36

Figure 2-6: Cyrillic alphabet system. Adapted from [Berlitz 1993]. 37

Figure 3-1: A sample of words from IRONOFF database 56

Figure 3-2: Simplified example of the proposed methodology 57

Figure 3-3: Block diagram for proposed methodology 60

Figure 3-4: (a) Morphological variation, (b) Directional variation, (c) Temporal variation 61

Figure 3-5: Resampling of the segmented characters in a fixed number of points 62

Figure 3-6: Local direction and curvature features 62

Figure 3-7: An allograph of ‘t’ with the latent stroke in dotted line 63

Figure 3-8: Examples of character prototypes of ‘f’ after clustering..... 63

Figure 3-9: Feature space with 4 prototypes p_1 (+1,-1), p_2(-1, +1), p_3 (-1,+1),p_4(+1,-1) and different positions of a test sample x following a spiral indexed by an angle [0°, 720°]..... 67

Figure 3-10: Partial membership inverse distance kernel functions for different positions with 4 prototypes, p_k 68

Figure 3-11: Nokia digital pen (Nokia SU-1B) and Esselte digital paper. 73

Figure 3-12: Features of a digital Pen, Nokia SU-1B..... 74

Figure 3-13: Example of a text passage from a reference document..... 75

Figure 3-14: Example of correct segmentation and labeling at the character level by the industrial text recognizer. 75

Figure 3-15: Example of the presence of segmentation errors 76

Figure 3-16: Partial membership Gaussian distance kernel functions for different positions with 4 prototypes, P_k..... 80

Figure 3-17: Partial membership exponential distance kernel functions for different positions with 4 prototypes, P_k..... 81

Figure 3-18: Partial membership exponential distance kernel functions for different 83

Figure 3-19: Graph of Identification Rate against Number of Clusters..... 88

Figure 3-20: Average number of misclassified writers against number of characters 90

Figure 4-1: A sample text passage from theReuters-21578 database (financial news category).104

Figure 4-2: A sample of the form used to collect the handwritten sample from the writers. 105

Figure 4-3: The word 'mln' that has been wrongly recognized 106

Figure 4-4: The 'e' in the wrongly recognized word that will be used for writer identification . 107

Figure 4-5: The sigmoid function in the $C_{\alpha_i} \times C_{\alpha_j}$ space 112

Figure 4-6: Performance of the writer recognition system on a database of 200 writers using different Alphabet Information Coefficient functions. 114

Figure 4-7: Identification rate as letters are removed based on their discriminative power 119

List of Tables

Table 1: Review of state-of-the-art in writer identification for the past decade. 40

Table 2: Review of publicly-available database for benchmarking writer identification algorithms 48

Table 3: Writer Identification Competitions 50

Table 4: Building of the term frequencies and inverse document frequencies. 70

Table 5: Performance of the Fuzzy c-means algorithm using different kernel functions 79

Table 6: Performance of writer identification using different distance metrics 86

Table 7: Discriminative power of different letters of the alphabet in writer identification on French handwritten documents 96

Table 8: Effect of alphabet knowledge on writer identification rate 108

Table 9: Discriminative power of different letters of the alphabet in writer identification on English handwritten documents (REUTERS-21578 database) 116

List of Symbols

α	letters of the alphabet, ‘a’ to ‘z’
x	instance of character x belonging to letter α
p_k	prototype k
$P_\alpha(p_k \mid x)$	conditional probability of a given character x assigned to prototype k
N	number of prototypes used
$dist\ (p_k, x)$	similarity distance between prototype k and character x
$tf_{\alpha,k}$	term frequency of prototype k in letter α
$idf_{\alpha,k}$	inverse document frequency of prototype k in letter α
λ_α	reliability factor
$n(x_i)$	number of characters of letter a in reference document i
$n(x_j)$	number of characters of letter a in test document j
$C_{\alpha i}$	total number of characters of all letters a in reference document i
$C_{\alpha j}$	total number of characters of all letters a in test document j
R	number of reference writers
T	number of test writers

1. Introduction

1.1 Motivation

Technology has become an integral part of modern lifestyles where our lives are becoming intertwined with technology itself. The adoption of mobile computing devices are seeing higher take-up rates as smart-phones or tablets such as ipads begin to integrate tightly with various facets of our everyday lives. Smart-phones are no longer confined to just being used as simple portable communication devices, but they are also being used to access, manipulate and handle a whole spectra of applications. For example, smart-phones and tablets are increasingly being used as direction-finders or Global Positioning Satellites (GPS) devices. Furthermore, with the penetration of 4G network as well as the assimilation of Augmented Reality (AR) technology into applications with localized services, one can expect more applications with seamless and natural interaction between the user and their mobile devices when being used to find directions. This dynamic shift in how such devices are being used demand a fundamental shift in mindset in how users interact with technology in a more natural and interactive manner, as well as how data in such devices are being designed, stored and accessed.

With such advancement of technology by leaps and bounds, the last decade has seen a major and dramatic shift in paradigms towards the usage of more natural and simplified ways of interacting with computers. Mobile tablets allow the tech-savvy businessman to optimize his time and plan for corporate meetings, make business presentations and sale pitches, communicate and send emails on his mobile devices while on the move, all using the same mobile devices. Businessmen are able to annotate, highlight and write sales figures while they are making sale reports on the move with such mobile devices. Even educators are doing away with thick and heavy textbooks and replacing teaching material with these mobile computing devices such as ipads or other Android-based devices. Such devices permit

handwritten equations to be written into the lessons while teaching mathematics; and even create annotations and comments to help in the explanation of certain concepts that are being taught. Even though the keyboard has been introduced for a long time, it is nonetheless not a natural or intuitive way of entering data into computers. It is virtually impossible to input a mathematical equation or sketch using the keyboard. Hence, numerous initiatives have been funded to research upon and develop more efficient algorithms, software and computing platforms to handle the surge in demand for seamless and natural interaction between human users and machines [Oviatt, Cohen et al. 2000], and to deliver interactive environments with a new level of intelligence.

All these led to the emergence and proliferation of a kind of document: handwritten online documents. Online handwritten digital documents are defined as those digital documents that not only provide information obtainable from offline digital documents, but also contain dynamic temporal information of the handwriting process [Jain and Namboodiri 2003]. This implies that additional features such as the velocity, pressure and even latent pen movements when the pen is up or when the pen is down can be extracted for indexing in online digital documents. This additional information that can be extracted from online handwritten documents is able to provide vital clues as to the identities of the writer. They are produced by state-of-the-art devices such as Tablet PC, mobile communication devices such as smart-phones with stylus input, mobile tablets such as ipads with natural user interfaces like stylus pens or even digital paper coupled with digital pens. The rapid increase in the number of such documents requires efficient management tools to properly index and retrieve them from databases.In this information age, we can expect an explosion of information and efficient means of storing and retrieving digital documents is no longer a luxury. Mounting pressure is therefore emphasized on finding innovative solutions towards faster processing, indexing and retrieval of digital documents. In order to alleviate the increasingly heavy

consumption of resources, indexing techniques can benefit storage of digital documents in a number of ways.

Firstly, even with the growing storage capacity and emergence of dual or even quad-processing cores in laptops, tablet PCs, smart-phones and other mobile communication devices that run at parallel processing speeds, these devices are still not optimized to handle large amounts of data, especially digital documents efficiently. In this regard, indexing techniques can help to further alleviate the resource constraints on such devices. There is an increasing trend of software consuming more systems resources, especially so for resource-hungry, memory intensive programs that run demanding computer graphics. Indexing the digital documents will not only ensure that the storage utilization is small, but also ensures sequential search and retrievals to be computationally less expensive and further optimized to be more efficient. This frees up system resources for other more critical multi-threaded processes. Therefore, battery life will be prolonged and power consumption will significantly decrease, which is currently the main gripe with most users using these mobile communication devices today.

Secondly, there is often a need to interface with external large data repositories to access and manipulate large amounts of information. For example, a digital database containing handwritten documents could easily number thousands of documents such as the one used by [Srihari, Cha et al. 2001]. Indexing will allow the search to be run in a more efficient manner and reduce the round-trip time it takes to send the queries to the large database and the time it successfully retrieves the digital documents. However, because of the nature of handwritten documents, traditional indexing techniques such as using hash tables or partitioning are not possible. Instead, subjective identifiers such as writer information can characterize and interpret the extracted content in a meaningful and useful manner. Indexing

techniques generally adopt some similarity measure to form clusters. Representative clusters are then used to store the indexes of the document, and using these representative clusters to access the indexes will be more efficient.

Thirdly, from information security's point of view, writer identification has ubiquitous applications in digital rights management and forensic analysis. The individuality of handwriting styles allows handwriting analysis to be considered as a behavioral biometric trait that can differentiate between different people [Morris 2000; Srihari, Cha et al. 2001; Srihari, Cha et al. 2002]. Handwriting analysis has traditionally been used in the field of forensic document analysis by forensic experts in the detection and to a large extent, the deterrence of fraud, identity theft and embezzlement cases. Furthermore, in environments where large amounts of documents, forms, notes and meeting minutes are constantly being processed and managed, knowing the identity of the writer would provide an additional value. The ability to retrieve all the documents of a specific user allows to browse Pencast websites where handwritten digital versions of personal notes are stored exactly as they were captured. One such application is to process and retrieve the identities of students for subsequent verification purposes. Typically, the handwritings of students can vary according to different cultures, occupations, physical attributes and even physiological factors as discussed by Huber, Headrick and Morris [Huber and Headrick 1999; Morris 2000]. This individuality of handwriting styles allows handwriting analysis of students taking tests and examinations to be considered as a behavioral biometric trait that can differentiate different people apart and ascertain their identities during tests and examinations, thus minimizing the occurrences of fraud or identity thefts.

Last but not least, we can also perform writer-adaptation to create, store or retrieve a profile of handwriting styles of the writers if we are able to automatically determine their

identities [Connell and Jain 2002; Chellapilla, Simard et al. 2006]. This way, the performance of the handwriting recognition system can be vastly improved since we are able to customize the recognition system to tailor to the writing profile and style of the writer. One can thus imagine a virtuous circle where knowing the writing identity from the bootstrap recognition system will help to improve the recognition results as already mentioned above.

1.2 Objectives

This research project is therefore highly motivated to build and improve upon a writer recognition system. A successful research on this thesis will breach the gap of the current findings and uncover new techniques for increasing the performance of online handwriting recognition systems. These considerations serve as motivating factors for this thesis, "Writing Style Modelling Based on Grapheme Distributions – Application to On-Line Writer Identification".

1.3 Major Contributions of the Thesis

1. **Character Prototyping Approach**

Our methodology proposes a prototyping approach at the character. This is the finer level which still holds a semantic meaning or at least a language based origin. Furthermore, prototypes that are built at the character level allows for an intuitive graphical inspection of the prototypes that can be invaluable for helping forensic experts to analyze handwriting. In addition, our character prototyping approach is not limited to just the Latin script, but the approach can be extended to other scripts that make use of a repeating set of characters for their writing systems such as the Greek and Cyrillic alphabet systems.

2. Interoperability Design

This thesis provides a design for a fully automatic framework that integrates an industrial engine into the writer identification system. The significant benefit from using an industrial recognition engine lies in being able to take full advantage of the level of maturity and accuracies of the current handwriting recognition technology. Going forward, advancements in handwriting segmentation, classification and recognition techniques will continue to advance and develop, leading to further benefits that can be derived from such off-the-shelf industrial recognition engines that can be integrated into the proposed framework.

By relying on such an approach, not only the writer identification task will be possible but additional functionalities may be permitted. Keyword extraction, document categorization, automatic summarization are among those tasks which take advantage of a textual transcription of a document originally available as a handwritten document.

3. Fuzzy C-means Model (FCM)

The traditional document retrieval methodology in using Information Retrieval models such as Vector Space Model is adapted with our proposed Fuzzy C-means approach. The Fuzzy C-means model is consistently used throughout the framework during clustering, building of prototypes, and assigning character prototypes to give superior performance compared to the existing Information Retrieval models. Various fuzzy kernels such as the exponential kernel function, Gaussian kernel function and the inverse kernel function, have been investigated to determine the effect of such an approach.

4. Alphabetic Information

One of the added advantages of working at the character prototype level is that we can make use of alphabetic information in various stages of clustering, classification or recognition. A proposed coefficient called the Alphabet Information Coefficient provides additional basis to improve the recognition rate during writer identification. In addition, using other alphabetic information such as discriminative letters of the alphabet is also explained to use as alphabet knowledge to improve the writer recognition rate.

In addition to having defined a general framework for proposing a solution to the writer identification problem as explained above, we have also more specifically addressed the following points, by asking the following questions:

a. Influence of the number of prototypes to define the different writing styles for a given letter of the alphabet

The underlying question is to define how many variants are useful to depict the way people produce the different letters of the alphabet. At school, we have been taught to draw the letters with a reference model, which was supposed to be unique. Actually, this is not the case, more than one pattern are used for writing a given letter. However, too few a representation of patterns can lead to inadequate depiction of the letter, yet over-representation of patterns poses the problem of over-fitting. Therefore, the real question is thus: how many prototypes are sufficiently enough, at the trade-off of performance versus computing time and complexity?

b. Influence of the number of letters (alphabet subset) to build the writing style distribution

Here, the question is to know if all the letters support the same amount of information regarding the writing style? Or, should we focus on some specific letters? For instance, does a 'o' letter which is very basic in its shape with a simple circle conveys as much information as the letter 'f' which is a much more complex pattern?

c. Influence of a hard or a soft prototype selection

Once several prototypes have been defined for representing a given letter, it is possible either to consider that a single one is selected has being the representative of a new instance of this letter or that all of them to a certain extent model this instance.

d. Influence of the metric used to match two distributions

Combined with the (term frequency) $\times$ (inverse document frequency) representation of a writing style, three different metrics have been investigated. They are the Euclidean distance, the Kullback-Leibler divergence and a Chi-square based distance.

e. Influence of the length of the texts

The longer a text is, the better it should be for extracting evidences of a specific writing style. We have conducted a sensitivity analysis of the writer identification system to illustrate this behavior and find the limits of the proposed system.

A concluding remark here is that many other interesting questions have not been investigated in this work. For instance, we have deliberately not considered the pen pressure as well as all other dynamic information, such as the writing speed or acceleration which are definitively important biometric features. The idea was to focus this study on the static representations of the letters and their identification capabilities. There is no doubt that if the goal would be to build a competitive system, a combination with such additional features would be beneficial.

1.4 Organization of the Thesis

This thesis is organized as follows:

In Chapter 2, a survey of the existing state-of-the-art methods for writer identification, with a focus on prototyping methods is presented. A comparison of their respective performance for some of the works over the past decade is also reviewed. In addition, a survey into the various publicly available databases, as well as recent trends in writer identification and their various competitions in the field of writer identification are also discussed.

Following this in Chapter 3, a study of how characters can be used to improve the recognition rate of writer identification is presented. This chapter provides the foundations of a framework for a writer identification system, and will be the basis of discussion for this thesis. This will then lead to the next chapter of an investigation of how alphabetic information can be used in writer identification systems.

Subsequently in Chapter 4, a proposed novel approach in employing alphabetic information to be used in writer identification systems is discussed. Various aspects of alphabet knowledge such as the discriminative power of different letters of the alphabet is explored in this chapter. Furthermore, a study of how different letters of the alphabet and its various combinations can be used to improve the recognition rate of writer identification is also proposed in this chapter. Finally in Chapter 5, conclusions are drawn and future research directions are proposed.

2. Literature Review

2.1 Introduction

With the advent of increasingly portable and mobile tablets, there is a growing trend and a shift in paradigm in using such devices for various reasons such as in education or during meetings in presentations. The underlying technology of handwriting recognition has however been around for decades. Nonetheless, the area of research into writer identification still remains very much an active region of interest in the document research community. This could be attributed to the ever-increasingly important applications writer identification plays. Such applications could be in the area of forensics, where an investigation into the identity of the author behind certain questioned documents can give rise to clues in solving crimes. This is of primarily interest in helping forensic scientists ascertain the authenticity of questioned documents such as suicide notes, inheritance wills or even historical documents.

2.2 Questioned document examination by human experts

The American Society of Questioned Document Examiners (ASQDE) regulates and outlines standards for questioned document examination by human forensic experts [Levinson 2001; Kelly and Lindblom 2006]. Some of the features that human forensic experts could make use in handwriting analysis could be prescribed at the character level. Character-level traits and characteristics such as entry and exit strokes, pen lifts, proportions, heights and loops are commonly used to determine the authenticity of questioned documents [Morris 2000]. The final conclusion drawn from a human forensic expert in questioned document examination must then be based on a sufficient number of common characteristics between the known and questioned writing samples, and requires a judgmental call by the human forensic expert in this process of comparison. It will thus be helpful to have a semi-automated recommender system to assist and help these human forensic experts to carefully examine

and further scrutinize the various characteristics. This would serve as a strong motivation to explore character-based writer identification approaches. However, the individuality of handwritten documents must first be established in order for such recommender systems to assist human forensic experts.

2.3 Individuality of handwritten documents

Qualitative studies of handwritings is said to date as far back to ancient philosophers such as Aristotle and Confucius. Such graphological analysis are more often than not very subjective and lack the scientific rigor, and they are highly biased and heavily influenced by the subjectivities of the person doing the analysis. It was not until recently that a more scientifically rigorous treatment to quantify the individuality of handwritten documents was made. One of the pioneering works in establishing a formal quantitative analysis of the individuality of handwritten documents can be credited to studies by Srihari et al. [Srihari, Cha et al. 2001; Srihari, Cha et al. 2002; Srihari and Leedham 2003; Srihari, Beal et al. 2005]. Pattern recognition techniques were used to provide objective assessments and scientific validations of the individuality in handwriting. In their work, both local and global features were extracted on a database of 1500 individuals stratified across different genders, age groups and ethnicities. A nearest neighbour algorithm and a 3-layered neural network are then used to train and classify the results for writer identification and writer verification respectively. Their results verified the hypothesis that individuality exists in handwriting with a 95% confidence from a statistical inference of the entire US population [Srihari, Cha et al. 2002].

Such individuality studies allowed expert testimonies from skilled document forensic specialists to be treated with scientific rigor and become admissible as court evidences.

Furthermore, these studies also paved the way for the development of numerous automatic writer identification and verification systems over this past decade. These studies provided scientific evidence to prove that the individuality of handwriting styles allows handwriting analysis to be considered as a behavioural biometric trait that can differentiate between different people [Morris 2000; Srihari, Cha et al. 2001]. Since then, the fundamental assumption underlying most state-of-the-art writer identification systems is that writers can be differentiated based on the uniqueness of their handwriting styles. The various writer identification systems then attempt to model the different styles of writings of writers by using a multitude of approaches. The existing literature of writer identification makes use of a vast array of methods such as the morphological approach [Zois and Anastassopoulos 2000], HMM-based approach [Schlapbach and Bunke 2004; He, You et al. 2008], transformation-based approaches [Pitak and Matsuura 2004; Fiel and Sablatnig 2012] and the allograph prototype approaches [Bensefia, Paquet et al. 2005; Sesa-Nogueras and Faundez-Zanuy 2012]. This will be further discussed in section 2.5.

2.4 Writer Identification in different languages

In the past decade, Srihari et al. [Srihari, Cha et al. 2001] established the individuality of handwriting through the design of a scientific and quantitative methodology using pattern recognition techniques. This laid the foundation for much of the current forensic principles when using pattern recognition techniques to identify handwriting. Research interest in writer identification techniques since flourished [Sungsoo, Seungseok et al. 2005; Namboodiri and Gupta 2006; Niels, Vuurpijl et al. 2007; Schlapbach, Liwicki et al. 2008] and many writer identification techniques were being developed. These developments were however, primarily in documents that contained English handwriting. Interestingly, writer identification in other languages started to advance, perhaps stemming from the need to

develop judicial processes of other languages into a more rigorous and scientific process in handwriting analysis as well. Writer identification techniques began to gain traction in languages such as French, Greek, Cyrillic, Portuguese, Chinese, Arabic, Farsi, Persian and even Brazilian [Chan, Yong et al. 2007; He, You et al. 2008; Helli and Moghaddam 2010; Chaabouni, Boubaker et al. 2011; Hanusiak, Oliveira et al. 2012]. This propelled strong legal motivation for forensic authorities to begin commissioning feasibility studies into using pattern recognition techniques for handwriting analysis and writer identification [Freitas 2008]. The Brazilian Forensic Letter Database contained Portuguese handwritten documents written by 315 writers, and contained numerous writing style information such as relative slant, relationship between letters and baseline, sentence indentation, and the use of whitespaces. The database is very peculiar and characteristic to Brazilian writers about their various writing habits such as the innate absence of certain letters ‘k’, ‘w’ and ‘y’ in everyday nouns because these special foreign letters are only used for personal names in Brazil.

With an increasing trend in writer identification techniques targeting different specific languages gaining traction, it then becomes worthwhile to take a step back to understand and categorize the different writing systems worldwide. The roots of major languages can be traced back and grouped into the following main writing systems: Alphabet-based, Syllabary, Abjad, Syllabic-Alphabet, Semanto-Phonetic. Figure 2-1 illustrates the various languages that belong to each of the above-mentioned writing system categories. It can thus be seen that a large portion of Latin-based languages such as English, French, German, Italian, Portuguese have its roots and writing system based on characters that define the alphabet set of that respective language.

As illustrated in Figure 2-1, scripts are defined differently from languages in that scripts form a larger subset of writing systems, whereas one script system can comprise of

various languages. Take for example in Figure 2-1, languages such as Portuguese, English, French and German share commonalities of similar letters, albeit with different grammer and diacritics, that belong to the Latin script. Likewise, Russian, Slavonic and Serbian languages share commonalities that belong to the Cyrillic script. Both the Latin and Cyrillic scripts, along with the Greek script, make use of the alphabet writing system which can always be represented by certain subset known as characters.

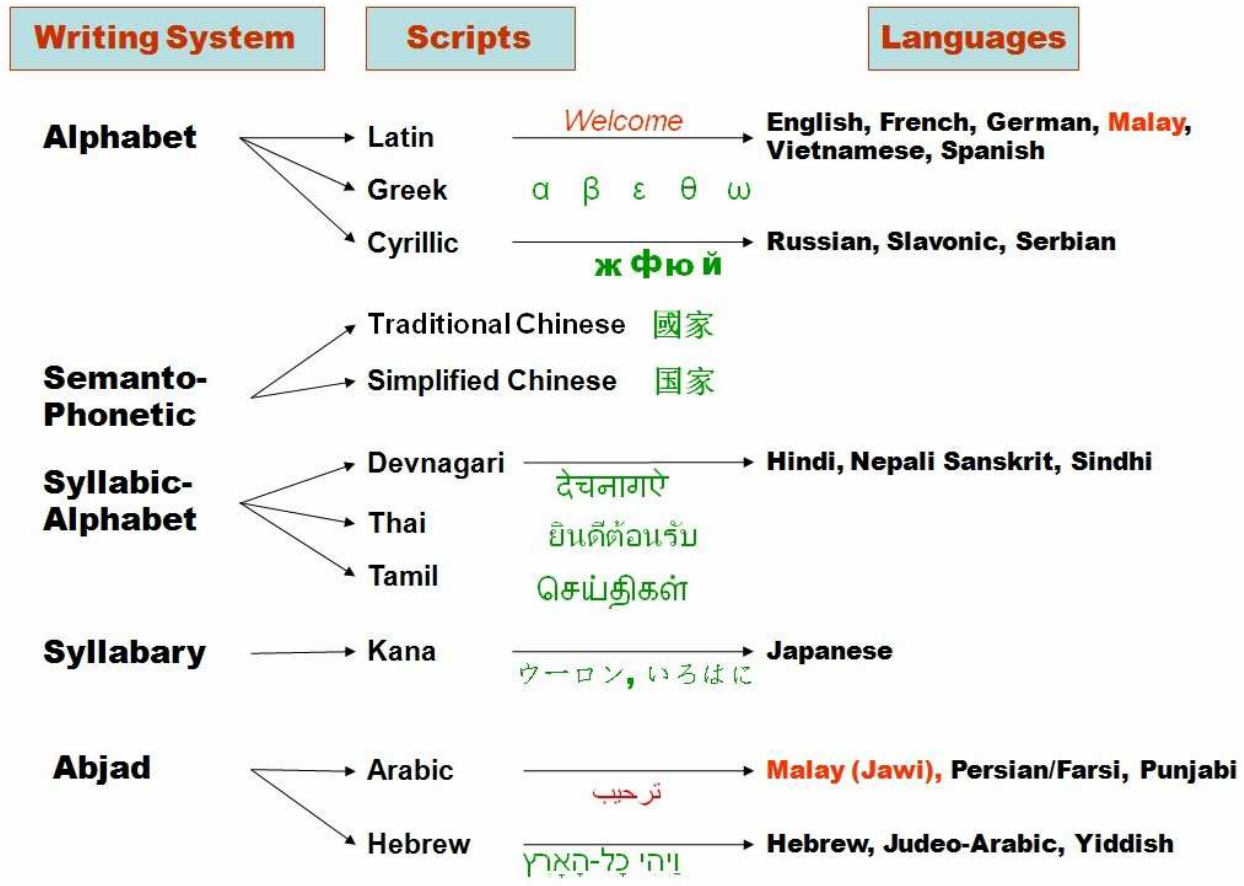


Figure 2-1: A description of some writing systems commonly used [Namboodiri and Jain 2004; Omniglot 2012].

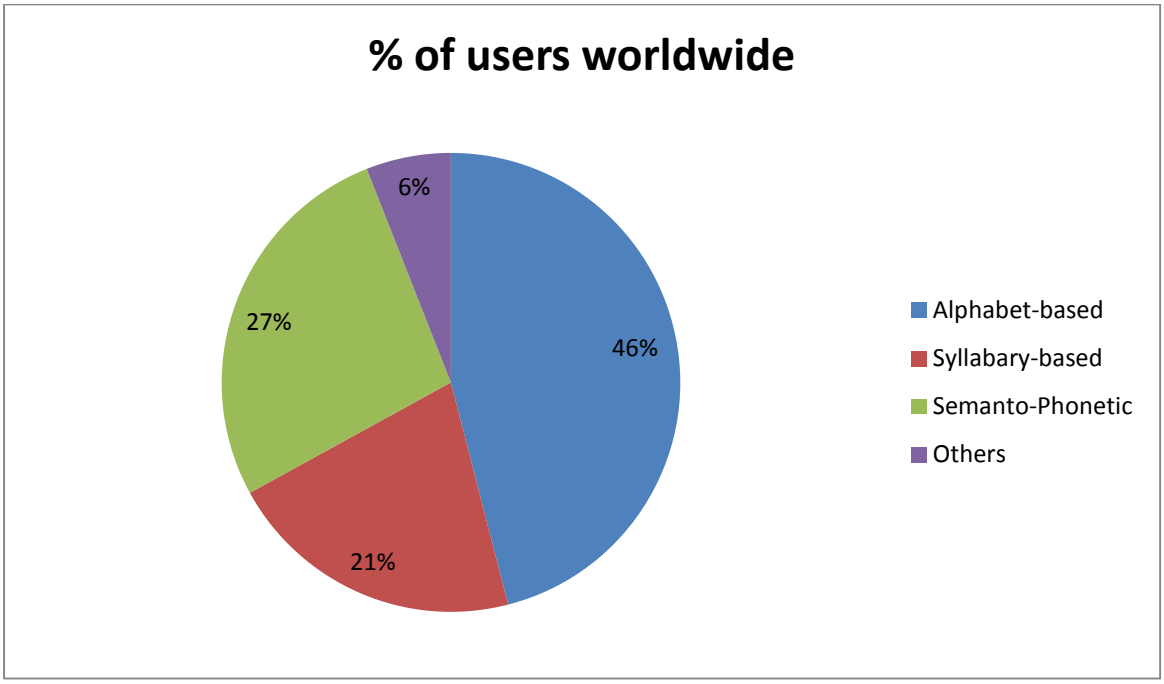


Figure 2-2: Graph of percentage of users worldwide using a particular writing system.

It is interesting to note the proportion of users worldwide that uses a particular writing system previously discussed. Figure 2-2 [Ethnologue 2012] shows that the highest percentage of worldwide users, standing at 46%, are making use of alphabet-based writing systems such as English, French, German, Greek and Cyrillic. The next largest group of users make use of the Semanto-Phonetic writing systems such as Chinese, Vietnamese, Japanese (Kanji), Korean (Hangul), standing at 27%. This is followed closely by Syllabary-based writing systems such as Tamil, Devanagari, Thai and Japanese Kana. This proportion of writing systems shown in Figure 2-2 gives an indication of the relative importance of alphabet-based writing systems compared to other writing systems, and that is also why there are currently more studies in the research area of handwriting analysis being conducted on alphabet-based writing systems. Furthermore, we are also seeing an increasing trend of research in the domain of document analysis and handwriting recognition being conducted in Semanto-Phonetic writing systems, especially in the Chinese language.

Now that we have appreciated the difference between languages, scripts and writings systems, let us further discuss on the intricacies of some of the alphabet writing system, namely Greek and Cyrillic. Figure 2-3 shows a sample handwritten text passage written in Greek and we can observe that the cursive handwritten text passage is somewhat similar to that of English text, and can be segmented into the character level. Figure 2-4 depicts a closer examination at the Greek alphabet set, comprising of 24 characters that make up this alphabet set. This is very similar to the Latin-based writing styles such as English, with one main difference in that Greek writing style does not make use of diacritics.

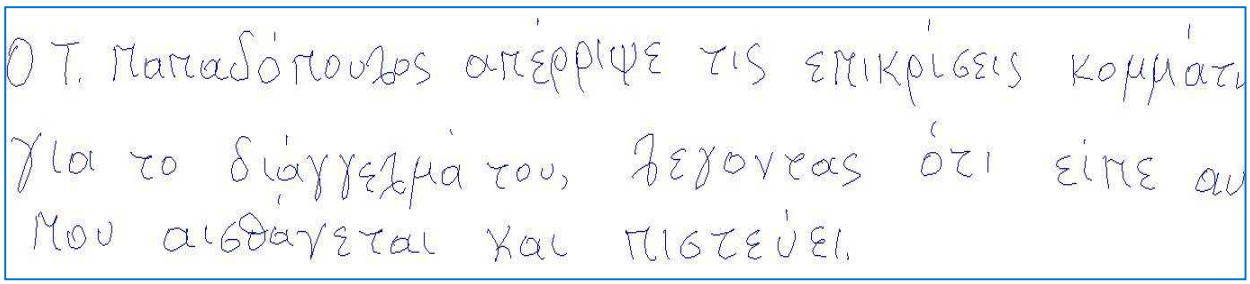


Figure 2-3: A sample Greek handwritten text passage

small letters	CAPITAL LETTERS	Names of letters	English pronunciation (phonetic equivalent)
α	A	alpha	a as in far
β	B	beta	b as in boy
γ[1]	Γ	gamma	g (hard) as in get
δ	Δ	delta	d as in dog
ε	E	epsilon	e as in get
ζ	Z	zeta	z (initial) as in zoo dz (medial) as in adz
η	H	eta	e as in obey
θ	Θ	theta	th as in think
ι	I	iota	i (short) as in pit i (long) as in machine
κ	K	kappa	k as in kite
λ	Λ	lambda	l as in lip
μ	M	mu	m as in man
ν	N	nu	n as in net
ξ	Ξ	xi	x as in box
ο	O	omicron	o as in not
π	Π	pi	p as in pit
ρ	P	rho	r as in rot
σ, ς[2]	Σ	sigma	s as in sell
τ	T	tau	t as in top
υ	Y	upsilon	u as in put
φ	Φ	phi	ph as in phone
χ	X	chi	ch as in German "ach"
ψ	Ψ	psi	ps (initial) as in psalms ps (medial or final) as in lips
ω	Ω	omega	o as in ode

Figure 2-4: Greek alphabet system, the first column shows the small Greek letters, and the second column shows the equivalent Greek letters in capital letters [Peter Allen Miller 2002].

A cursive sample of the Cyrillic writing style is shown in Figure 2-5. A common feature among these two text passages in Figure 2-3 and Figure 2-5 is that they make use of the alphabet writing system, and such writing styles can be decomposed into a recurring set of basic characters in such alphabet writing systems. Such basic characters then build up their vocabularies and lexicons by combing these set of basic characters in various combinations to form words. Likewise, Figure 2-6 shows a representation of the Cyrillic alphabet set of characters, along with their diacritics. Certain languages such as French or Portuguese may

also contain diacritics to express more meanings to their words formed, even though they are all derived from Latin roots.

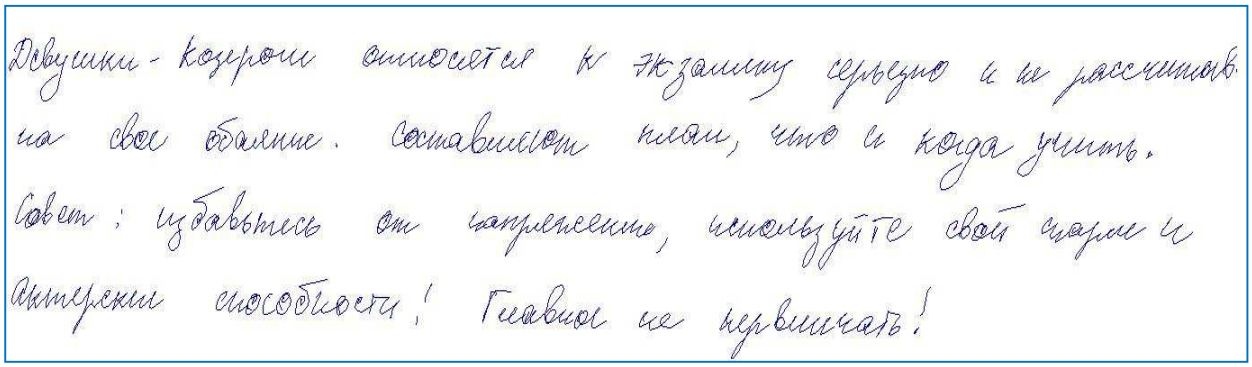


Figure 2-5: A sample Cyrillic handwritten text passage

The Russian Alphabet			
Character	Sound	Vowel	Pronunciation
А а	ah	*	a in car
Б б	b		b in bit
В в	v		v in vine
Г г	g		g in go
Д д	d		d in do
Е е	yeh	*	ye in yet
Ё ё	yo	*	yo in yonder
Ж ж	zh		zh in pleasure
З з	z		z in zoo
И и	ee	*	ee in see
Й й	y	*	y in boy
К к	k		k in kitten
Л л	l		l in lamp
М м	m		m in my
Н н	n		n in not
О о	o	*	o in hot
П п	p		p in pot
Р р	r		trilled
С с	s		s in see
Т т	t		t in tip
У у	oo	*	oo in boot
Ф ф	f		f in face
Х х	kh		ch in loch
Ц ц	ts		ts in sits
Ч ч	ch		ch in chip
Ш ш	sh		sh in shut
Щ щ	shch		sh followed by ch
Ъ ъ	(hard sign)		
Ы ы	ih	*	i in ill
Ь ь	(soft sign)		
Э э	eh	*	e in met
Ю ю	yoo	*	u in duke
Я я	yah	*	ya in yard
7.62x54r.net			

Figure 2-6: Cyrillic alphabet system. Adapted from [Berlitz 1993].

Now that we have established the different writing systems noted in current literacy studies, the sections that pursue go on to describe past and current works being done in the area of online writer identification and verification systems for languages that make use of the alphabet writing system (predominantly English and French). Their achievements, limitations and outstanding issues are given a thorough examination.

2.5 Writer identification and verification

Much progress has been made in both the fields of writer identification and writer verification in the last decade. In writer recognition systems, care must be taken to clearly distinguish writer identification systems from writer verification systems. Writer verification performs a one-to-one matching between a test writer and an already enrolled writer, and attempts to ascertain the authenticity of the test writer. On the other hand, writer identification involves executing a one-to-many match and returns a ranked list of results for the search. The difference, though subtle, lies in the applications in which they can be utilized in.

Writer recognition systems can typically make use of global features such as texture, curvature and slant features [Hochberg, Kelly et al. 1997; Busch, Boles et al. 2005; Bulacu and Schomaker 2007; Hanusiak, Oliveira et al. 2012] as well as a combination of local features such as graphemes, allographs and connected components [Schomaker and Bulacu 2004; Pervouchine and Leedham 2007; Srinivasan, Kabra et al. 2007] to identify the writers. They can be generally classified into approaches that utilize text-dependent or text-independent techniques. Signatures are examples of text-dependent systems since the writers have to write the exact same text as what they have written previously for the system during the enrolment process. Srihari et al.'s [Srihari, Cha et al. 2001; Srihari, Cha et al. 2002] works falls into this category of text-dependent approaches. They proposed the use of two levels of features; one at the macro level, making use of features such as the average slant, aspect ratios and entropies at the paragraph or document level. The other level functions at the micro level and makes use of features such as gradient, structural and concavity at the word or character level. They then used a multi-layer perceptron for writer verification and obtained an accuracy of 98% with this text-dependent approach that only required limited amount of text to be present.

Our work falls into the latter category of text-independent techniques where the writers are not bounded by any specific lines of text in order for the system to recognize them. Instead, the system analyzes their handwriting styles through a series of automated processes, regardless of what they have written. This kind of writer recognition systems included previous works such as the method proposed by [Pitak and Matsuura 2004] which adopted a Fourier transformation approach. The extracted features are the velocities of the barycenter of the pen movements and they are transformed into the frequency domain using Fourier transform. The advantage in adopting such a model is that it is text-independent, but at the expense of a lower noise tolerance. The noise must be filtered out as much as possible in the pre-processing stage, otherwise the noise might be mistaken for high velocity components once the features are transformed into the frequency domain. Text-independent writer recognition systems can also make use of stochastic approaches like the Hidden Markov Models (HMM) technique presented in the works of Schlapbach et al. [Schlapbach and Bunke 2004]. They built one HMM model for each writer and extracted nine features at the line level from a database of 100 writers. An identification rate of 96% was attained based on 8600 text lines from the 100 writers.

Table 1: Review of state-of-the-art in writer identification for the past decade.

Author	Year	Approach	Accuracy	Language	Domain
[Zois and Anastassopoulos 2000]	2000	Morphological approach	96.5% on 50 writers	English	Offline
			97% on 50 writers	Greek	
[Said, Tan et al. 2000]	2000	Gabor filters & grayscale co-occurrence approach	95% on 20 writers	English	Offline
[Srihari, Cha et al. 2001]	2002	Text-dependent multi-layer perceptron approach	98% on 1000 writers	English	Offline
[Pitak and Matsuura 2004]	2004	Fourier transformation approach	98.5% on 81 writers	Thai	Online
[Schlapbach and Bunke	2004	Hidden markov models approach	96% on 100 writers	English	Offline
[Bensefia, Paquet et al. 2005]	2005	Grapheme-based clustering approach	95% on 88 writers	French	Offline
			86% on 150 writers	English	
[Bulacu and Schomaker 2007]	2007	Textural and allograph prototype approach	92% on 650 writers	English	Offline
[Niels, Gootjen et al. 2008]	2008	Allograph prototype matching approach	100% on 43 writers	English	Online

[Chan, Viard-Gaudin et al. 2008]	2008	Discrete Character prototype distribution approach	95% on 82 writers	French	Online
[Garain and Paquet 2009]	2009	2D Auto-Regression techniques for writer identification	62.1% on 382 writers	French Arabic	Offline
[Siddiqi and Vincent 2010]	2010	Using redundant patterns in writing with orientation and curvature information	92% on 100 writers	Arabic	Offline
[Quang Anh, Visani et al. 2011]	2011	Tf-idf approach in comparing grapheme-based writer identification to character-based writer identification.	85% on 300 writers	French	Online
[Fiel and Sablatnig 2012]	2012	Codebook approach using SIFT features	90.8% on 650 writers	English	Offline
[Hanusiak, Oliveira et al. 2012]	2012	Texture-based global approach using SVM	96% on 115 writers	Portuguese	Offline
[Shivram, Ramaiah et al. 2012]	2012	Modeling of shared writing styles using a three-level hierarchical Bayesian model	89.33% on 43 writers	English	Online
[Marcelli, Parziale et al. 2012]	2012	Modeling of latent processes behind offline handwritten traces	77.27% on 18 writers	Italian	Offline
Our approach	2010	Continuous Character prototype distribution approach	99% on 120 writers	French	Online

Of paramount importance to this thesis are the approaches that make use of graphemes or allographs, which has been gaining popularity in writer identification. Such state-of-the-art algorithms and techniques make use of a template matching approach that assigns handwriting styles to prototype templates which are representative of the handwriting styles [Bensefia, Paquet et al. 2005; Chan, Yong et al. 2007; Niels, Vuurpijl et al. 2007; Niels, Gootjen et al. 2008; Helli and Moghaddam 2010]. The prototypes attempt to model the writing styles of the writers as close as possible, based on features extracted from the online handwritten documents. Identification of the writer is then achieved based on the comparison of some similarity measures between the extracted features from the reference writers and the test writer in question.

Bensefia et al. [Bensefia, Paquet et al. 2003; Bensefia, Paquet et al. 2005] proposed using a sequential clustering approach at the grapheme level for offline texts to categorize different writers for their writer identification system. This approach attained an identification rate of 86% on an English database of 150 writers and 95% on a French database of 88 writers. The advantage of this method is that it does not depend on any lexicon and is therefore language independent. However, working at the grapheme level requires the definition of the granularity of the considered graphemes. In their work, Bensefia et al. have selected as the finer level a sub-character level. In contrast, as will be presented later, our proposal will be to work systematically at the character level. Another work that involves graphemes is by Bulacu et al. [Bulacu and Schomaker 2007]. They have generated a codebook of graphemes and combined them with features at the texture level to attain an identification rate of 92% on a data set of 650 writers. Both of these works make use of information retrieval techniques in their systems.

More recently, Hanusiak et al. [Hanusiak, Oliveira et al. 2012] reported a promising result of 96% overall accuracy by taking into account texture-based features, as well as dissimilarity representation on a dataset of 315 writers from the Brazilian Forensic Letter Database [Freitas 2008]. The work was based on five feature descriptors such as entropy, homogeneity, dissimilarity, inverse variance and energy, and Hanusiak et al. went on to further discuss that the feature classifiers they used have no conditional dependency on one another, based on combining their feature classifiers in the ROC space. However, the main drawback of this technique is that a full page of handwritten text is required to be extracted into global features, and this amount of information might not always be readily available in real-life scenarios.

Interestingly, another recent trend in the research area of writer identification seeks not to just simply look at different sets of features or algorithms in identifying writers, but attempts to provide an alternative perspective on how writer identification processes should be modeled. Shivram et al. [Shivram, Ramaiah et al. 2012] approach of modeling commonalities of writing styles across different writers to characterize the different writer profile departs from the traditional works in that they believe the handwritings of different writers are unique, but their writing styles can share common traits. Hence, they attempt to form distributions that describe different writers based on the degree of writer-style commonalities such as degree of slant and the amount of loop. Shivram et al.'s work [Shivram, Ramaiah et al. 2012] reported an accuracy of up to 89.33% when 30 of such writer–styles were used in their three-level hierarchical Bayesian structure. Another recent work that deviates from the norm by Marcelli et al. [Marcelli, Parziale et al. 2012] attempts to look at modeling offline handwriting styles through the latent processes behind how the various handwriting traces are formed, instead of looking at a set of feature sets that describe writer profiles based on their handwriting features. The contribution of their paper lies in

having a simple model that provides a writer profile by focusing only on the latent processes that creates the various offline handwritten traces, instead of modeling the handwritten traces themselves. Marcelli et al. [Marcelli, Parziale et al. 2012] reported writer verification accuracies of up to 77.27%. Interestingly, they even validated their results through the participation of experienced human forensic document examiners where only one out of 25 of these forensic document examiners were able to correctly perform a writer verification. However, it is to be noted that their database is very limited, and consists of only a total of 32 documents written by 18 different writers with no results to report.

Information retrieval (IR) techniques are gaining popularity in writer identification. Among many popular types of IR models such as the fuzzy model, Boolean model or the probabilistic model [Radecki 1983; Chen and Horng 1999], the vector space model approach that was first proposed by Salton et al. [Salton, Wong et al. 1975] remains to this day one of the most dominant approaches in IR due to its relatively simple, yet effective design. This vector space model involves two stages; an indexing phase and retrieval phase in a high dimensionality feature space. The indexing phase involves representing the set of documents with a set of occurrence vectors, the term frequency (tf) and inverse document frequency (idf). The underlying principle of the tf-idf combination relies on how frequent (or infrequent for the case of idf) a feature occurs in the document to represent the relevance of that feature towards the retrieval of the document.

Consequently, the retrieval phase then compares the tf-idf vector of the query document with that of the indexed document for retrieval of the document. This IR approach was later adapted for use by Bensefia et al. [Bensefia, Paquet et al. 2003] in the context of writer identification. Their contribution was to apply the concepts of tf and idf using graphemes as their feature set for the problem of writer identification. They proposed that the

invariant handwriting styles of writers can be viewed as features that are defined within the writer's set of allographs. A clustering algorithm can then be used to define the groups of patterns that model the invariant handwriting styles of writers. Subsequently, a tf-idf score will be calculated based on these prototypic groups of patterns for the indexing and retrieval phase. The works by Bensefia et al. laid down the foundations of numerous current writer identification techniques that adopt the IR approach. Among recent research in writer identification using IR techniques lies two noteworthy works by Niels et al. [Niels, Vuurpijl et al. 2007; Niels, Gootjen et al. 2008] and Chan et al. [Chan, Viard-Gaudin et al. 2008] that yielded promising results.

Niels et al. [Niels, Vuurpijl et al. 2007] used dynamic time warping to hierarchically cluster allographs and build a set of membership vectors, which contains the frequency of occurrence of each allograph for each character. This prototypic template of membership vectors then represented the handwriting styles of the different writers. However, dynamic time warping approaches are computational expensive. Furthermore, it is difficult to cover all variations in handwriting during the training of the prototypes and dynamic time warping is highly sensitive to the absence of prototypes. Therefore, dynamic time warping will not be able to give the expected results if a new variation in handwriting that was not previously covered during the training of the prototype were to appear [Niels, Vuurpijl et al. 2007].

Another distinctive difference is that the framework proposed in this thesis is fully automatic whereas the work done by Niels et al. relies on a manual segmentation process. The work by Chan et al. [Chan, Viard-Gaudin et al. 2008] handled this issue of previously missing handwriting variations during prototype building by adopting a statistical approach. They made use of a character prototype distribution to model the specific allographs used by a given writer and created statistical distributions to model the handwriting styles of writers.

A top-1 accuracy of 95% was achieved based on this text-independent approach which considered 82 reference writers. Even though working at the character level as opposed to using the grapheme (sub-character) or word level appears to be quite challenging, character based approaches are able to produce a more consistent set of templates for writer identification provided that recognition and segmentation are performed accurately.

From the review of the current state-of-the-art described in this section, the following conclusions can be derived. Firstly, the text-dependent approaches allow high accuracies to be achieved even from a limited sample of text. However, one serious drawback of this is the issue of feasibility in implementing this kind of systems in reality. Writers will have to know the exact text to write, thus restricting its applicability to limited real-life situations. Text-independent approaches, on the other hand, circumvent this issue by using statistical methods that extract writer-specific features that are insensitive to the textual content of the documents. The drawback is that a minimum amount of text needs to be present for such methods to be statistically sufficient. Since our proposed methodology falls into the latter category, we have conducted a study to determine the minimum amount of text required, which is presented in section 4.5. Secondly, the literature review allows us to conclude that prototype-matching based approaches are gaining popularity in recent years, as they are able to provide high levels of accuracies. Table 2 provides a brief summary of some of the recent works in writer identification for the past decade.

Interestingly, we observe a trend of more and more IR-based approaches [Bensefia, Paquet et al. 2005; Bulacu and Schomaker 2007; Niels, Vuurpijl et al. 2007; Chan, Viard-Gaudin et al. 2008] being proposed for writer identification in recent years. This gain in popularity is proof of the potential that such IR models can achieve for writer identification, in spite of the simplicity in its design. This upsurge in popularity can first be credited to early

Bensefia et al.'s works [Bensefia, Paquet et al. 2003] for adapting IR models into writer identification.

Yasuda et al. [Yasuda, Takahashi et al. 2000] proposed a HMM-based approach that is invariant to noise and small shape change. However, this model is text-dependent. Jain et al.[Jain and Namboodiri 2003] proposed using Dynamic Time Warping (DTW) on the stroke direction, curvature and the height as features to do word matching. They were able to achieve high precision of 92.3% at a recall rate of 90%. The advantage of this method is that it is language independent, but at the expense of a slower processing time. Bulacu et al. [Bulacu and Schomaker 2007] used a two-pronged approach for automatic writer identification and verification. Textual information such as the slant and curvature were transformed into probability density functions and allographic information such as the graphemes were clustered into codebooks. This combinatorial approach yielded a top-1 accuracy of 87% and a top-10 accuracy of 96%. Chan et al. managed to achieve higher top-1 accuracy of 95% based on a character level prototype distribution methodology that is text-independent. An unresolved issue mentioned in [Chan, Viard-Gaudin et al. 2008] mentioned the applicability of using genetic algorithms to search for discriminative alphabets most descriptive of the writer in question.

2.6 Publicly available databases

The current state-of-the-art techniques and systems use differing standards of test data. While there are standardized databases available on the public domain for the research community to use such as IRONOFF, IAM, PSI or TRECVID, many past works tend to collect and collate and use their own set of research data. Caution must therefore be exercised when comparing among works with differing test conditions. For example, the work

presented by Said et al.[Said, Tan et al. 2000] resulted in a correct writer identification rate of 96% based on their data of 40 writers and some text lines per writer. Zois et al. [Zois and Anastassopoulos 2000] however used a database of 50 writers and 45 samples of the same word per writer to obtain a correct identification rate of 92.48%. The largest database is used by Srihari et al. [Srihari, Cha et al. 2001] where 1000 writers were asked to copy the same text 3 times. The accuracy of the writer identification system is influenced by the length of text in the document. A longer text allows for better recognition capabilities, thereby increasing the accuracy of the identification system. Currently, there are few attempts to correlate this relationship between the accuracy of the system using character level prototyping and the minimum length of text used.

Table 2: Review of publicly-available database for benchmarking writer identification algorithms

Name of database	Year Created	Reference	Language	Size of writers
IRONOFF	1999	[Viard-Gaudin, Lallican et al. 1999]	French	373 writers
CEDAR	2000	[Srihari, Cha et al. 2001]	English	1000 writers
IAM	1999	[Marti and Bunke 1999]	English	657 writers
PSI	2001	[Bensefia, Nosary et al. 2001]	French	88 writers

PUCPR	2008	[Freitas 2008]	Portuguese	315 writers
SCUT-COUCH-2009	2009	[Jin, Gao et al. 2011]	Simplified Chinese	190 writers
IfN/Farsi Database	2008	[S. Mozaffari and Amirshahi 2008]	Farsi	600 writers

Table 2 summarizes the various size of the database and some of the properties of these publicly available databases. It is interesting to point out that in the earlier years; most of the publicly available databases are based in either the English or French language. It is not until in recent years, that we observe a trend of more publicly available database being created in other languages such as Arabic, Farsi, Chinese and even Portuguese. There seems to be a growing trend of the handwriting recognition community looking towards establishing a common database for benchmarking purposes in other languages.

2.7 Competitions

Since 2011, in order to promote research on writer identification, a series of competitions on writer identification have been organized by the handwriting analysis community in conjunction with the various conferences such as ICDAR or ICFHR. It is heartening to know such competitions on writer identification are being actively organized by the document analysis community. The purpose of such competitions is to document the state-of-the-art advances in the field of writer identification, and test and benchmark the performances of various writer identification systems on a common set of challenging

databases. However, up to know to the best of our knowledge all these competitions were focused on off-line handwriting. None of them have been proposed for on-line documents.

What is remarkable about such competitions and the various proposed algorithms is that it provides us with an insight into the current state-of-the-art in writer identification algorithms. For example, the ICDAR 2011 Arabic writer identification contest that runs on the Kaggle platform, a commonly used data prediction contest platform, has attracted over thirty teams working in the field of Arabic writer identification and a total submission of 139 entries [Hassaïne, Al-Maadeed et al. 2011].

Table 3: Writer Identification Competitions

Name of competition	Year	Winning Results	Remarks
ICDAR 2011 Arabic Writer Identification Contest [Hassaïne, Al-Maadeed et al. 2011]	2011	UCL team achieved 100% accuracy using orientated Basic Image Feature columns (oBIF) approach	Database contained 50 writers
ICDAR 2011 Writer Identification Contest [Louloudis, Stamatopoulos et al. 2011]	2011	Tsinghua team achieved an overall rank of 1 using a grid microstructure feature (GMSF) approach.	Database contained 26 writers.

ICFHR 2012 Arabic Writer Identification Contest [Hassaïne and Al-Maadeed 2012]	2012	UCL team achieved 95.3% accuracy using oBIF approach. Zhang et al. also obtained 95.3% using kernel PCA and SVM to classify.	Database contained 200 writers. Joint 1st place with Zhang et al.
ICFHR 2012 Writer Identification Contest [Louloudis, Gatos et al. 2012]	2012	Tebessa team achieved an overall rank of 1 based on a combination of multi-scale edge-hinge and multi-scale run-length features	Database contained 100 writers.

As described in Table 3, the winner of the ICDAR 2011 Arabic writer identification contest is from the University College London, UK and their state-of-the-art system managed to attain a remarkable average error rate of 0% based on the offline handwritten documents provided from more than 50 writers. Griffin et al. [Crosier and Griffin 2010] used a mainly global texture analysis approach called the orientated Basic Image Feature (oBIF) columns, which extracts and assigns seven features such as dark line on light, light line on dark, dark rotational, light rotational, slop, saddle-like or flat. These seven global features are then passed through a series of six Derivative-of-Gaussian (DoG) filters to perform an edge-detection operation. Furthermore, some local features such as the local orientation were also used. Finally their classification stage made use of a nearest neighbor classifier, which gave the 100% writer identification rate.

Interestingly, it is noteworthy to also point out that the writer identification system with the next highest accuracy for the ICDAR 2011 Arabic writer identification contest attained an accuracy of only 89.2%, but had to involve a computationally expensive approach of using Monte Carlo simulation involving multiple runs of the simulated annealing process. This is in contrast to the simple K- nearest neighbor classifier used by Griffin et al. which is very much computationally light, because Griffin et al. were able to suitably choose a set of good features based on their oBIF approach, and that led to their high accuracy of 100%.

More recently in ICFHR 2012 Arabic writer identification contest which is a follow-up of ICDAR 2011 Arabic writer identification contest, the database previously used in ICDAR 2011 was extended and enlarged to over 200 writers, which increased the challenge of the task at hand. Even with a more challenging task, this contest saw a close to five-fold increase in the number of participating teams to 48 teams, with a total submission of 582 entries.

Even with the larger dataset available, Griffin et al.'s oBIF approach attained the highest identification rate of 95.3%, and was tied at top position with another approach by Zhang et al. Zhang et al.'s approach extracts features at the grapheme level and performs a kernel Principal Component Analysis (PCA) followed by using support vector machines for classification.

Another interesting point to take note is that the top ten teams with the highest identification rate all attained an impressive identification rate above 90%, and a further analysis shows that more than three quarters of the participating teams were able to achieve an accuracy of more than 80%. Even though Arabic writer identification can be said to be in

its infancy stage, having only gained popularity in the past decade, we are already looking at decent writer identification rates for commercialization to be done.

As for writer identification contests on the Latin and Greek alphabet systems, the contest involves two scenarios comprising of a total of twelve experiments (six experiments in each scenario) that involves English, French, German and Greek languages. The overall winner of this contest is the cumulative winner of all twelve experiments. In ICDAR 2011 writer identification contest [Louloudis, Stamatopoulos et al. 2011], the overall winning team is team Tsinghua where Ding et al. uses a grid microstructure feature (GMSF) approach [Xu, Ding et al. 2011] to calculate a probability distribution of microstructures that model the writing styles of the writers. The classification was done using a Chi-square distance metric. This approach performs very well across most experiments, even with the challenging cropped images. In the ICFHR 2012 writer identification contest [Louloudis, Gatos et al. 2012], team Tebessa used a combination of multi-scale edge hinge and multi-scale run length features to model the different handwriting styles of writers [Chawki and Labiba 2010]. This overall winner used a Manhattan distance to classify the English and Greek handwritten documents.

2.8 Summary

In recent years, there has been a fervent increase in the amount of research done in the area of writer identification, as well as an ardent emergence of more publicly-available database and competitions in different languages to promote writer identification, as described in this chapter. This seems to suggest that both the document analysis and forensic science communities are keen to further explore research in writer identification, since there

still remains much room for improvement in increasing the performance of online handwriting recognition systems.

Currently, the state-of-the-art in writer identification can be broadly divided into global-based methods such as textural features, or allograph-based methods such as graphemes, as summarized in Table 1. Unfortunately, such existing methods are not immediately intuitive to interpretation by humans, and thus do not instinctively help human forensic experts in their process of comparison. A more intuitive representative could therefore exist at the character description level.

This area is where this thesis hopes to improve by using writer information at the character description level. This is because such information is complementary and collective information between them offers the ability to improve the performance of online handwriting systems that is not possible with other conventional methodologies, especially in helping human forensic experts in their handwriting analysis task. A framework and understanding of how such character level based descriptors could be used for writer identification are proposed for the rest of this thesis.

3. Writer Identification: Character level

3.1 General Overview

We begin this chapter by briefly describing a simplified example of how our writer identification system works. Let us take for example some samples of handwritten words that are written by two sample writers, writer 1 and writer 2, found in the IRONOFF database [Viard-Gaudin, Lallican et al. 1999], as illustrated in Figure 3-1.

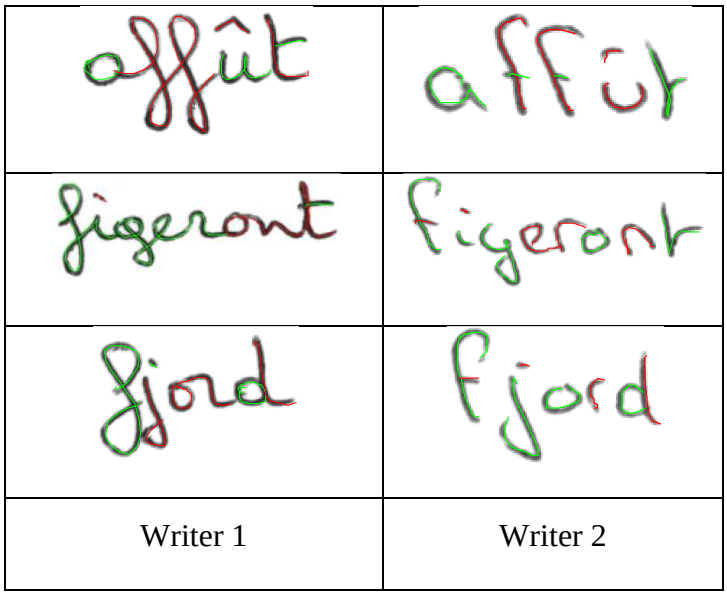


Figure 3-1: A sample of words from IRONOFF database

Let us examine the three sample French words, “affût”, “figeront” and “fjord” as shown in Figure 3-1 in details. As observed, there is great intra-writer stability for both writers where we can tell that the two writers tend to have a dominant style of writing the letter ‘f’ and ‘t’, even across different words. Even to the naked eye, one can easily distinguish between the words written in column 1 to be vastly different from the words written in column 2. We can thus easily imagine that we can make use of a distribution of different allographic prototypes of a letter of the alphabet, and then assigning each instance of the letter to the prototype that most closely resembles the allographic style that is being written. It will then be possible to calculate a distribution of such allographic prototypes to

model and profile the handwriting styles of the writers. In order to further explain how this distribution of allographic prototypes can be used to model handwriting styles, another simplified example is shown in Figure 3-2.

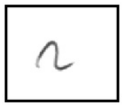
		1	**2**	**3**	**4**
Available allographs for character 'r'					
Reference Documents	frequency vector of writer *1*	0.71	0.00	0.00	0.29
	frequency vector of writer *2*	0.10	0.00	0.60	0.30
Test Document	frequency vector of writer *T*	0.70	0.20	0.00	0.10

Figure 3-2: Simplified example of the proposed methodology

Typically, one could expect different writers to have different styles and writing preferences of writing the letter 'r'. As an illustration shown in Figure 3-2, there are four different prototypes of allographs for 'r' displayed. One writer might have a preferred tendency to write the letter 'r' using the style depicted by allograph one 71% of the time, and depending on various intrinsic or extrinsic factors previously discussed by Huber et al. [Huber and Headrick 1999; Morris 2000], the same writer might write the letter 'r' using the style shown by allograph four 29% of the time. This could be indicative where his style of writing represents a strong dominant handwriting style and that there is a tendency for the writer to use a certain highly skewed distribution of allographs. Comparing this with another writer 2 as illustrated in Figure 3-2, writer 2 might have another slightly less dominant handwriting style compared to writer 1, and utilizes allographs one, three and four 10%, 60% and 30% of the time respectively when writing the letter 'r'. A writer identification system such as the one proposed by this thesis, would therefore be able to profile and store the

writing styles of both writers 1 and 2, and when an unknown document with test writer T as depicted in Figure 3-2, needs to be identified to a certain writer in the database, the writer identification system should be able to compare and classify the distributions of writing styles to the most similar writer. In this case, writer T with a preference for using allograph one 70% of the time, would clearly be classified correctly as writer 1 because writer T has a dominant writing style of allograph one and he doesn't use allograph three at all, as opposed to writer 2.

Similar to the simplified example explained, the proposed framework in this thesis adopts a fuzzy c-means algorithm to create a distribution of frequency vectors that statistically models the handwriting styles of the writers. The distribution of handwriting styles then undergoes classification to identify the writer of the documents. Writer identification is then essentially accomplished by a matching process between the allographic prototypes of the writers in question to templates of allographic prototypes of the reference writers found in the database. The reference documents provided by writer *i* and writer *j* are transformed into a frequency vector based on the distribution of different styles of allographs for '*r*'. The test document from writer *T* would undergo the same transformation. Following this, distances would be computed using the test document's vector with those stored in the reference. The top-1 ranked reference writer would therefore be identified as the writer of the test document.

Having briefly introduced the essential notion of our writer identification system works, the following sections that ensure will then deal with a more in-depth treatment of our writer identification system. The methodology used can be broadly divided into three stages, namely the prototype training stage, the reference and test document labeling stage and finally, the classification stage. A detailed account of each of the three stages in the character prototyping framework is given in section 3.2 below. A description of a Fuzzy C-Means

(FCM) model is used in the proposed framework to improve the results of the character prototyping approach. A sensitivity analysis, followed by the limitations of this system is also discussed in section 3.4.

3.2 Character Prototyping Model

One of the key originalities of the proposed writer identification framework is the usage of an approach that performs an automatic segmentation and labeling of the text at the character level. Handwriting recognition has in recent years, reached a level of maturity where readily available commercial and industrial text recognition engines are able to provide us with reasonably high recognition accuracies [Fujisawa 2008]. We envisage that future research directions in writer identification will be developed based on the foundations of existing technology in order to fully take advantage of the level of maturity and accuracies of the current handwriting recognition technology. This allows us to leverage on increasingly more accurate and efficient state-of-the-art recognition engines to improve the performance of the writer identification systems. Therefore in this thesis, we integrate an industrial text recognition engine into the writer identification framework.

An industrial character segmentation and recognition engine, “MyScript SDK” [MyScript 2012], with the proper linguistic resource (French or English in our case) attached for increased accuracy, has been used for this purpose. MyScript is an off-the-shelf product which was used to integrate into our framework. The proposed framework can be divided into three stages, namely the prototype building stage, the reference and test document labeling stage and finally, the retrieval stage, as illustrated inFigure 3-3.The following sections deal with discussions on using prototypes for writer identification, specifically using characters instead of graphemes as the prototypes for comparative analysis.

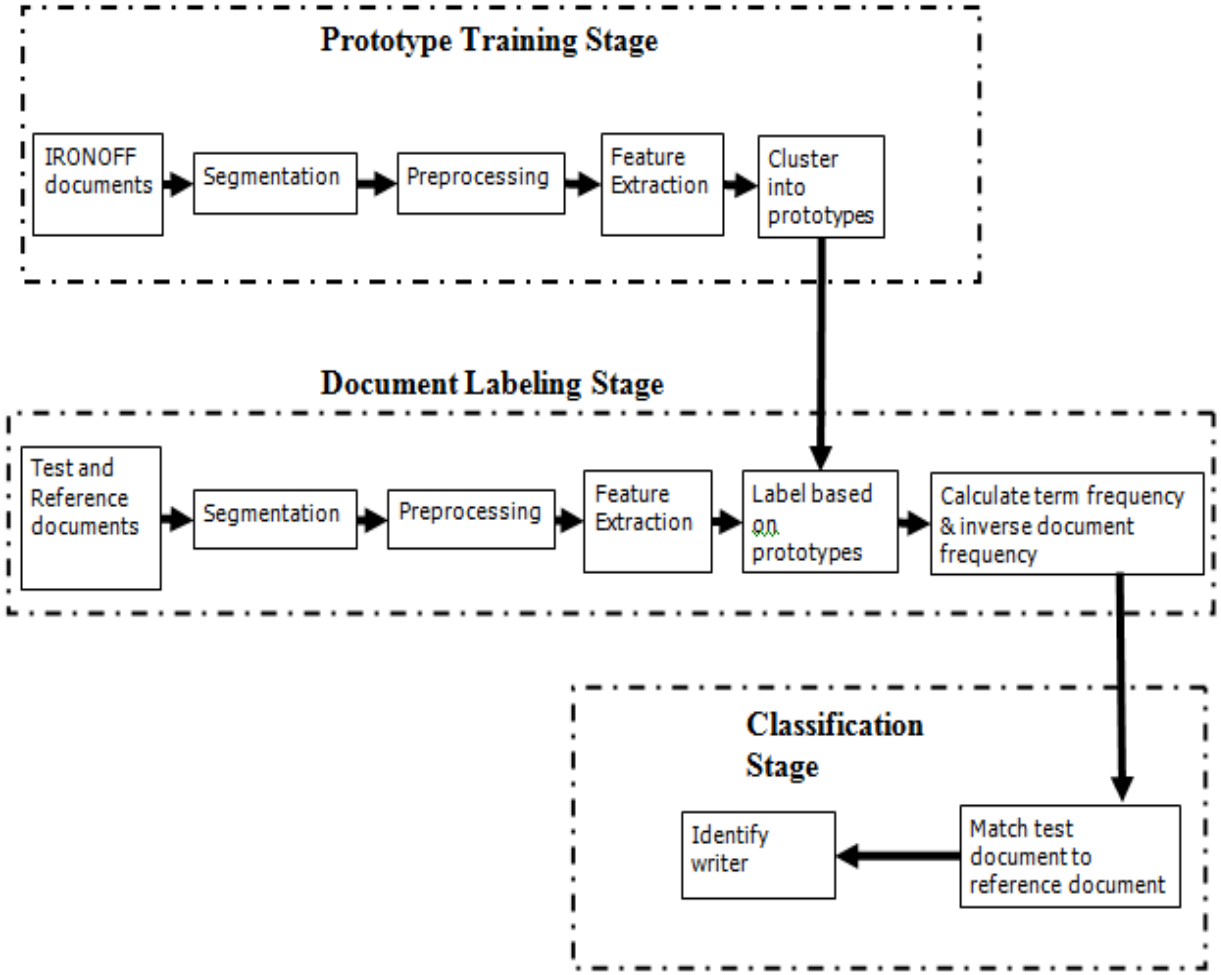


Figure 3-3: Block diagram for proposed methodology

3.2.1 Prototype training stage

During the prototype training stage, prototypes are clustered at the character level, using the IRONOFF database [Viard-Gaudin, Lallican et al. 1999] of 16585 isolated French words that are written by 373 subjects. The purpose of this stage is to build a set of character prototypes by using characters in context of words to model the different allographs of the 26 Latin letters of the alphabet (‘a’ to ‘z’). The industrial text engine automatically segments the isolated words from the IRONOFF database into a total of 87719 characters, after which the

twenty-six respective subsets that have been obtained are used to build allographic prototypes that model the handwriting styles of different writers. Allographic prototypes at the character level can exploit three different types of handwriting variations to perform writer identification, specifically (1) morphological variations, illustrated in Figure 3-4a, (2) directional variations and (3) temporal variations, as illustrated in Figure 3-4b and Figure 3-4c. We therefore make use of the natural existence of such diversities in handwriting to differentiate between different styles of writing during the clustering of our prototypes and in the subsequent stage of document indexing.

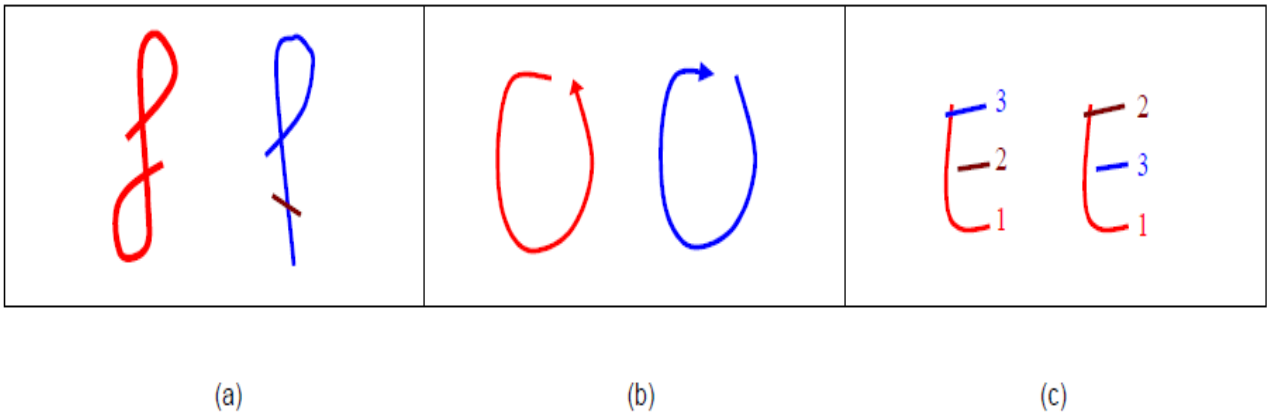
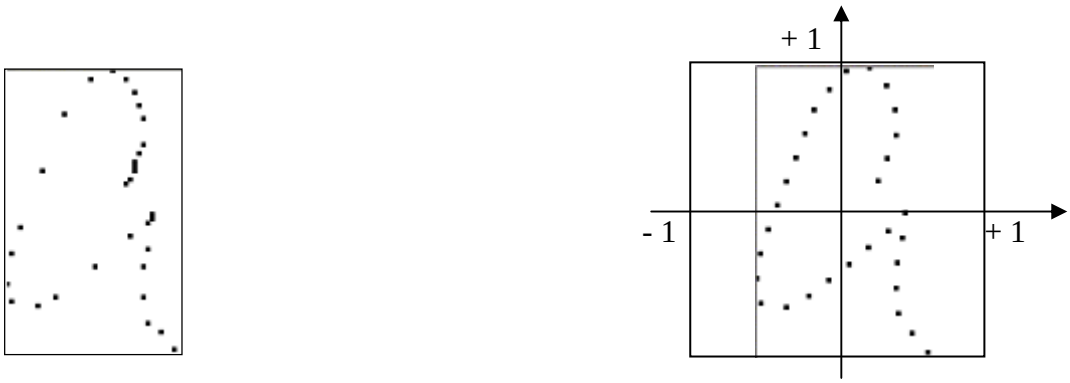


Figure 3-4: (a) Morphological variation, (b) Directional variation, (c) Temporal variation

3.2.2 Feature extraction

After the characters are segmented, the segmented characters then underwent further preprocessing where the size of each segmented character is normalized and resampled to a fixed number of 30 points as displayed in Figure 3-5[Guyon, Albrecht et al. 1991; Vuurpijl and Schomaker 1997]. A process of feature extraction on each of the resampled points is then carried out. The features being used are the x and y co-ordinates (2 features), the directions of the local tangent expressed by cosine and sine of the angle θ (2 features), the local curvatures

Φ (2 features) as presented in Figure 3-6 and the binary Pen-up or Pen-down information (1 feature).



a) Raw trajectory (b) Resized and resampled trajectory with 30 points

Figure 3-5: Resampling of the segmented characters in a fixed number of points

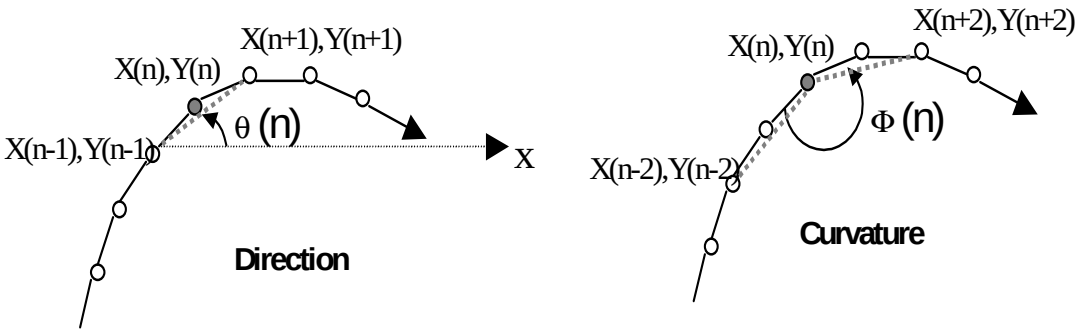


Figure 3-6: Local direction and curvature features

In online handwriting systems, dynamic features such as latent strokes are captured into the system as well. Latent strokes are defined as strokes resulting from pen movements that occur when strokes are being written while the stylus is in the 'pen-up' position, such as in writing the characters 't' or 'i'. The latent stroke is illustrated by the dotted lines in Figure

3-7. Latent strokes contain useful information about the individual writing styles of different writers.

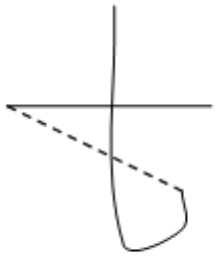


Figure 3-7: An allograph of 't' with the latent stroke in dotted line

This effectively allows us to work using a feature space of dimension: 30 points x7 features = 210. Out of these seven features, four of the features (direction of x-coordinate, direction of y-coordinate, curvature of x-coordinate and curvature of y-coordinate) can in fact be derived from the x & y coordinates. The extracted set of features derived from the IRONOFF isolated words database are then clustered into representative character prototypes, using the well-established k-means clustering algorithm [Han and Kamber 2006]. Figure 3-8 illustrates some character prototypes of 'f' obtained after clustering are shown in Figure 3-8. The dotted lines represent the trajectory when the pen is in the Pen-up position. The k-means clustering algorithm is performed on an alphabet basis, thereby giving us 26 letter of the alphabet x N prototypes. We have chosen N=10 experimentally, which is discussed in details in section 3.4.1.

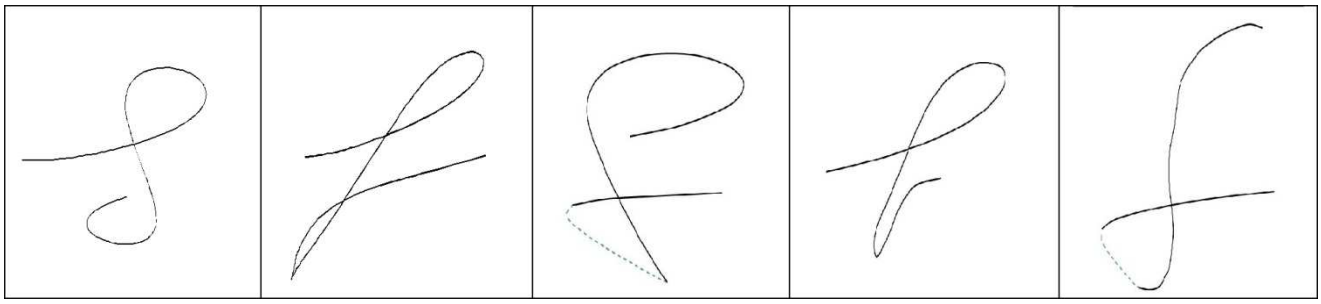


Figure 3-8: Examples of character prototypes of 'f' after clustering

With this first stage, we have a set of representative prototypes of all of the lower case letters of the alphabet. This process can be easily extended to any other symbols, such as upper case letters, punctuation marks, math symbols provided that the segmentation/recognition tool is able to support these symbols. We have focused this study on the lower case letters since they represent the most common way of writing a textual document, additional symbols should be not so significant from that point of view.

It is also worth to note that these prototypes are built on a dataset strictly independent with respect to the writer set on which identification task takes place. In that sense, the proposed method is omni-writer. Should we want to add new writers in the reference dataset, this stage is not affected. Another solution would have been to build the prototypes from the reference writer documents. In that case, the prototypes would be more accurate but very specific to these writer set. We expect to cover correctly all the possible writing styles using a large independent dataset. In this experiment, 373 writers from the IRONOFF dataset are considered.

3.2.3 Document labeling stage

In the document indexing stage, the sets of reference and test documents are automatically segmented and recognized into characters by the industrial engine; similar to what was undertaken in the prototype building stage. More details on the data set used are given in section 3.3. The main purpose of this indexing stage is to represent the handwriting style of each writer with a statistical distribution of term frequencies (tf) and inverse document frequencies (idf) used in the IR model by mapping the features in the documents to the set of prototypes built previously during the earlier prototype building stage. Further

details of tf and idf will be covered in the next section 3.2.4. This mapping of the segmented and recognized characters into a statistical distribution of prototype frequencies is accomplished by a fuzzy c-means algorithm.

3.2.4 Fuzzy C-means Model

The previous prototype building stage serves to identify common individual handwriting traits and styles to create individual prototypes at the character level. Subsequently, the document indexing stage now utilizes these prototypes to estimate individual statistical distributions of handwriting styles for each of the test and reference documents in the database. Based on the results of the distributions, they provide statistical information about the handwriting styles of each writer. Our proposed method adopts a fuzzy c-means algorithm which uses a kernel function to estimate the probability that a character *x* has been generated by a prototype *p*. Three different kernels are proposed and described in equation (3-1) to equation (3-3). The first one is an exponential kernel with a tuning parameter β to adjust the selectivity of the exponentials, the second one is the Gaussian kernel and the third one is the inverse kernel [Hoppner, Klawonn et al. 1999].

$$P_{\alpha}(p_k \mid x) = \frac{\exp(-\beta \times dist(p_k, x))}{\sum_{k'=1}^{N} \exp(-\beta \times dist(p_{k'}, x))} \tag{3-1}$$

$$P_{\alpha}(p_k \mid x) = \frac{\exp\left(-dist^2(p_k, x)\right)}{\sum_{k'=1}^{N} \exp\left(-dist^2(p_{k'}, x)\right)} \tag{3-2}$$

$$P_{\alpha}(p_k \mid x) = \frac{1/dist(p_k, x)}{\sum_{k'=1}^{N} 1/dist(p_{k'}, x)}, x \neq p_k \quad (3\text{-}3)$$

$P_{\alpha}(p_k \mid x)$ is the probability that a given segmented character *x*, which has been recognized as of the alphabet set $\alpha, \alpha \in \{'a','b',...,'z'\}$, is assigned to prototype $p_k, k \in [1, N]$. This represents the partial membership of prototype p_k discussed earlier. The function $dist(p_k, x)$ represents the Mahalanobis distance. In equation (3-1), β is a tuning parameter which has been experimentally set to be 0.01 and *N* is the number of prototypes used. Equation (3-2) describes the Gaussian kernel function where the distribution of the feature vectors was assumed to be Gaussian with zero mean and unit variance. Equation (3-3) describes the formulation for the inverse kernel function where the notations have the same meaning as in equation (3-1).

Figure 3-9 gives an illustration of the behaviour of the inverse kernel function described by equation (3-3) in a situation where 4 prototypes are available in a 2D feature space and considering a trajectory for the test character *x* following a spiral curve in this same space as displayed in Figure 3-9. The 4 prototypes p_1 (+1, 1), p_2 (-1, +1), p_3 (-1, 1) and p_4 (-1, +1) are located at the 4 corners of a square, where the co-ordinates (*i*,*j*) are defined in the Cartesian co-ordinate system. The spiral which defines the positions of a test character is indexed by the parameterized value of angle θ varying from 0° (where the corresponding point (0, 0) is at the center of the square), to 720° (with a corresponding point having (2, 0) as coordinates).

When the test point is just at the center of the 4 prototypes, each of the 4 prototypes gives the same partial membership (0.25) as defined by the inverse kernel function in Figure

3-9 whereas as the point is moving along the spiral, each of the 4 prototypes contributes differently. When the test point is very close to one of the prototypes, this prototype takes nearly all the membership mass, this is the case for instance for the values of the angle of $45^o + n \times 90^o, n \in \mathbb{N}$. Take for example if n=3, where θ=315^o as illustrated by the vertical line in Figure 3-10, the test character *x* will be defined at the position of the spiral nearest to prototype p_4 as shown by Δ in Figure 3-9. The behaviour of the kernel function in this case, assigns the largest partial membership of 0.53 to prototype p_4, and lower values of partial membership of 0.17, 0.125 and 0.17 to prototypes p_1, p_2 and p_3 respectively. However, this kind of greedy winner-takes-all approach might not necessarily give better results. A further detailed comparison and analysis of the performance of these three Fuzzy C-Means kernel described in equations (3-1), (3-2) and (3-3) will be further discussed in the later sections

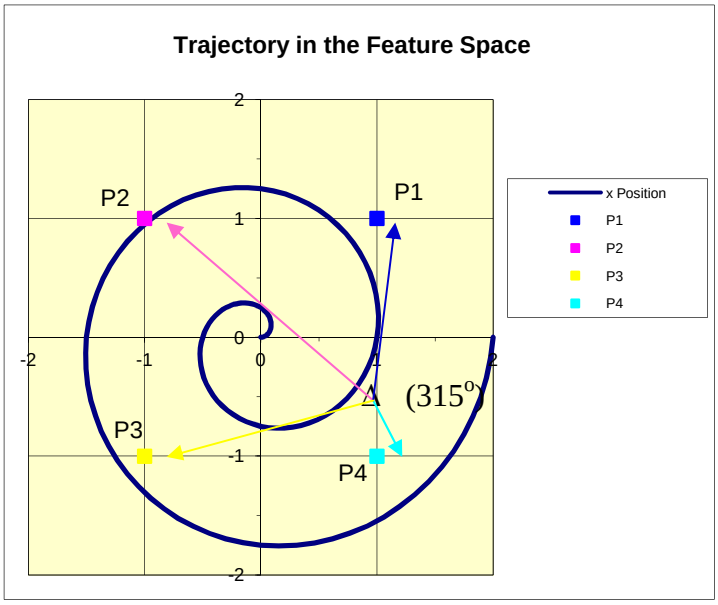


Figure 3-9: Feature space with 4 prototypes p_1 (+1,-1), p_2(-1, +1), p_3 (-1,+1),p_4(+1,-1) and different positions of a test sample x following a spiral indexed by an angle [0°, 720°]

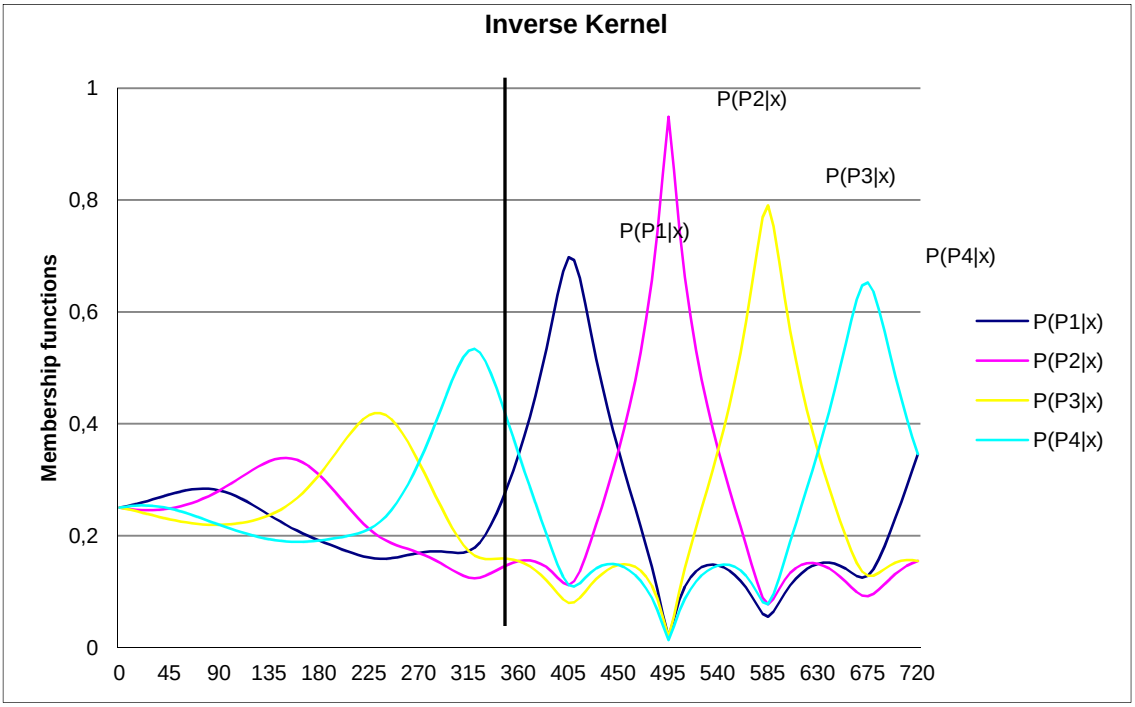


Figure 3-10: Partial membership inverse distance kernel functions for different positions with 4 prototypes, p_k

In this thesis, we have adapted the tf-idf model first proposed by Salton et al. [Salton and Buckley 1988] into a character prototyping approach. This means that our definition of tf-idf will be on a character basis, depending on the letter of the alphabet. Through our experiments, which will again be discussed later, we observe that a modification to tf-idf initially proposed by Salton et al. is in order due to the absence of certain letters of the alphabet in most documents, noteworthy letters such as 'w' and 'q'. This implies that we should assign characters that are more probable to be similar to a certain prototype p_k a higher weightage. Hence, the adaption of Salton et al.'s term frequency and inverse document frequency follows equation (3-4) and equation (3-5) respectively.

$$tf_{\alpha,k} = \frac{1}{M}\sum_{m=1}^{M} P_\alpha\left(p_k \mid x\right) \tag{3-4}$$

$$idf_{\alpha,k} = \log \frac{\sum_{k'=1}^{N} \sum_{i=1}^{R} tf_{\alpha,k',i} + \varepsilon}{\sum_{i=1}^{R} tf_{\alpha,k,i} + \varepsilon} \tag{3-5}$$

As described in equations (3-1) to (3-3), characters from the reference and test documents are assigned a partial membership to the prototypes based on their distance metric to the prototypes. Therefore, characters which lie further away from certain prototypes are assigned a lesser degree to that particular prototype. $P_\alpha(p_k \mid x)$ is then used to calculate on a letter of the alphabet basis, the distribution of frequency vectors; the term frequency as described in equation (3-4) and the inverse document frequency as described in equation (3-5), to be used during classification. Hence, $P_\alpha(p_k \mid x)$, $tf_{\alpha,k}$ and $idf_{\alpha,k}$ are all calculated on a letter of the alphabet basis. In equation (3-4), M is the number of characters corresponding to the letter of the alphabet *α.* In equation (3-5), R is the number of reference writers and ε is a small value to prevent any numerical problems. Then it is possible to compute, as presented in Table 4, the term frequencies, $tf_{\alpha,k,i}$, which represent the probability that reference writer *i* uses prototype p_k for writing alphabet letter α, with $\sum_{k=1}^{N} tf_{\alpha,k,i} = 1$. From these values, which are computed on the reference document writer set, it is possible to compute the inverse document frequencies, $idf_{\alpha,k}$, which are a measure of the uniqueness of the usage of prototype p_k for writing alphabet letter α among the different writers of reference set.

Table 4: Building of the term frequencies and inverse document frequencies.

α		‘a’				‘b’				α	‘z’			
k		1	2	…	N	1	2	..	N	…	1	2	..	N
i	1	$tf_{'a',1,1}$	$tf_{'a',2,1}$	…	$tf_{'a',N,1}$	$tf_{'b',1,1}$	$tf_{'b',2,1}$	..	$tf_{'b',N,1}$	…	$tf_{'z',1,1}$	$tf_{'z',2,1}$	..	$tf_{'z',N,1}$
	2	$tf_{'a',1,2}$	$tf_{'a',2,2}$	…	$tf_{'a',N,2}$	$tf_{'b',1,2}$	$tf_{'b',2,2}$	..	$tf_{'b',N,2}$		$tf_{'z',1,2}$	$tf_{'z',2,2}$	..	$tf_{'z',N,2}$
	..			…				..		$tf_{\alpha,k,i}$			..	
	R	$tf_{'a',1,R}$	$tf_{'a',2,R}$	…	$tf_{'a',N,R}$	$tf_{'b',1,R}$	$tf_{'b',2,R}$	..	$tf_{'b',N,R}$		$tf_{'z',1,R}$	$tf_{'z',2,R}$	..	$tf_{'z',N,R}$
		$idf_{'a',1}$	$idf_{'a',2}$	…	$idf_{'a',N}$	$idf_{'b',1}$	$idf_{'b',2}$	..	$idf_{'b',N}$	$idf_{\alpha,k}$	$idf_{'z',1}$	$idf_{'z',2}$	..	$idf_{'z',N}$

The physical significance of our adapted term frequency provides statistical information on the different writing styles of the writers at the character level. The term frequency presents a distribution that can be used to match the test document in question with all the reference documents in the database. Therefore, the identified writer will be the one with the most similar term frequency between the set of test and reference documents. Each and every document is being represented by a vector of 26xN terms consisting of 26 letters of the alphabet and N number of prototypes for the term frequency. The inverse document frequency takes into account the relative importance of one prototype compared to another, across all the documents present during training or testing.

It can be shown from equation (3-5) that more emphasis is being placed onto those prototypes that are infrequent. This suggests that such particular styles of writing is distinct and is indicative of the uniqueness and importance of such character prototypes in identifying

writers. Both the term frequency vector and the inverse document frequency are then used in the classification stage.

It can be seen that with our fuzzy c-means approach, a character is not labeled to one particular prototype but rather, each character is distributed over every prototype of the corresponding letter of the alphabet. The experimental results obtained serves as strong evidence to attest that our proposed fuzzy c-means approach for assigning prototypes during the document indexing stage yielded a higher level of accuracy. Finally, in the last stage of retrieval, the frequency vectors are used for classification in order to identify the writer corresponding to the test document.

3.2.5 Classification stage

The classification stage is then performed using a k-means classifier based on a distance measure given below. The distances are then ranked and the identification of the reference writer *i*, corresponding to the test writer *j* is taken as the writer with the shortest Euclidean distance as calculated in equation (3-6). A comparison of different distance metrics (Kullback-Leibler divergence and the Chi-square distance) will also be discussed in the later sections 3.3.5 and 3.3.6 respectively to see how their performance compares against the Euclidean classifier.

$$\text{Euclidean Dist (reference writer } i\text{, test writer } j) = \sum_{\alpha='a'}^{'z'} \lambda_\alpha \times \sqrt{\sum_{k=1}^{N} \left[idf^2{}_{\alpha,k} \left(tf_{\alpha,k,i} - tf_{\alpha,k,j} \right)^2 \right]} \qquad (3\text{-}6)$$

where λ_α is a weighing factor to normalize the amount of characters present in each document. The formulation for λ_α is described in equation (3-7). λ_α represents a reliability

factor of using any one particular letter of the alphabet. α takes into account the effect of different amounts of the presence of a particular letter of the alphabet in the document. For example, if the letter 'z' seldom appears in the particular document or does not appear at all, instances of the letter 'z' used in calculating the term frequencies may cause distortions in the statistical distribution of the handwriting style. Likewise, having more instances and samples of a letter of an alphabet in the document increases the confidence and reliance of using this letter of the alphabet in the classification and identification of the writer.

$$\lambda_{\alpha} = \sqrt{\frac{x_{\alpha,i}}{C_{\alpha,i}} \times \frac{x_{\alpha,j}}{C_{\alpha,j}}} \tag{3-7}$$

Where $x_{\alpha,i}$ is the number of characters of letter *a* in reference document *i*.

$x_{\alpha,j}$ is the number of characters of letter *a* in test document *j*.

$C_{\alpha,i}$ is the total number of characters of all the letters in reference document *i*.

$C_{\alpha,j}$ is the total number of characters of all the letters in test document *j*.

3.3 Character Prototype Model Evaluation

3.3.1 Data Acquisition

The database of online handwritten documents was collected using a digital pen and digital paper technology, as shown in Figure 3-11 (image taken from [PCMag 2012]). The digital pen used in our experiments was the Nokia SU-1B, and the digital paper used in our experiments was an Esselte A5 digital notepad, where both makes use of the Anoto digital pen and paper technology. The unique dot pattern printed on the digital paper allows for quick and easy capturing of online information such as dynamic data, pen-up and pen – down data, writing speed, writing acceleration and even pressure when coupled with a digital pen such as the Nokia SU-1B digital pen.

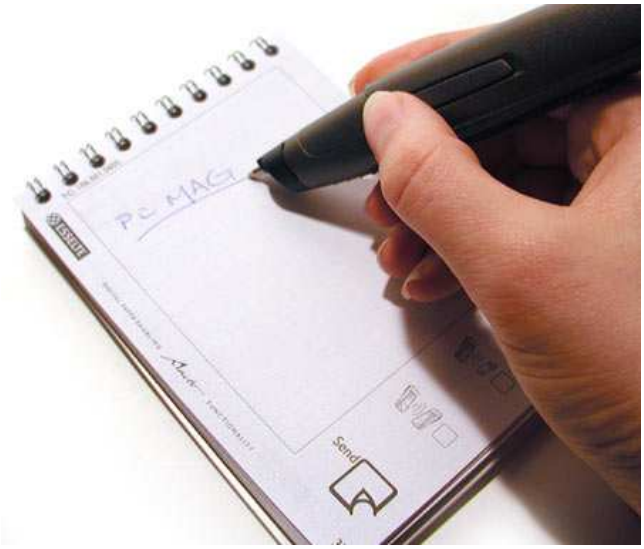


Figure 3-11: Nokia digital pen (Nokia SU-1B) and Esselte digital paper.

The advantage of the digital pen, allows for handwritten data to be captured as an online handwritten document directly into the pen's memory. This is accomplished via a combination of ordinary ink, and a digital infrared camera concealed near the ball-point of the digital pen, as shown in Figure 3-12 (image taken from [Gizmowatch 2012]). This is different from a typical digitizer-and-tablet input where no special digital paper is required to capture the online handwritten text. However, the advantages of such a digital pen and digital paper technology is that it allows the writer to use exactly the same natural experience of writing using a normal pen and paper, as well as it significantly speed up the capture of large amounts of data, and help accelerate the process of data acquisition.

Figure 3-12: Features of a digital Pen, Nokia SU-1B

3.3.2 Database

Online handwritten documents were collected from 120 writers, where each writer wrote two documents: one is considered as a reference document and the other one is taken as a test document. The contents of the two documents are different where the length of the reference and test documents varies from 86 characters to 972 characters. Each writer has to copy a given text passage taken from a variety of sources such as literacy works, financial news and short notices. In addition to copying from a given text, the writer has to provide his/her own text. This allows a large variety of content to exist in the database, which does not impose any constraints on the dependency of the domain.

These reference documents and test documents that were collected belong to a separate dataset from the IRONOFF dataset. The rationale for this is that the IRONOFF dataset is primarily used to build the set of allographic prototypes during the previous prototype building stage. Furthermore, a separate database needed to be collected because the

IRONOFF database contains only isolated words and hence are not representative of actual online documents. Therefore as a consequence, the prototype set is generic and independent with respect to the actual reference database of documents from which the writer is to be identified. The advantage of building the prototypes from an independent database is that the prototypes need only be trained once, thus making it much more robust and scalable to be deployed across a large number of systems.

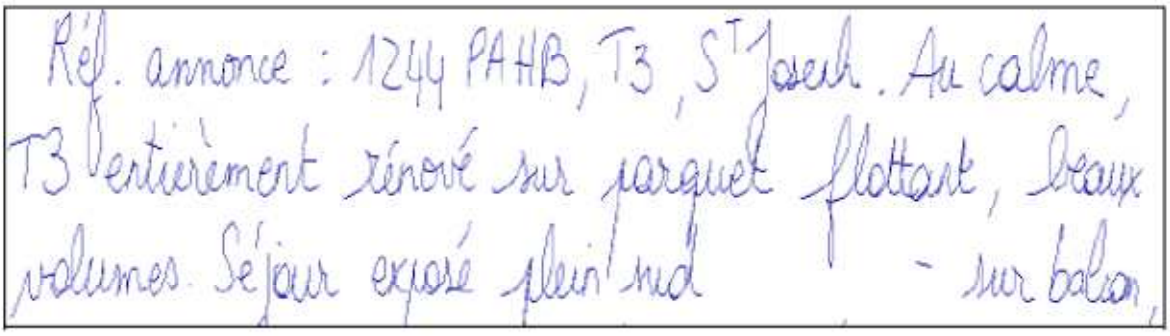
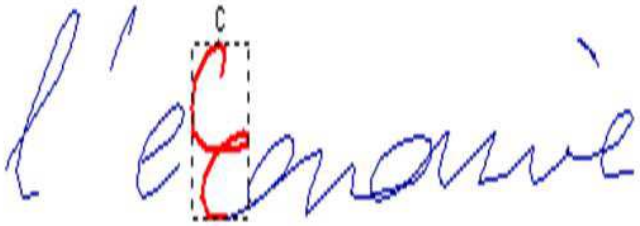


Figure 3-13: Example of a text passage from a reference document.

Figure 3-14: Example of correct segmentation and labeling at the character level by the industrial text recognizer.

a) Text to be recognized

b) Correctly recognized text “economie” by industrial text recognizer



c) Incorrect segmentation

Figure 3-15: Example of the presence of segmentation errors

Figure 3-13 illustrates a sample of a text passage found in a reference document and Figure 3-14 depicts a word that has been recognized by the industrial engine, which will also then automatically provide the segmentation into characters. For example, as seen in Figure 3-14, the industrial engine first recognized the word ‘flottant’ from the text passage given in Figure 3-13. Next, all the characters of this word ‘flottant’ needed to be segmented into the respective characters ‘f’, ‘l’, ‘o’, ‘t’, ‘t’, ‘a’, ‘n’ and ‘t’ so that we can have difference instances of the various characters. This process of segmentation into characters is also done by the industrial engine. However, the segmentation and recognition capabilities are not perfect. There exist instances when the recognition is correct, yet the segmentation has been performed wrongly. For example, as illustrated in Figure 3-15b, the word “économie” might have been recognized correctly, but due to the presence of the accent found in ‘é’, the industrial text engine has wrongly segmented the following character, ‘c’, as shown in Figure

3-15c. This creates an instance of noise in the allograph set and will distort the result of the prototype distribution computation.

We have found the character recognition accuracy of the industrial text engine to be 91% on the whole sets of reference and test documents, which indicates that 9% of the characters have been assigned to the wrong alphabet letter. This result was automatically computed by taking the true label of the documents and comparing them with the output of the industrial recognition engine. The wrongly labeled characters are included in the experiment because this will reflect a real-life scenario, where it is realistic to expect a certain degree of recognition errors. It is worth to note that the recognition engine is already trained and ready-to-use, hence no specific training is required on the documents.

3.3.3 Fuzzy C-means Model (FCM) Discussion

The experiments were conducted on a set of 120 test and 120 reference documents from 120 different writers. The proposed methodology using fuzzy c-means model (FCM) to label the test and reference documents to the prototypes with the Euclidean distance metric resulted in a high accuracy of 98.3% for writers that are ranked correctly in the top-1 position. This translates into a misclassification error of two misclassified writers out of 120, with both of them being classified in the top-2 position. This indicates that the writer identification system has confused the misclassified documents with only one other document from a different writer. Comparisons with previous results obtained by Chan et al. [Chan, Viard-Gaudin et al. 2008] on the same dataset, who also performed writer identification based on character prototyping, show a significant improvement over their proposed methodology. An accuracy of 96.7% (four misclassified writers out of 120) was obtained using their method where the four writers were wrongly identified ranked at top-2, 4, 9, and 12 positions. This means that their writer identification system has confused the misclassified

documents with up to 11 other documents. Therefore, our proposed methodology is able to perform with significantly higher accuracies.

This improvement over Chan et al.'s results [Chan, Viard-Gaudin et al. 2008] can be explained as follows. Their methodology hinges on the concept that each character can only be assigned to one particular prototype for which a distribution of handwriting styles is built. This is flawed in reality because overlapping handwriting styles for different writers can be commonly found. Our observations reveal that there are numerous instances when the characters are close to more than one prototype in the vector space. This can be explained by the fact that a writer can have strong, dominant handwriting style and weak handwriting styles.

Weak handwriting styles change according to various circumstantial and temporal states [Huber and Headrick 1999; Morris 2000], which can affect the strong dominant handwriting style and lead to reminiscence of multiple overlapping handwriting styles. For such instances, the discrete allocation of prototypes used by Chan et al. does not yield good results. A writer whose dominant handwriting style does not fit well into an existing prototype will be weakly modeled using their approach. Therefore, in our proposed methodology, each character is not just assigned to one particular prototype, but rather, each character is assigned a certain degree of all the prototypes depending on how close they are to that prototype, allowing us to realize a higher accuracy. The more similar the character is to a certain prototype, the greater the degree that prototype has on the character.

3.3.4 FCM Kernel Evaluation

Experiments were also conducted using different kernel functions [Hoppner, Klawonn

et al. 1999] for the fuzzy c-means algorithm to determine the kernel function that can perform best in our writer identification system. We compared the use of three different FCM kernel functions as previously described in equations (3-1), (3-2) and (3-3). Table 5 shows the comparison in performance of the writer identification among using the three different FCM kernel functions and our results show that the exponential kernel function performs the best in our writer identification system.

Table 5: Performance of the Fuzzy c-means algorithm using different kernel functions

Identification rate using exponential kernel function	99.2%
Identification rate using Gaussian kernel function	97.5%
Identification rate using inverse kernel function	96.7%

It can be seen that the inverse kernel function performs poorly, which can be explained by the poor behavior of such kernel functions as it approaches the centroid of the clusters. The results obtained using the Gaussian kernel function performs better than the inverse kernel function, notably because the Gaussian kernel function has the advantage of a smooth distribution as it approaches the centroid of the clusters.

This poor asymptotic behavior of the inverse kernel function is further illustrated in Figure 3-10 at $\theta=495^{o}$ and 585^{o} where the character approaches very close to the centroid and gives a partial membership that is completely dominated by the nearest prototype, with the other prototypes being given a partial membership that is much lower. Comparing this to handwriting styles, that would indicate that this inverse kernel function does not sufficiently represent weak handwriting styles of the writers as the character approaches situations close to the centroid.

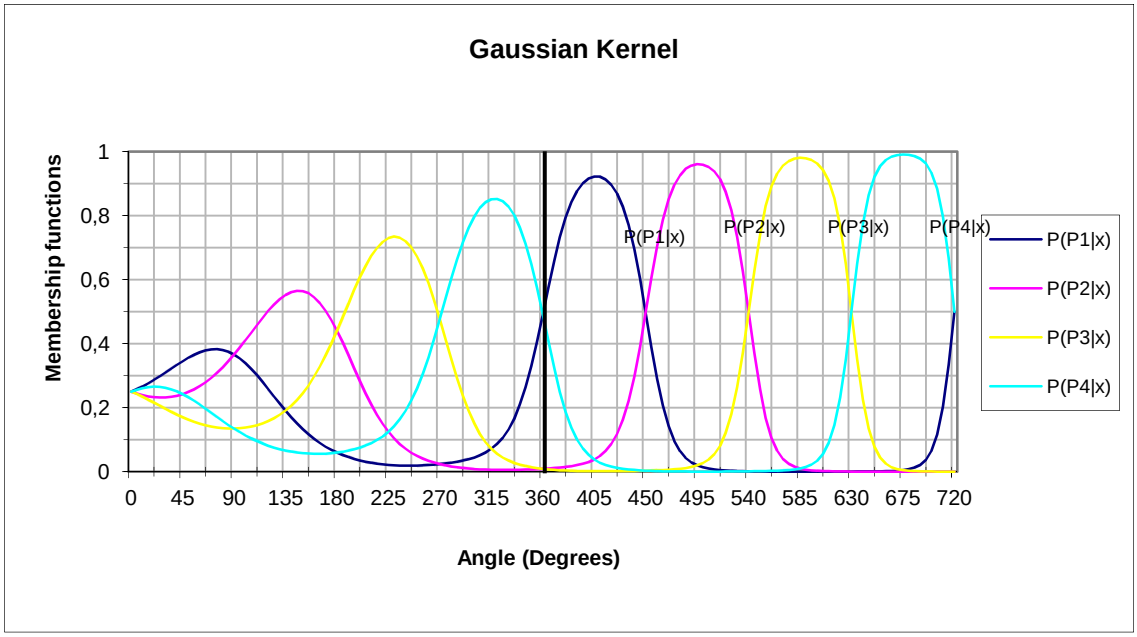


Figure 3-16: Partial membership Gaussian distance kernel functions for different positions with 4 prototypes, P_k

A further examination of the Gaussian kernel function shown in Figure 3-16 reveals that such a kernel function still allows for a partial membership to be allocated to all the prototypes if the relative position of the other prototypes are nearby. This is in stark contrast to the inverse kernel function discussed earlier, where even if some prototypes are nearby, as long as the relative location of the character x is exactly at the position of the prototype (ie. $x = p_k$), then that prototype will take all the weight. The choice of a Gaussian kernel function therefore allows the moderation of a partial membership function depending of the relative locations of the prototypes to one another in the feature space, ie. if other prototypes are far away, the considered prototype takes most of the weight, but not if some prototypes are nearby.

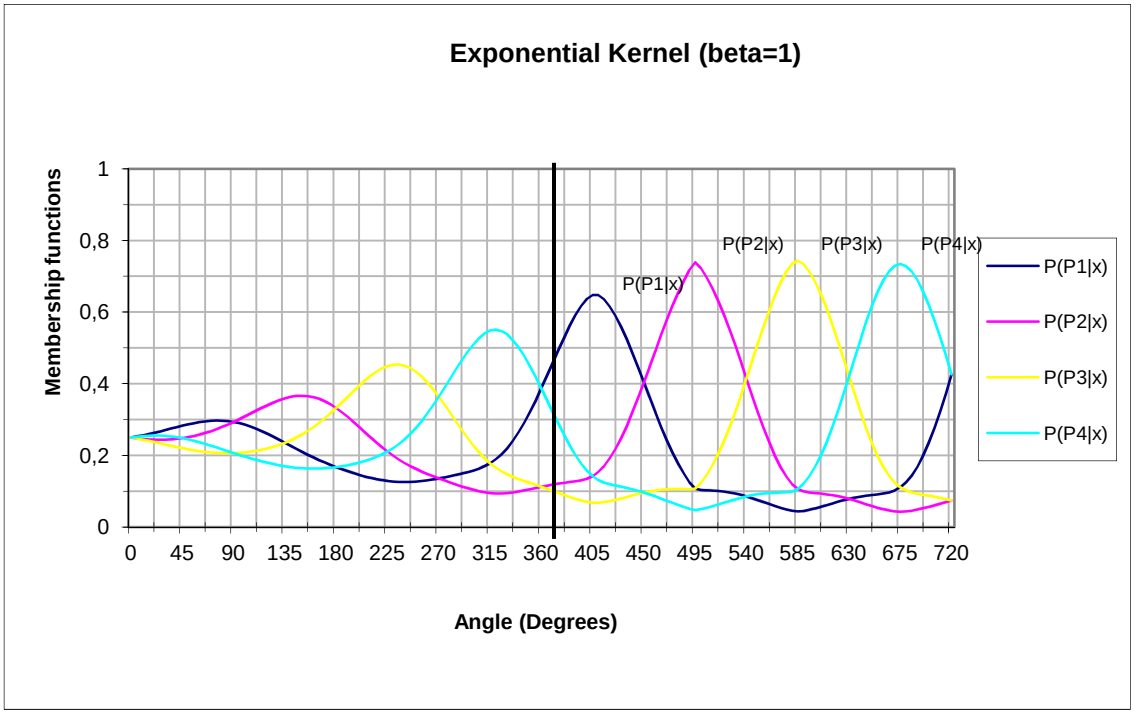
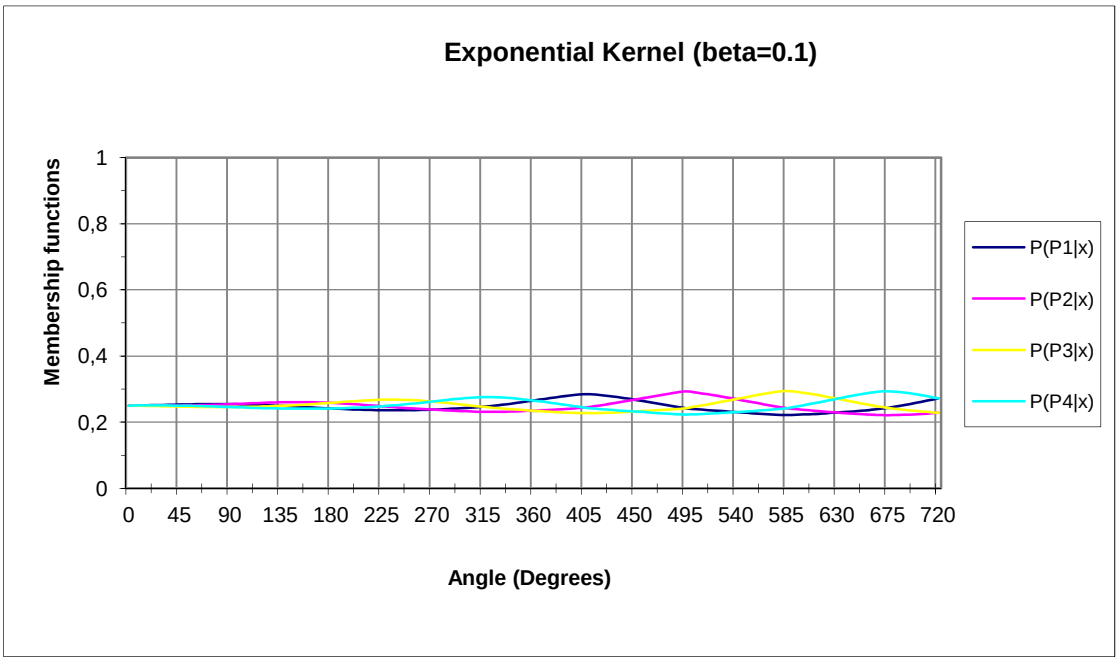
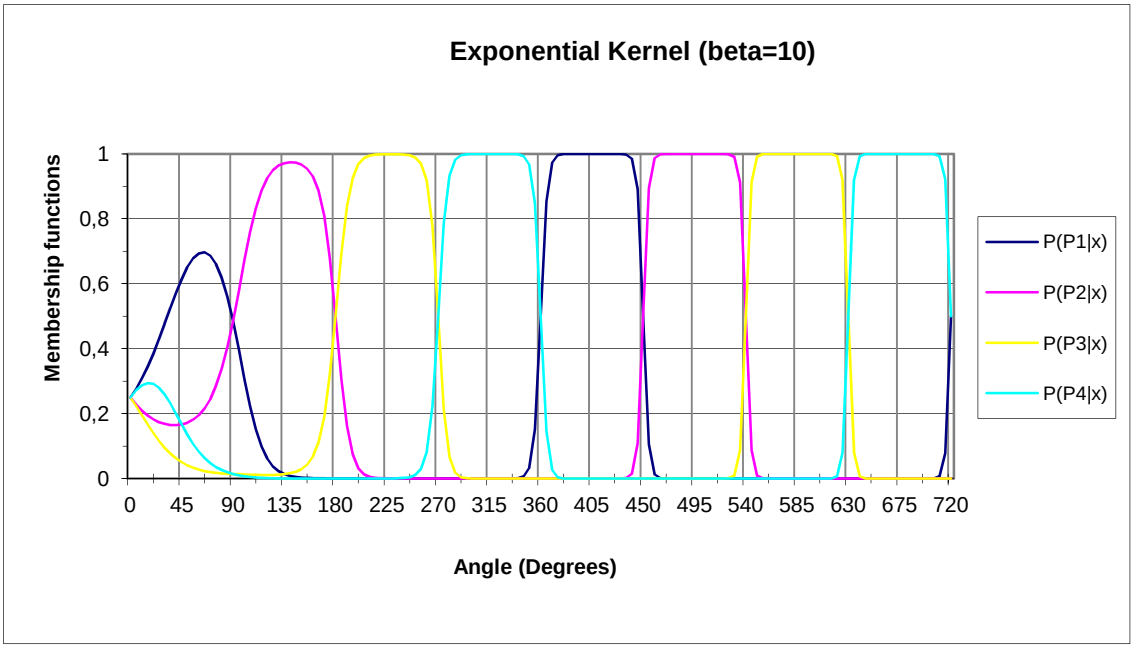


Figure 3-17: Partial membership exponential distance kernel functions for different positions with 4 prototypes, P_k

Our results indicate that the exponential kernel function was able to achieve the best result of 99.2% accuracy in our writer identification system. This can be seen in Figure 3-17 where because of the specific value of beta that is small enough to cover a larger spatial range in the feature space domain, the exponential kernel function does not completely take the whole partial memberships for character points that are very close to a particular prototype as indicated by the vertical line at $\theta=315^{o}$. Instead the kernel function still assigns a smaller partial membership to the other prototypes that are further away from the character point. A partial membership of 0.55 to prototype p_4, and lower values of partial membership of 0.17, 0.18 and 0.1 to prototypes p_1, p_2 and p_3 respectively is assigned using this exponential kernel function. Comparing this result of the exponential kernel function and the Gaussian kernel where the partial memberships are 0.072, 0.006, 0.072 and 0.85 for prototypes p_1, p_2, p_3 and p_4 respectively, is evidence that the bandwidth of the kernels with respect to the

feature space heavily influences how the partial membership of the prototypes are being modeled.





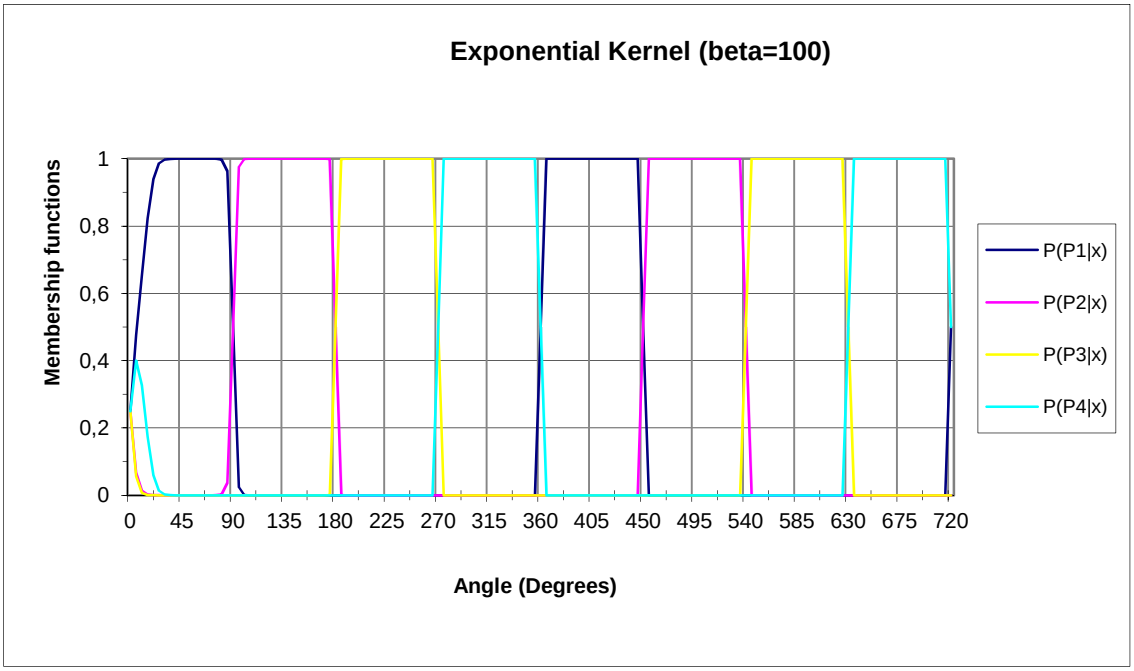


Figure 3-18: Partial membership exponential distance kernel functions for different

An in-depth understanding of how the beta behaves in the exponential kernel function allows one to be able to fine-tune the degree of partial membership of the kernel function. The effect of different values of beta can be observed in figures . The beta parameter controls the spread of the different exponential membership functions, and determines how fast the membership function saturates. A low value of beta results in a slow saturation of the membership functions and this essentially results in an uniformly distributed partial membership to all the prototypes. Therefore, too low a value of beta will not be able to effectively distribute the partial membership to the different prototypes. Conversely, a large value of beta will see the membership functions quickly saturating to the case where it becomes the case of a winner-takes-all scenario, and the partial fuzzy C-means model essentially becomes a discrete assignment of membership. Therefore, the experimental tuning of the beta parameter becomes important in controling the spread of the partial membership assigned by the exponential kernel function.

3.3.5 FCM with Kullback-Leibler Divergence

Extensive testing and experimentations were also performed on using other metrics in obtaining more representative distributions of the writers' handwriting styles. The Kullback-Leibler (KL) divergence, also known as information gain and frequently utilized in the field of information theory, is one such useful metric in measuring how different one distribution is from another. The formula for KL divergence is given in equation (3-8):

$$\sum p(x) \log \frac{p(x)}{q(x)} \tag{3-8}$$

where *p(x)* is the distribution of term frequencies for a test writer and *q(x)* is the distribution of term frequencies for a reference writer. It can be seen from the above formula that the KL divergence is asymmetric. This means that the KL divergence for *p(x)* to *q(x)* is different from that of *q(x)* to *p(x)*. However, it should be noted that in our case, we are only interested in obtaining an idea of how different the distribution between reference writers and test writers is, in order to identify the reference writer corresponding to the test document in question. Hence, it is not necessary for us to factor in the asymmetric behavior of KL divergence.

Applying equation (3-8), we obtain:

$$\text{KL(p,q)=KL Div (reference writer } i, \text{ test writer } j) = \sum (idf \times tf)_j \log \frac{(idf \times tf)_j}{(idf \times tf)_i} \tag{3-9}$$

But since the inverse document frequency is independent of the identity of the writer, (i.e we have $idf_i = idf_j$), and we can rewrite the formula as:

$$\text{KL(p,q)=KL Div (reference writer } i, \text{ test writer } j) = \sum (idf \times tf)_j \log \frac{tf_j}{tf_i} \qquad (3\text{-}10)$$

The summation has to be done on all the bins (1 to N and 'a' to 'z') of the distributions, hence we obtain:

$$\text{KL Div (reference writer } i, \text{ test writer } j) = \sum_{\alpha='a'}^{'z'} \sum_{k=1}^{N} \left[idf_{\alpha,k} \times tf_{\alpha,k,j} \log \frac{tf_{\alpha,k,j} + \varepsilon}{tf_{\alpha,k,i} + \varepsilon} \right] \qquad (3\text{-}11)$$

In this formula, we have to take care of possible zero values for some bins if the estimated distributions when a prototype is not used in a reference document ($tf_{\alpha,k,i} = 0$). To prevent any numerical problems, as with equation (3-5), we have added a small value (ε), a technique which is inspired from the interpolated backing-off method to take into account the sparsity of certain distributions, where this technique stems from its origins in language modeling techniques [Witten and Bell 1991; Kneser and Ney 1995].

In our data, there are numerous instances when the writer distributions, *p(x)* or *q(x)* are zero. The physical significance behind a value that is zero signifies that there are no instances of the writer having a particular style of writing. The results will be inaccurate if this is ignored. Hence, the formula for KL divergence was adapted to take this into account.

In our experiments, ε is set to 1×10^{-9} for the purpose of our simulations. Equation (3-11) takes into account the spikes in the distribution caused when either *p(x)* or *q(x)* are zero and effectively smoothens the distribution. The identification rate obtained when this KL divergence was used for classification was a low 91.7%, as shown in Table 6. This is even lower than the results of 96.7% obtained by Siew et al.'s 1-nearest neighbor approach. This could be due to the fact that the KL divergence results in a larger distance values compared to

when Euclidean distance was used. Hence, the Euclidean distance might be a better choice for this writer identification system.

Table 6: Performance of writer identification using different distance metrics

Fuzzy C-Means Model (FCM) approach		
Euclidean distance eq. (3-6)	**KL divergence** eq. (3-11)	**Chi^2 distance** eq. (3-12)
98.3%	91.7%	99.2%

3.3.6 Chi-square Distance

Experiments were also performed using the Chi-square distance, as described in equation (3-12), as a different metric for the minimum distance classifier to determine the best performing metric for our writer identification system. We can observe from Table 6 that the Chi-square distance measure outperforms the Euclidean measure in our writer identification system, achieving a top-1 writer identification rate of 99.2%. This is equivalent to a misclassification error of only one misclassified writer, with the misclassified writer being in the top-2 position.

$$\text{Chi}^2\text{Dist (reference writer } i\text{, test writer } j) = \sum_{\alpha='a'}^{'z'} \sum_{k=1}^{N} \frac{idf_{\alpha,k}\left(tf_{\alpha,k,i} - tf_{\alpha,k,j}\right)^2}{tf_{\alpha,k,i} + tf_{\alpha,k,j}} \qquad (3\text{-}12)$$

The rationale, which can explain the better result obtained with the Chi-square distance, is that the Chi-square distance considers a relative difference between the two components of the distributions instead of an absolute difference as with the Euclidean

distance. This relative difference is more meaningful with respect to the style of writings that we would like to distinguish. Our results also support Schomaker et al.'s results [Schomaker and Bulacu 2004] that the Chi-square measure outperforms the Euclidean distance measure. Therefore, in our system, the better performing metric to use for our classifier is the Chi-square distance metric.

3.4 Sensitivity Analysis and Limitation of Proposed Models

3.4.1 Effect of Number of Character Prototypes on Accuracy

It is hypothesized that different letters of the alphabet require different number of prototypes to effectively model all the possible writing styles of that character. For example, there are more ways and styles to write 'f' as compared to writing 'c'. For the sake of simplicity, a preliminary level of analysis has been performed to find a global optimal value for the number of prototypes needed. This analysis will allow us to better understand the behavior of the system as the number of prototypes varies. In order to verify the results of this experiment so that it can be applicable even to other databases, a cross-validation approach was used. The 120 test document database was randomly subdivided into two equal partitions of 60 test documents each, while keeping the database of 120 reference documents constant. Hence, both partitions will be making a writer identification based on the same set of 120 reference documents. The global number of prototypes is varied from 2 to 60 for every character of that corresponding letter of the alphabet.

Figure 3-19 shows the effect of varying the number of global prototypes used with the identification rate. Similar results attained by the two subsets were notably observed. As seen from Figure 3-19, the identification rate is highest when the number of prototypes used is from 10 to 30. Additional number of prototypes beyond the 30 prototypes will result in a drop in the performance of the identification system. Similarly, there is degradation in the performance of the system when less than 10 prototypes are used. This can be explained by

the principle of Occam's razor. A large number of prototypes create sparse dimensionality which deteriorates the performance of the classification. Likewise, insufficient number of prototypes will be unable to effectively separate between inter-class variations. Based on the above analysis, the optimum number of global prototypes is taken to be 10 in the experiment.

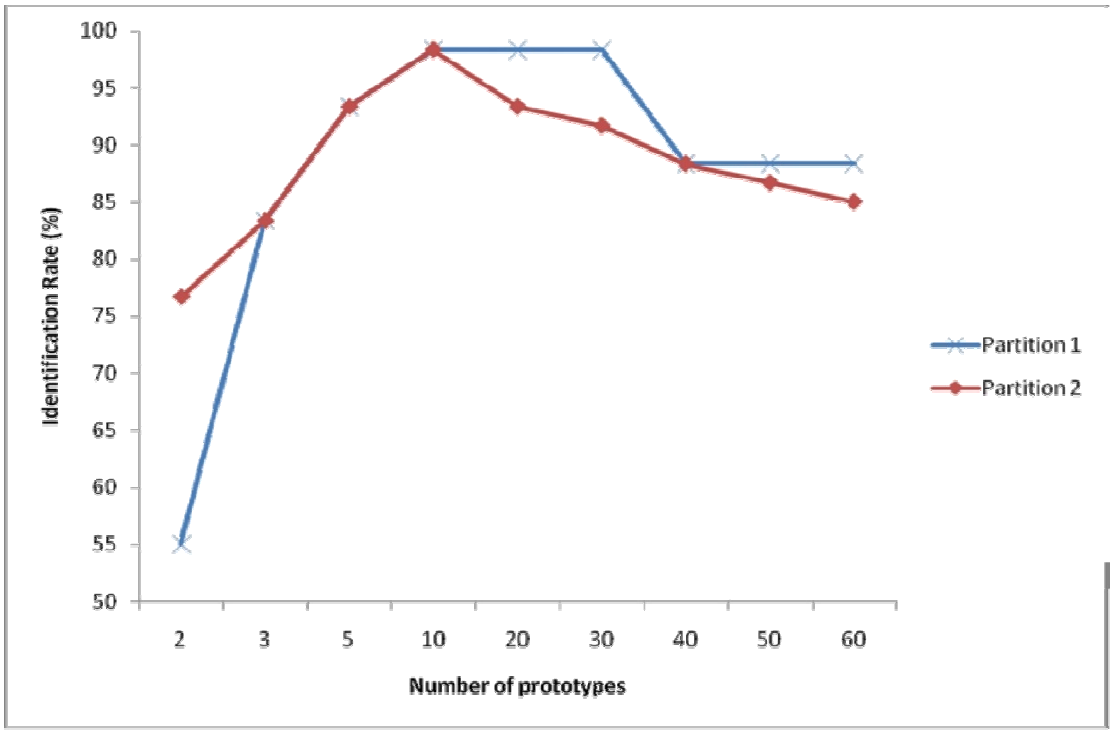


Figure 3-19: Graph of Identification Rate against Number of Clusters

3.4.2 Effect of Length of Text

Writer identification systems that adopt stochastic approaches generally require a minimum amount of data to be present in order for the stochastic modeling to be sufficiently representative of the actual data. Therefore, it is imperative to gain an understanding of the length of text that needs to be sampled so as to facilitate the derivation of useful information that closely characterizes the handwriting styles of the writers. A series of experiments have

thus been conducted to investigate the amount of text required for sufficient accuracy of the writer identification system.

The experiments have been designed as follows. Only characters in the test documents have been varied, leaving the reference documents; which varies from 168 characters for the shortest reference document to 808 characters for the longest reference document, unchanged. This is because in a typical writer identification scenario, requests to identify the test writer in question occur much more frequently than enrolment for the reference documents. Furthermore, keeping the reference documents unchanged allows a fair comparison when identifying from the same set of reference writers.

Figure 3-20 shows the average number of misclassified writers attained by varying the number of characters in the test documents. The desired number of characters for conducting the experiments is achieved by reducing the characters in the original test documents randomly across all the 26 letters of the alphabet. Writer identification is then performed on the test documents containing the reduced number of characters. In order to ensure that the results are more robust, this experiment is repeated over 10 different runs to obtain an average value. It is clearly shown in Figure 3-20 that the number of misclassified writers remains approximately constant when a minimum of 160 characters are present in each test document. However, there is a severe drop in the accuracy of the writer identification system once the number of characters in each test document falls below this threshold. This can be justified by the fact that there is insufficient allographic information available to be effectively representative of the various handwriting styles.

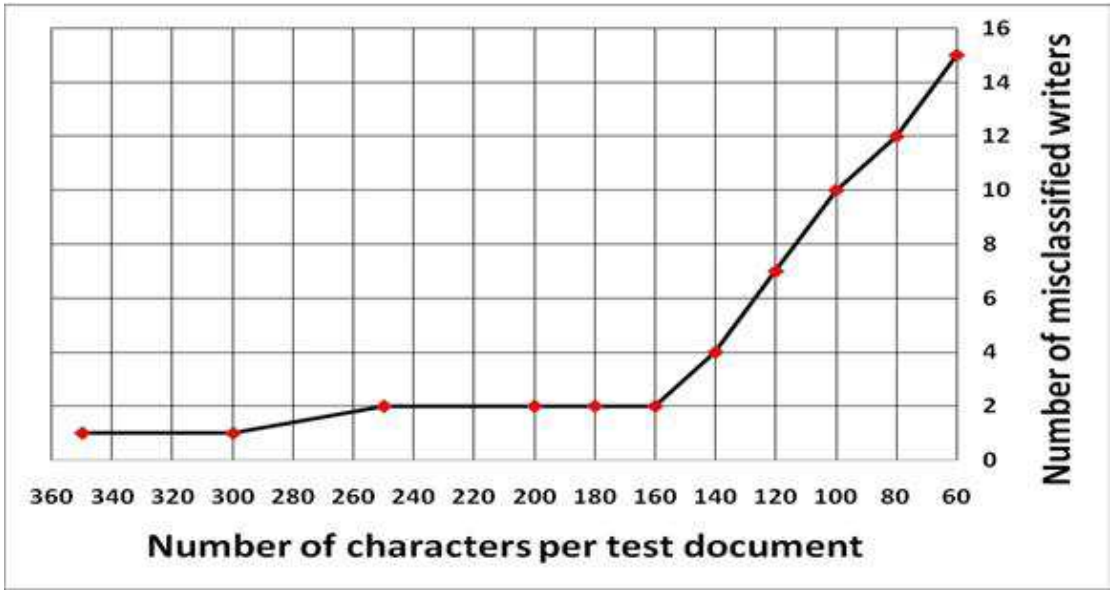


Figure 3-20: Average number of misclassified writers against number of characters

A minimum length of text is necessary to perform a reasonably accurate statistical representation of the handwriting styles. Figure 3-20 also highlights another interesting point where beyond this minimum threshold, any further increase in the amount of allographic information does not result in any performance enhancements in the accuracy of the system. In consequence, a minimum threshold of 160 characters or the approximate equivalence of 30 words or 3 lines in each test document is required for sufficient performance in our methodology.

3.5 Summary

The proposed online writer identification using Fuzzy C-Means Character Prototyping framework consisted of three main stages, the prototype training stage, the document labeling stage and the classification stage, as described in details above. The prototype training stage serves to identify common individual handwriting styles into individual prototypes at the character level. This stage is accomplished using a handwritten data derived from a large database of 373 writers. The next document labeling stage then utilizes these prototypes to

create individual distributions of handwriting styles for each of the 120 test and 120 reference documents in the test database. Based on the results of the distributions, they provide statistical information about the handwriting styles of each writer. Identification of the writer is then achieved in the last classification stage. We achieved an accuracy of 99.2% based on the 120 writers.

Furthermore, a sensitivity analysis provides us with an insight of the limitations of such an online writer identification framework. There exists a minimum length of text for a decent performance of this writer identification system because this approach is a statistical approach, and insufficient data will build a poor stochastic representation of the distribution of the handwriting styles of the writers. In this system, we have determined the minimum length of text to be about 3 lines of text. A sensitivity analysis was also conducted to investigate the effect of varying amounts of prototype to model the handwriting styles of the writers. From the experiments, this optimum number of prototypes has been ascertained to be 10.

4. Writer Identification: Alphabet level

4.1 General Overview

We begin by clearly defining the notion of alphabets and characters to avoid any ambiguities with the terminologies used here. An alphabet is *a set of characters* used in a writing system (eg. {‘a’, ‘b’, ‘c’, …, ‘z’} in the Latin writing system). A character is therefore an instance belonging to a letter of the alphabet, such as an instance of ‘a’, ‘b’ or ‘c’ in the Latin writing system.

One of the questions raised in this chapter [Siddiqi and Vincent 2010] is to determine whether it is important to assume the identity (the label) of a given instance of a letter. Of particular interest is whether alphabet knowledge can provide any concrete physical evidence in facilitating the identification of writers in writer identification systems. Multiple studies relating to the study of the discriminative power of certain letters of the Latin alphabet in the identification of writers [Srihari, Cha et al. 2002; Cha, Yoon et al. 2006; Pervouchine and Leedham 2007] have been conducted with no particularly strong conclusive evidence. Therefore, we attempt to view this problem from another perspective and use alphabet knowledge as an aiding tool to help the clustering of allograph features into prototypes instead. Our motivation lies in establishing if alphabet knowledge contains additional clues that can help ascertain the identity of the writer. In doing so, we can justify the promising results obtained by allograph prototype approaches that make use of characters to build the allograph templates.

4.2 Alphabet Knowledge Model

4.2.1 Discriminative Power of different letters of the alphabet

It is not unconceivable that different writers can write certain letters of the alphabet with a more distinct and unique handwriting style than other letters of the alphabet. This difference in handwriting styles for certain letters of the alphabet can be attributed towards multiple intrinsic and extrinsic factors. Extrinsic factors include cultural and societal influences; physical influences such as left-handedness or right-handedness, grasp and dextrality. Intrinsic variables to consider can include literacy influences and educational system influences [Huber and Headrick 1999; Morris 2000]. All these influences play a part to influence the writer in his handwriting style. It will then be a natural progression to hypothesize that different letters of the alphabet will have the capability to provide varying degrees of discrimination in identifying writers. For example, the letter ‘c’ might not have a lot of allographs and variations in its approach of writing. We thus hypothesize that such letters of the alphabet will have a low discriminative power in writer identification. In this thesis, we performed experiments to verify this hypothesis and showed evidence of the discriminative power of certain letters of the alphabet in writer identification.

In this experiment on French handwritten documents from the database containing 120 writers as previously described in section 3.3.2, only one letter of the alphabet was used at a time in writer identification. Hence, when we are investigating the discriminative power of ‘a’, only the letter of the alphabet ‘a’ was used from the reference and test documents in the document indexing stage and the retrieval stage. In this case, the size of the tf-idf vector prototypes is reduced from 26 X N = 26 X 10 = 260 to only N = 10. The top-1 accuracy in writer identification was then obtained by considering only writers that have the letter of the alphabet ‘a’ in both their reference and test documents. Writers that do not have any letter ‘a’ in either the reference or test document are omitted in the ranking results. The number of characters of each letter of the alphabet found in either the reference or test document ranges from 10 to 150 characters. This process is then repeated for 19 more letters of the alphabet.

Six letters, namely, ‘w’, ‘k’, ‘z’, ‘j’, ‘y’ and ‘h’ were omitted for the purpose of this experiment since these letters of the alphabet rarely appear in the documents and will skew the results if included. The outcome of this experiment is illustrated in Table 7.

Table 7: Discriminative power of different letters of the alphabet in writer identification on French handwritten documents

Letter	Top-1 Accuracy	% of total Char
a	43.33%	7.41%
s	43.33%	8.18%
d	42.86%	3.66%
t	39.17%	7.73%
r	35.83%	6.82%
e	35.00%	15.76%
o	33.33%	6.11%
i	31.67%	6.81%
p	30.00%	3.04%
n	27.50%	8.11%
l	23.33%	5.23%
x	20.88%	0.53%

q	19.83%	1.14%
u	19.17%	6.68%
g	18.26%	1.38%
m	16.67%	2.89%
f	14.29%	1.01%
v	13.56%	1.61%
c	12.50%	3.26%
b	11.22%	0.98%

From Table 7, the second column indicates the top-1 accuracy obtained when only that particular letter of the alphabet is used in performing writer identification. The top-1 accuracy is the percentage of writers that are ranked correctly in the top-1 position; hence a top-1 accuracy of 100% will indicate that all the writers have been identified correctly. The results supported our hypothesis that certain letters of the alphabet like ‘v’ , ‘c’ and ‘b’ might have few variations in its allographs and style of writing and hence will have a low discriminative power in writer identification. Likewise, letters of the alphabet like ‘a’, ‘s’, ‘d’ and ‘t’ are highly discriminatively in writer identification. As such, more emphasis should be placed on such letters with high discriminative powers and less emphasis on those with low discriminative powers.

The third column in Table 7 shows the frequency of occurrence of such letters of the alphabet in both the test and reference documents. This distribution of frequency of

occurrence is similar to the results obtained by Rosenbaum et al. [Lyne 1985; Rosenbaum and Fleischmann 2003] for characterizing the distribution of letters in general French linguistic resources, where the most frequent letter of the alphabet is ‘e’ and the least frequent letters are ‘w’ and ‘k’. Therefore, the set of documents in our reference and test database is representative of general French linguistic resources.

A further examination of the third column in Table 7 reveals an interesting note, where certain frequently appearing letters of the alphabet such as ‘e’ do not provide high discriminative powers, whereas rarely occurring letters such as ‘x’ is able to be more discriminating in writer identification. This indicates that characters need not appear very often in order to have high discriminative powers. From this result, we can design algorithms that give more focus and emphasis to letters of the alphabet that have high discriminative powers, regardless of whether they occur frequently in the document. Therefore, our results strongly indicate that different letters of the alphabet do indeed have different capabilities to provide varying degrees of discrimination in identifying writers.

Our initial observation is that some letters of the alphabet such as ‘f’ are ranked in a rather low position for its discriminative power even though we expect a high discriminative power for it. These expectations initially arose because intuitively, the presence of both the descender and the ascender in ‘f’ should naturally allow for more handwriting variations [Han and Kamber 2006]. However, it is interesting to note that our reported results are counter-intuitive to this expectation. The reason could be attributed to our observation that this letter of the alphabet suffers from a very low number of instances of characters: only around 1%. This leads to a poor estimation of the prototype distribution, resulting in the same problem as the experiment reported in Figure 3-20 in the previous chapter, i.e. when the number of characters is low the performance drops severely. Nonetheless, this experiment

clearly shows that different letters of the alphabet have different identification capabilities, which supports findings from Cha et al. [Srihari, Cha et al. 2002; Cha, Yoon et al. 2006]. More emphasis should be placed on such letters of the alphabet with high discriminative powers and less emphasis on those with low discriminative powers.

In order to further explore this notion of the relationship between frequency of occurrence and discriminative powers, we further note that frequently appearing letters of the alphabet such as ‘u’ (19.17% accuracy with a character frequency of 6.68%) do not provide high discriminative powers, whereas letters of the alphabet that do not appear as frequently such as ‘d’ (42.86% accuracy with a character frequency of 3.66%) is more discriminating in writer identification. This implies that the frequency of occurrence is not directly correlated to the discriminative power of different letters of the alphabet. In our case, our *tf-idf* is not related to the frequency of occurrence of the letters but to the frequency of the prototypes being used, since our tf-idf has been normalized on a letter of the alphabet basis. Hence each letter of the alphabet is able to be processed independently.

Interestingly, our results contrast with works by Niels et al. [Niels, Gootjen et al. 2008] where they conclude, without experimentation, that frequently occurring letters of the alphabet like ‘e’ are least suitable for distinguishing between writers and rarely occurring letters of the alphabet like ‘q’ are most suitable for distinguishing between writers. Based on our results, we argue that the suitability of different letters of the alphabet for writer identification should not be based solely on the frequency of occurrence, but rather, take into account various other intrinsic factors and extrinsic factors which can affect the discriminating power of different letters of the alphabet as well.

4.2.2 Proposed Methodology in using Alphabet Knowledge

In section 4.2.1, we established that different letters of the alphabet do have different discriminative power. Hence, we propose to use this information of alphabet knowledge to aid in the clustering of the allographs in order to build the distribution of prototypes that are representative of the handwriting styles of writers. An Alphabet Information Coefficient is also introduced during classification to take into consideration the effects of different letters of the Latin alphabet. Two writer identification systems are built; one that uses alphabet knowledge and one where alphabet knowledge is omitted. The writer identification system without using alphabet knowledge allows for comparison with the system that uses alphabetic information. The approaches for both systems are described and compared in the sections that follow.

Methodology without alphabet knowledge

The purpose of this methodology without alphabet knowledge is to investigate the effect of whether the absence of a-piori information of the alphabet matters in the writing identification process. To investigate this, the same framework where the three stages (cf. prototype training stage, document labelling stage and classification stage) previously described in Figure 3-3 are similarly used. However, in the document labelling stage, the segmented characters are instead clustered into 260 prototypes all at the same time in the feature space, without any alphabetic information. Likewise, in the last classification stage, frequency vectors are then classified to identify the writers using a variation of the chi-square distance measure, as depicted in equation (4-1).

$$\text{Distance}\left(\text{writer}_i, \text{writer}_j\right) = \sum_{k=1}^{N} \frac{idf_k \left(tf_{k,i} - tf_{k,j}\right)^2}{tf_{k,i} + tf_{k,j}} \qquad (4\text{-}1)$$

where idf_k is the inverse document frequency and tf_{ki} represents the term frequency of the reference and test documents, where k refers to the k^{th} prototype and i refers to writer i. N stands for the number of prototypes that are being clustered, where N=260 in this case. This differs from equation (3-12) because here in equation (4-1), we are not making any assumptions or using any a-priori information about the alphabet knowledge in clustering. Whereas for equation (3-12), we have assumed a-priori information about the alphabet knowledge during the clustering process when we cluster them on a letter of the alphabet basis, hence the value of N for equation (3-12) is to cluster across 10 prototypes instead. In equation (4-1), we have again adapted the standard *idf* and *tf* measures from traditional document analysis literature [Salton, Wong et al. 1975; Salton and Buckley 1988] to address our writer identification problem. In our framework, idf_k and $tf_{k;i}$, are used to create a statistical distribution that models the handwriting styles of different writers. This distribution then enables us to classify and perform an identification of the writer in question.

Methodology using alphabet knowledge

The purpose of this methodology with alphabet knowledge is to investigate the effect of whether the presence of a-piori information of the alphabet matters in the writing identification process. To investigate this, the same framework discussed in chapter 3 is used. The main difference lies in that in the last classification stage, the Alphabet Information Coefficient is introduced during this stage to specifically weigh each letter of the Latin alphabet as shown in equation (4-2). The reason is that with this second method, the distributions over the prototypes are computed on a letter-by-letter basis. This contrasts with

the methodology without alphabet knowledge in the earlier section where in the document labelling and classification stages, no a-priori information on the alphabet is used. Consequently, 26 distributions are obtained instead of a single one as with the first method. Some of these distributions are more reliable than others and it is the goal of the Alphabet Information Coefficient to give more weight to the more reliable distributions.

$$\text{Distance}\left(\text{writer}_i, \text{writer}_j\right) = \frac{\sum_{\alpha='a'}^{'z'}\left(AIC_\alpha \times \sum_{k=1}^{N} \frac{idf_k\left(tf_{k,i} - tf_{k,j}\right)^2}{tf_{k,i} + tf_{k,j}}\right)}{\sum_{\alpha='a'}^{'z'} AIC_\alpha} \tag{4-2}$$

$$Stepwise\ AIC = \begin{cases} 0, & if\ C_{\alpha_i} \times C_{\alpha_j} = 0 \\ 1, & otherwise \end{cases} \tag{4-3}$$

$$Geometric\ mean\ AIC = \sqrt{C_{\alpha_i} \times C_{\alpha_j}} \tag{4-4}$$

$$Sigmoid\ \ AIC = \frac{1 - \exp(-\lambda \times C_{\alpha_i} \times C_{\alpha_j})}{1 + \exp(-\lambda \times C_{\alpha_i} \times C_{\alpha_j})} \tag{4-5}$$

In equation (4-2), idf_k and $tf_{k;i}$ have the same connotations as equation (4-1), N is the number of prototypes that are being clustered, per letter of the Latin alphabet and is chosen to be N=10 as explained previously. The Alphabet Information Coefficient is being calculated for all letters α of the Latin alphabet, $\alpha \in \{'a','b',...,'z'\}$, which we have designed three different expressions for modelling this Alphabet Information Coefficient function. These expressions are defined by equations (4-3), (4-4) and (4-5), where $C_{\alpha i}$ represents the number of characters or instances of the letter α of the Latin alphabet that appears in the document of reference writer *i*. Likewise, $C_{\alpha j}$ represents the total number of characters for test writer *j* and in equation (4-5), λ is a tuning parameter which is experimentally set to be 0.01. The design

of these expressions for modelling the Alphabet Information Coefficient function will be discussed more in details in sections 4.3.2

The distance defined by equation (4-2) is derived from a chi-squared metric, where it is used to match the two distributions: one from a test writer, another from a reference writer. It introduces a relative comparison between the components of the distributions, instead of focusing on the absolute difference between them like with a Euclidian distance. This is more meaningful with respect to the different handwriting styles that we are attempting to make a distinction of. It has also given the best results in the study conducted by Schomaker [36].

Here, we have introduced the concept of the Alphabet Information Coefficient in equation (4-2). The Alphabet Information Coefficient is a measure of reliability of the character prototypes being built from the distribution of the letters of the Latin alphabet available in the documents. As the distributions are estimated from the counting of the number of instances of each letter, it is evidently clear that the reliability of these distributions are dependent on the number of samples encountered. The higher the number of samples of a letter of the Latin alphabet that are present in a text, the more reliable the distribution for this letter should be. Conversely, with very few samples, eventually none that exist for this letter of the Latin alphabet in either the test document or the reference document of a given letter of the alphabet, it becomes undesirable that the corresponding distribution has an important weight in the global distance function and should eventually be totally discarded. Hence, the Alphabet Information Coefficient has to take into account the number of characters that are present in both the reference and test document during classification.

4.2.3 Experiment

Online handwritten documents were collected from 200 writers where the writers copied text passages from the Reuters financial news corpus, Reuters-21578 [Saldarriaga, Viard-Gaudin et al. 2009]. Figure 4-1 illustrates a typical online handwritten sample that was collected. Reuters-21578 is a very popular publicly available financial news corpus initially intended for text categorization research and it contains a collection of different categories of financial newswire from Reuters. Each writer was requested to copy two different text passages, with a digital pen and paper technology: one which is considered as a reference document and the second one is taken as a test document from randomly selected news categories of Reuters-21578.

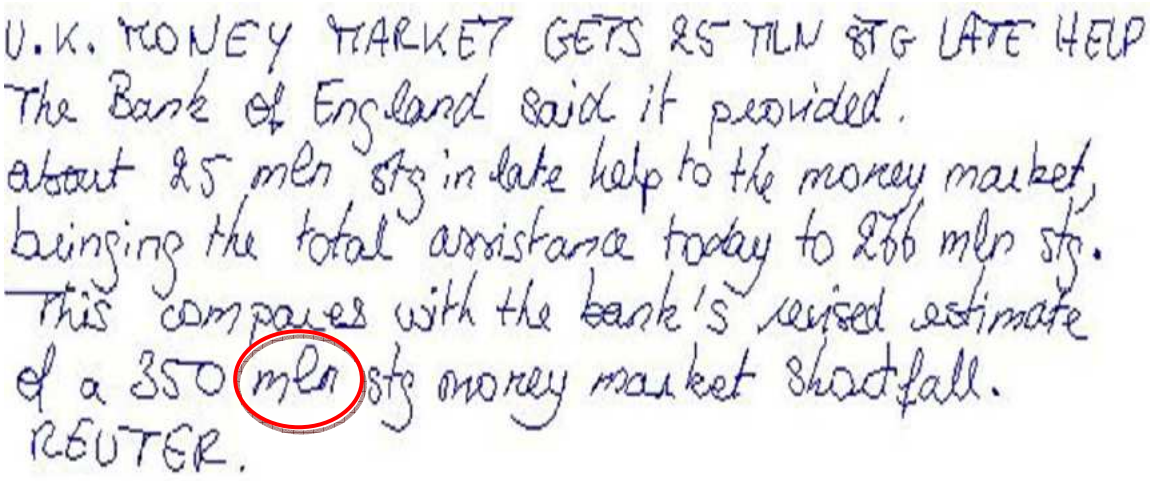


Figure 4-1: A sample text passage from theReuters-21578 database (financial news category).

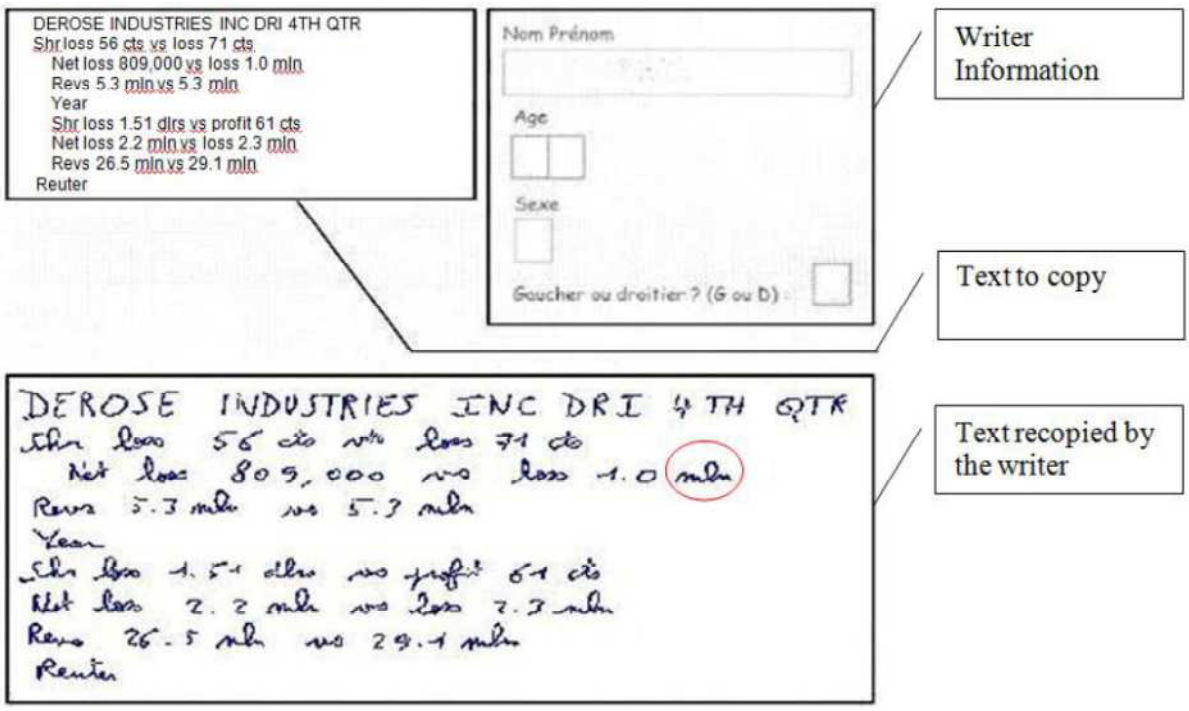


Figure 4-2: A sample of the form used to collect the handwritten sample from the writers.

Figure 4-2 shows the form used to collect the samples from the writers. These reference and test documents that were collected belong to a separate dataset from the IRONOFF database, which is only used for building the prototype set of characters. The IRONOFF database was chosen for this task of building the character prototypes because it contains a wide selection of handwriting styles from 373 different writers that is consequently large enough to be representative of the different variations in handwriting styles. As a result, the character prototype set is generic and independent with respect to the actual reference database of documents from which the writer is to be identified from. Furthermore, a separate database needed to be collected because the IRONOFF database contains only isolated words and hence are not representative of actual online documents. As can be seen from Figure 4-1and Figure 4-2, each online document on average consists of only 3 to 10 lines of text and contains shorthand that exists in financial news jargons. For example, mln (circled in Figure

4-1) which is a typical short-hand to represent the word ‘million’, is commonly used in financial terms. As illustrated in Figure 4-3, the automatic recognizer tool might confuse the word `mln' for `men', since it is an out-of-lexicon word. Next, Figure 4-4 shows the wrongly recognized character `e' that will be utilized later during writer identification. Such character recognition errors are not removed before trying to identify the writers so that a realistic and practical scenario can be reflected since such errors are inevitable in reality.

A character recognition rate of 89% was achieved on this database using the MyScript [MyScript 2012] industrial text recognition engine. This implies that, when using alphabet knowledge as defined in section 4.2.2 to help in the assignment to the corresponding prototype set, 11% of the characters will be assigned wrongly. For instance, in the previous example that illustrated the incorrect recognition of the word `mln', a `l' which is wrongly recognized as an `e' will be erroneously clustered to the set of prototypes for letter `e' of the Latin alphabet. Hence, the prototype frequency estimation will be corrupted by noise because of the presence of such inevitable recognition errors.

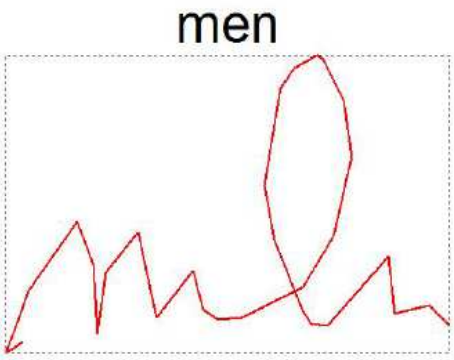


Figure 4-3: The word 'mln' that has been wrongly recognized

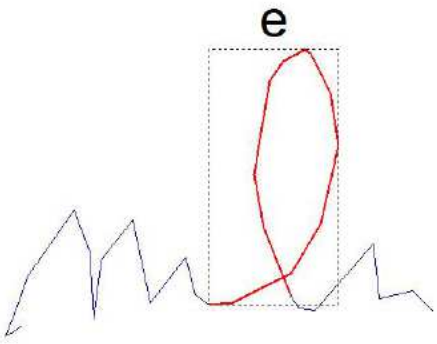


Figure 4-4: The 'e' in the wrongly recognized word that will be used for writer identification

Nonetheless, the ensuing choice of the recognition engine has no specific link or any bearings on the proposed method. In our experiments, “MyScript” was chosen because it is an off-the-shelf product that contains convenient built-in resources, while offering the possibility to create specific linguistic resources. Any other recognition system can equally be used to implement the proposed method. The better the segmentation and recognition capabilities of the recognition engine is at recognizing the letter, the higher the writer identification rate will be. Hence, the proposed method has no reliance or dependences on any one specific text recognition engine and other recognition engines can be used in tandem if a writing style is very odd and not well supported by any one particular recognition engine.

4.2.4 Alphabet Knowledge Model

Table 8: Effect of alphabet knowledge on writer identification rate

Top-1 Writer Identification Rate of 200 reference writers (Reuters database)				
Clustering without Alphabet Knowledge	Clustering Using Alphabet Knowledge			
66.0%	No Alphabet Information Coefficient used	Using stepwise Alphabet Information Coefficient shown in eq. (4-3)	Using geometric Alphabet Information Coefficient shown in eq. (4-4)	Using sigmoid Alphabet Information Coefficient shown in eq. (4-5)
	73.5%	76.5%	86.5%	87.0%

Table 8 compares the top-1 writer identification rate between the approach without using any alphabet knowledge to help in clustering and the approach with using alphabet knowledge to help in clustering. When alphabet knowledge was omitted in helping to cluster, an accuracy of 66.0% was achieved (68 misclassified writers out of 200). In contrast, with the help of alphabet knowledge to cluster the distribution of prototypes at the character level, we see that the alphabetic approach resulted in a higher top-1 identification of 73.5%. This translates into a misclassification error of only 53 out of 200 writers in the top-1 position. Alphabet knowledge provides additional knowledge on the handwriting styles. There exist different morphological variations even among different letters of the Latin alphabet. For

example, the letter `f' has more morphological variations and styles to write it as compared to the letter `c', where only a limited number of variations exist. Therefore, it will help in the identification of writers if we are able to examine certain letters more closely. With the assistance of this alphabet knowledge, clustering the graphemes or allographs into prototypes to model the handwriting styles of writers will be able to provide a higher writer identification rate, as evidenced from our results.

This improvement of the writer identification rate serves as further evidence to suggest that alphabet knowledge do influence the recognition of writers. This can be explained by the fact that in our proposed methodology, we make use of alphabet knowledge to cluster the prototypes into clusters of `a' to `z'. It must be highlighted that even though the identification rate reported here are much lower compared to the work done in the previous chapter 3 by 100 writers on a set of French handwritten documents, as well as works by Chan et al. [Chan, Viard-Gaudin et al. 2008] and Niels et al. [Niels, Gootjen et al. 2008], the reader should bear in mind that the Reuters database of 200 writers is much more challenging to recognize due to the shorter length of text in the documents and a higher segmentation error, as previously discussed in section 4.2.3. The effect of the transition from no Alphabet Information Coefficient being used (73.5%) to using Alphabet Information Coefficient (76.5% - 87%) will be discussed next in section 4.3

4.3 Alphabet Information Coefficient

4.3.1 Effect of Alphabet Information Coefficient

It is observed that certain letters of the Latin alphabet such as `j', `q' or `z' have infrequent occurrences for the English language since they rarely appear in most English texts. This might cause distortions to the prototype distributions created if the instances of such letters are not sufficiently consistent to completely model the writing style of that letter of the Latin alphabet. The Alphabet Information Coefficient serves to eliminate this bias by attributing less importance during classification to those letters which rarely occur in either the reference or test documents. This effectively places more emphasis to those letters that are able to sufficiently represent the writer's style of writing. Hence, the Alphabet Information Coefficient is a measure of reliability where more importance is attributed to prototypes that can reliably represent the handwriting styles of that writer in order for a higher writer identification rate to be attained.

When no Alphabet Information Coefficient is used in equation (4-2), all letters are treated as equally reliable in representing the writer's handwriting style. This is equivalent to assigning a same weight of 1 to the Alphabet Information Coefficient for all letters of the Latin alphabet. However for the stepwise Alphabet Information Coefficient shown in equation (4-3), letters which do not appear in either the reference or test documents are deemed as unreliable and thus omitted. As seen from Table 8, the stepwise Alphabet Information Coefficient improves the top-1 writer identification rate from 73.5% to 76.5% when the stepwise Alphabet Information Coefficient is used. This set of results highlights the fact that the reliability of the character prototypes that we use to model the handwriting styles are in fact dependent on the distribution of the letters available in the documents. When no instances of the letter appear in either the test or reference document, then that character

prototype should not be used in the determination of the writer identity. This is consistent with the results obtained where an increase in accuracy is attained when the stepwise Alphabet Information Coefficient is used as compared to when no Alphabet Information Coefficient is used.

However, the failure of the stepwise Alphabet Information Coefficient becomes apparent when letters of the Latin alphabet containing very few samples are used to model the handwriting style. In such cases, it becomes detrimental to the identification rate if such distributions still have an important weight in the global distance function. This is also why we have designed other Alphabet Information Coefficient functions to take into consideration the reliability of the prototypes under such circumstances. A good choice of Alphabet Information Coefficient should give higher reliability to letters of the Latin alphabet that occur frequently and reduce the reliability to rarely occurring letters which are not sufficient enough to effectively model the handwriting style. As seen in Table 8, a good choice of Alphabet Information Coefficient will result in the further increase in the accuracy from 76.5% to 87%. The design of the Alphabet Information Coefficient will be discussed next in section 4.3.2

4.3.2 Design of the Alphabet Information Coefficient

A series of experiments have been conducted to further investigate the design of different Alphabet Information Coefficients. Equation (4-3) uses a step-wise Alphabet Information Coefficient function so that it prevents using a letter of the Latin alphabet when no instances of this letter are encountered in either the test or reference document. All other letters are given the same weight. In this case, the results attained 76.5% for the top-1 identification rate. When equation (4-4) is considered, the geometric mean of the sample

numbers is used as a weighting factor, thereby preserving the previous property of discarding letters when no instances of the letter of the Latin alphabet are observed while increasing the weight for the most representative letters. A significant improvement in the top-1 identification is observed where an accuracy of 86.5% is attained when equation (4-4) is used. Equation (4-5) uses a sigmoid function, henceforth to be termed sigmoidal Alphabet Information Coefficient, to saturate the influence of the number of occurrences of a given letter. It is in our interest to limit those letters which are extremely frequent such as `a' and `e' from being overemphasized by the Alphabet Information Coefficient during the classification.

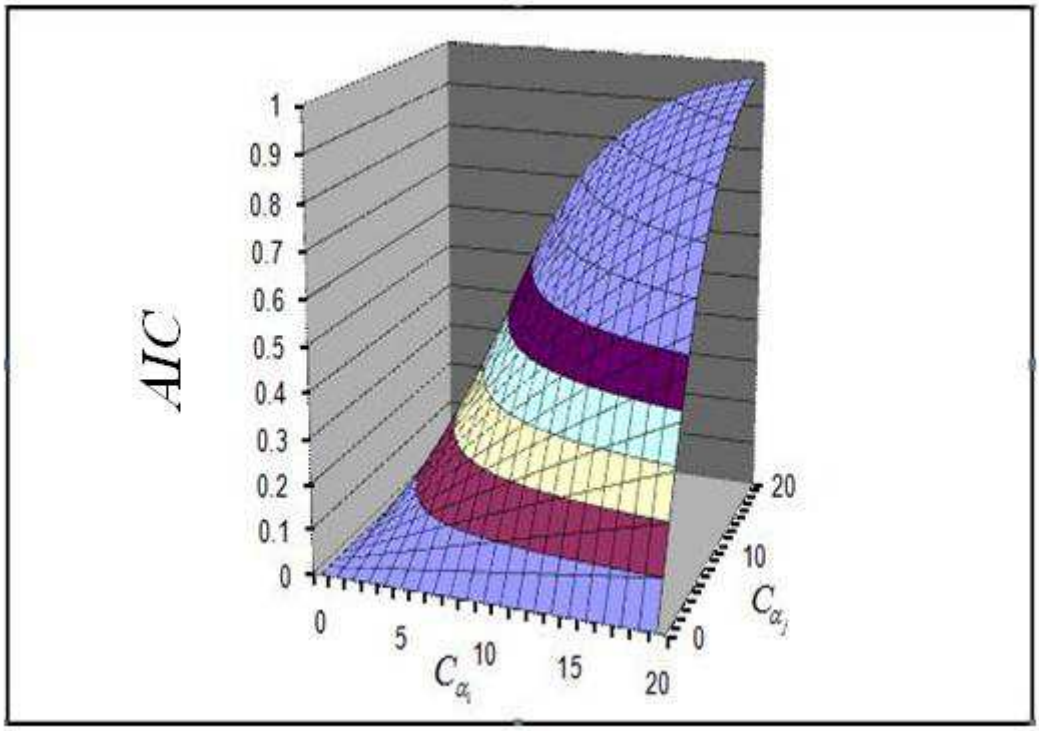


Figure 4-5: The sigmoid function in the $C_{\alpha_i} \times C_{\alpha_j}$ space

The tuning parameter λ, when experimentally set to be 0.01 results in the sigmoid function that is illustrated in Figure 4-5. With such a value of λ, the Alphabet Information Coefficient quickly becomes saturated to a value of 1 when approximately 20 occurrences of the same letter have been encountered in each document. This effectively means that we do

not require more than 20 occurrences of a letter from each document in order to reliably model the allograph distribution for that letter. In addition, we can see from Table 8 that the usage of this sigmoidal Alphabet Information Coefficient achieved a top-1 accuracy of 87.0%. This is a slight improvement compared to the Alphabet Information Coefficient using the geometric mean where a top-1 accuracy of 86.5% was attained. Upon closer examination, we observed that an identification rate of 98% can be attained with the sigmoidal Alphabet Information Coefficient if we relax our constraints and consider the top-9 positions, as shown in figure 7 (ie. The writers are considered correctly identified if they can be ranked within the top-9positions). However, for the Alphabet Information Coefficient described by equation (4-4), the same identification rate of 98% can only be attained if we consider the top-19 positions. Therefore, our experiments show that the sigmoidal Alphabet Information Coefficient can effectively improve the performance of the writer recognition system.

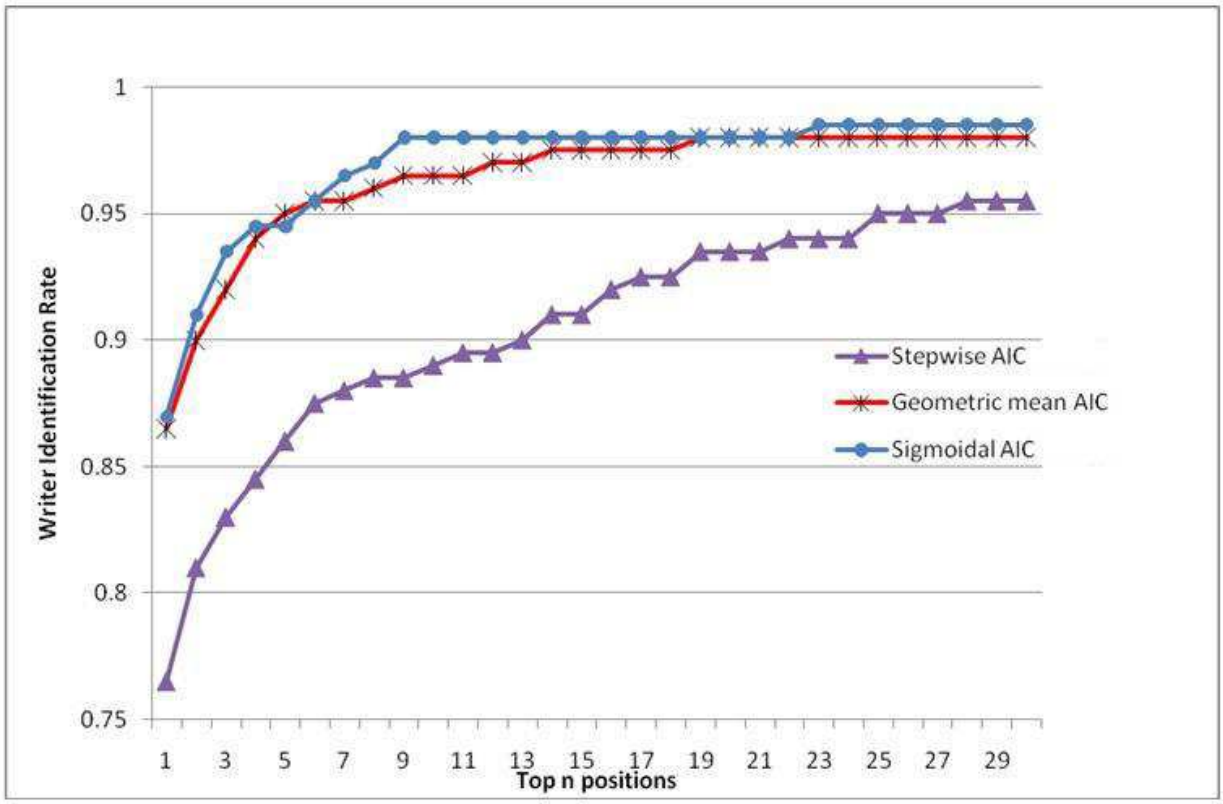


Figure 4-6: Performance of the writer recognition system on a database of 200 writers using different Alphabet Information Coefficient functions.

The choice of λ affects how quickly saturation of the sigmoid function is reached. As λ approaches a very large value, it behaves similarly to a step function as described in equation (4-3). The physical interpretation of such a step function Alphabet Information Coefficient is that we exclude those cases when no instances of such letters exist in either the reference or test documents. Likewise, we attribute equally all other letters of the Latin alphabet that contain some instances of the letters in both the reference and test documents. From Table 8, we see that the step function Alphabet Information Coefficient performs only slightly better than the case when no Alphabet Information Coefficient is used; achieving an identification rate of only 76.5% (47 misclassified writers) compared to an identification rate of 73.5% when no Alphabet Information Coefficient is used. This result is to be expected because those letters where only 1 or2 instances appear in the documents should be given less

importance as they might not be sufficiently consistent to represent the prototype distribution. Hence, the value of λ should not be too large. Conversely, if the value of λ is too small and approaches 0, this will have the same effect as when Alphabet Information Coefficient is not used. Based on our experiments, we have found the optimum value of λ to be 0.01.

4.4 Discussion

4.4.1 Discriminative power of different languages in the Latin alphabet system

Certain letters of the Latin alphabet allow for more variations to be written compared with other letters, thereby allowing different writers to express their individuality of handwriting with a style that is more distinctive and differentiated. For example, the letters of the Latin alphabet `f' has more morphological variations and styles in its approach of writing compared to the letter `c', where only a limited number of variations exist. This implies that most writers might inadvertently write the letter `c' with a similar style. Therefore, we hypothesize that different letters of the Latin alphabet will have different capabilities in identifying writers and we refer to this term as the discriminative power of letters.

Experiments were conducted to verify this hypothesis by using only one letter at a time to identify the writers. For example, in order to investigate the discriminative power of the letter `a', only the letter `a' was used from the reference and test documents in the document indexing stage and the retrieval stage. The top-1 accuracy in writer identification was then obtained by considering only writers that have the letters of the Latin alphabet `a' in both their reference and test documents. Writers that do not have any letter `a' in either the reference or test document are omitted in the ranking results. This process is then repeated for

the other letters of the Latin alphabet. Four letters, namely, `z', `q', `x' and `j' were omitted for the purpose of this experiment since these letters rarely appear in the documents and will skew the results if included. The outcome of this experiment is illustrated in Table 9.

Table 9: Discriminative power of different letters of the alphabet in writer identification on English handwritten documents (REUTERS-21578 database)

Letter	Top-1 Accuracy	% of total characters in documents
d	22.45%	4.45%
a	20.81%	8.17%
n	17.00%	7.41%
r	17.00%	7.21%
o	16.50%	7.89%
s	14.50%	7.64%
t	14.00%	8.67%
g	13.66%	1.57%
k	11.88%	0.73%
y	11.76%	1.47%
e	11.00%	12.17%
i	11.00%	7.55%
p	9.33%	2.31%
h	8.63%	3.39%
f	7.94%	2.21%

b	7.79%	1.37%
l	7.54%	4.20%
m	6.45%	2.44%
u	6.03%	3.16%
v	5.81%	0.95%
w	5.44%	1.16%
c	4.64%	3.25%

The results previously reported in Table 7 on French documents are similar to the results obtained here as shown in Table 9 for English documents; with the notable exceptions that `s' and `t' are more discriminating in the French documents (top-2 and top-4 respectively as described earlier in section 4.2.1) and that `b' was the least discriminating. These slight differences are due to the fact that English and French documents are still inherently different. Nonetheless, this experiment clearly shows that different letters of the Latin alphabet have different identification capabilities, which also supports findings from Cha et al. [Srihari, Cha et al. 2002; Cha, Yoon et al. 2006]. Therefore, more emphasis should be placed on such alphabets with high discriminative powers and less emphasis on those with low discriminative powers. The third column of table 3 shows the frequency of occurrence of such letters in both the test and reference documents. This distribution of alphabet frequency is similar to the results obtained by Foster [Foster 1982] based on the Brown Corpus of US English words, where the most frequent letters are `e', `t', `a' and the least frequent letters are `q' and `z'.

In many pattern recognition scenarios, one of the critical tasks in any large-scale practical system involves reducing the computational complexity of the system such as

seeking to minimize features that are redundant. The method proposed involves clustering in a feature space of dimensionality 210 for each letter, compounded by the fact that this has to be repeated for all 26 letters of the Latin alphabet. This issue can be addressed by using only a subset of Latin alphabet instead. We demonstrate the feasibility of using the discriminative power of the letter to determine a subset.

Figure 8 shows the drop in performance as letters of the Latin alphabet are removed based on the order of their discriminative power. The experimental results show that the performance remains constant as the least discriminating letters are removed (`c', `w', `v', etc). Conversely, we suffer a drastic drop in the performance as the most discriminating letters are removed (`d', `a', `n', etc). Degradation in the performance is already observed even with the removal of one discriminating letter. Our results indicate that a choice of letters based merely on their discriminative power can be utilized to select a subset of alphabets without adverse impact to the performance of the writer identification system. This will help to reduce the dimensionality and decrease the computational complexity of the system. As evidenced from our results, this method of using the discriminative power to select a subset, albeit sub-optimal, is a much simpler and effective approach compared to other more complex and time-consuming sub-set selection algorithms.

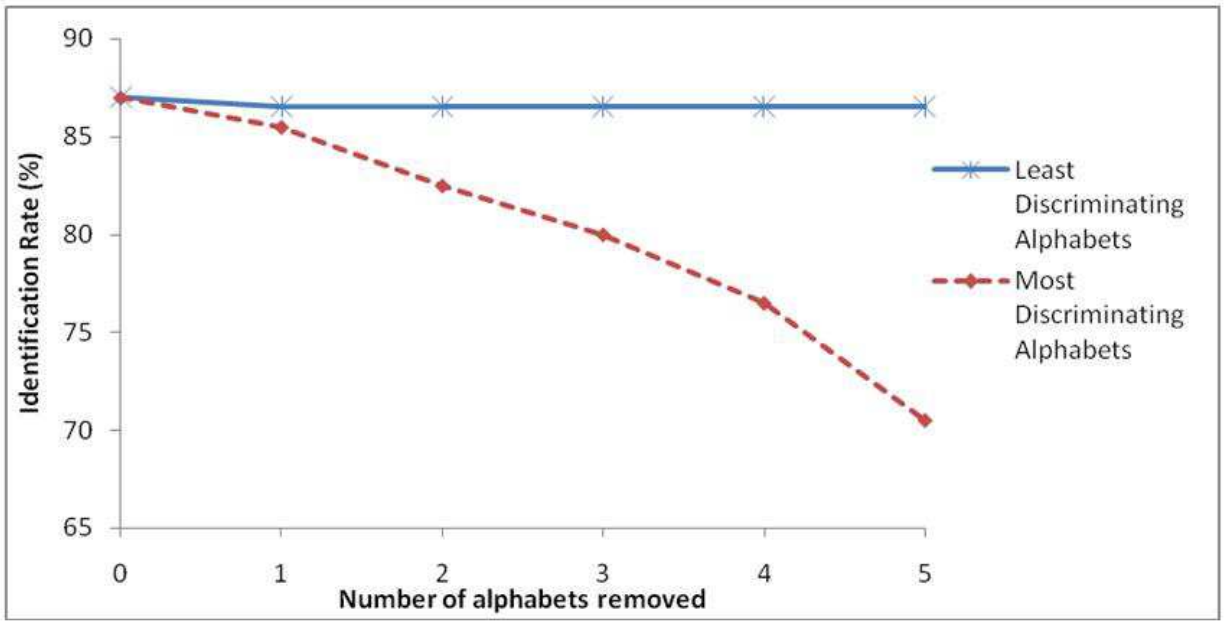


Figure 4-7: Identification rate as letters are removed based on their discriminative power

4.4.2 Discussion of Alphabet Information

Alphabet knowledge contains valuable information pertaining to the identity of the writer. This is demonstrated from the experimental results of an improvement in the writer identification rate from 66.0% to73.5% based on a set of 200 reference writers using the Reuters-21578 corpus when alphabet knowledge is considered. Furthermore, with the introduction of the alphabet information coefficient Alphabet Information Coefficient, we are able to achieve higher accuracies of 87.0%. A study into the Alphabet Information Coefficient design reveals that it is the sigmoidal variation of the Alphabet Information Coefficient which enables to give us this slightly higher top-1 identification rate of 87.0% when compared with the 86.5% obtained with the geometric mean version. Upon closer examination, our results reveal that a 98% identification can be attained at the top-9 positions with the sigmoidal Alphabet Information Coefficient as compared to the top-19 positions

with the Alphabet Information Coefficient in equation (4-4). The optimum choice of λ was also found to be 0.01 from our experiments.

A large value of λ is similar to a stepwise Alphabet Information Coefficient and our results show that the stepwise Alphabet Information Coefficient performs poorly with a top-1 identification rate of only 76.5%. This poor result is expected because cases where only 1 or 2 instances of the letter in the document are not sufficiently consistent to properly represent the prototype distribution. On the same note, a small value for λ will result in a poor identification rate as well. In fact, as λ approaches 0, it has the same effect as when Alphabet Information Coefficient is not used.

4.5 Summary

The individuality of alphabet knowledge has been demonstrated from our experimental results, where building prototype templates at the character level retains significant information relevant to the writer. Since alphabet knowledge exists inherently when the prototype templates are built at the character level, this verifies our hypothesis that alphabet knowledge contains individuality about the writer information. Therefore, alphabet knowledge helps in the identification of writers by allowing certain letters of the Latin alphabet that are more relevant to certain writers to be closely examined. For this reason, the results presented in this thesis explain why works that made use of character prototype approaches for writer identification are able to attain promising results. Hence, we foresee advantages in exploiting this individuality of alphabet knowledge, especially in automatic writer identification systems.

In this chapter, we have also established the notion of the discriminative power of different letters of the Latin alphabet through two different languages; the English language and the French language, as well as ascertain the feasibility in utilizing such information towards reducing the dimensionality and complexity of the writer identification. Furthermore, similarities and differences between the discriminative powers of both languages are compared. This thesis has established that the discriminative power of different letters is indeed language-dependent, and it is certainly dependent on the grammatical structure of the language being used.

A writer identification system can, through the effective use of only the more discriminative letters by the systematic removal of the less discriminative letters, result in a writer identification system where the accuracy of the writer identification system are still stable even if we do not use the complete set of letters of the Latin alphabet writing system. This result is important in creating a writer identification system that is both accurate and computationally efficient.

5. Conclusions

5.1 Framework for writer identification on Alphabet-based writing systems

The objective of this thesis was set out to deal with the problem of writer identification on online handwritten documents. This objective was met through the original contribution of this thesis in which a framework was proposed for writer identification based on a character prototyping approach, at the character level. Working at the character level allowed us to have more intuitive insights into the writing styles of various handwritten documents as compared to graphemes. The design of the fuzzy C-means approach was also addressed to further improve the performance of such a writer identification system.

In the first category, we propose the framework for a character-prototype approach, and achieved an accuracy of 99.2% based on the 120 writers. This was accomplished with an adapted Chi-square distance measure using an exponential fuzzy C-means kernel in the proposed writer identification framework. Compared with conventional grapheme–based prototype approaches, the consistency and superiority of our proposed approach has been experimentally demonstrated. Our proposed methodology does require a minimum amount of approximately 3 lines of text in order for a stochastically robust representation of the distribution of the writing styles.

5.2 Establishment of a language-specific information criterion - Alphabet Information Coefficient

In the second category, we further introduce two concepts of discriminative power of different letters of the alphabet, and the notion of an Alphabet Information Coefficient term. To this end, this thesis set a foundation for the concept of an “Alphabet Information Coefficient” in Alphabet-based writing systems, which was introduced as a measure of

reliability of how robust and consistent the prototypes are in representing the handwriting styles of that writer. Through the design of a sigmoidal Alphabet Information Coefficient term, we are able to achieve a good accuracy of 87% on a challenging dataset of short Reuters financial news written by 200 writers.

The individuality of alphabet knowledge has been demonstrated from our experimental results, where we have established that a writer identification system need not make use of the full set of letters of the alphabet and still attain a stable accuracy without a significant degradation or loss in accuracy. This is remarkable and useful in reducing the computational complexity of the system.

As a final concluding remark, it is interesting to note that many of today's modern languages (such as English, French, German, Italian, Cyrillic, Greek, Portuguese etc) with Latin-based writing systems attribute for more than 45% of the world's population, which is the largest percentage for any writing system, as illustrated earlier in Figure 2-2 [Ethnologue 2012]. It would therefore be worthwhile to study and develop a generic writer identification framework that can be effective across different languages that make use of the alphabet writing system.

As such, the framework discussed in this thesis can be easily extended to not just the English language, but the other languages with an alphabet writing system as well since such writing systems can be decomposed at the character level. Therefore, further studies still needs to be carried out on the how a character prototyping approach react particularly to the intricacies of that language group. This is because different languages tend to have different frequent, as well as infrequent characters, and the knowledge of such information can become useful in designing an Alphabet Information Coefficient suited for that language group.

Consequently, certain languages such as Brazilian Portuguese only employ certain letters like ‘k’, ‘w’ and ‘y’ in Brazil only for personal names. Such information will prove useful as a basis in designing a language-dependent writer identification system for forensic applications.

5.3 Future Directions for Writer Identification: Character level

The current proposed online writer identification using Fuzzy C-Means Character Prototyping framework relies on the premise that the handwritten document can be decomposed into the character level. This approach of writer identification can thus be easily imagined to be extendable to other languages of the Latin alphabet system, such as Italian, Portuguese or German. However, this extension is not trivial and not inconsequential because of certain peculiarities of different languages. Take for example in Portuguese, letters such as ‘k’, ‘w’ and ‘y’ are used only for personal names [Freitas 2008]. Hence, the existence of such peculiarities can further help in the writer identification process. A study into more languages that use the Latin style of writing will therefore provide us with greater insights into the stability and robustness of such a character prototyping approach in writer identification.

On the same note, it is therefore also worthwhile to look into the research area of character-based prototyping approaches for scripts that make use of non-Latin based alphabet writing systems such as Greek and Cyrillic. Since both Greek and Cyrillic scripts fall under the alphabet writing system and can thus be decomposed into the character level, it is hypothetically possible to create character prototypes to model the various handwriting styles of such handwritten documents. However, the current challenge in utilizing this character prototype approach is to find a sufficiently large enough database of Greek and Cyrillic handwritten documents that is statistically representative of the various writing styles of the letters used in these writing systems, much akin to how the IRONOFF database of French

words provides a sufficiently large representation of writing styles of the various letters of the Latin alphabet system. The importance of having statistically representative variants of the different writing styles is paramount to building robust and consistent character prototypes. Hence, a future study in this direction of other scripts such as Greek and Cyrillic that makes use of the alphabet writing system is warranted.

A challenge would then be to investigate a novel way to adapt this character prototyping approach to a non-Alphabet based writing system where the concept of a fixed representation of the writing system using a finite set of characters does not exist. Take for example the Arabic style of writing. Arabic writing systems are very dynamic and context-dependent where the allographs constantly changes depending on the context of the text passage, as well as the position of the text. Strokes of the word can either combine together to form a newer stroke and allograph depending on the context. The same word can therefore have drastically different variants of shapes, depending on the context of how it is being used, and there is no finite set of characters to represent such a writing style.

This challenging problem may be possible if we are able to consider a different interpretation of characters that combine to form a word. For example, Chinese words are formed using various combinations of basic building blocks known as radicals. A smaller set of finite and exhaustive prototypes to model some of the more common radicals could theoretically be possible to adapt a character prototyping approach even for such non-alphabet writings systems. This will allow one to take advantage of the high accuracies, robustness and large consistencies of character prototypes for applications such as in the forensic examination of parts of a word or radicals. All in all, a thorough and complete understanding of the peculiarities of the specific writing system will definitely aid in

designing and adapting a character prototype approach in Semanto-Phonetic, Syllabary or Abjad writing systems.

Last but not least, since we have proposed some future directions of how such a character prototyping approach used in this thesis can be extended to other languages of the alphabet writing systems and even to other writing systems using other scripts, it may therefore be interesting to be able to perform a preprocessing stage of script and language identification to automatically filter and route the different handwritten documents to the corresponding writer identification systems. Script and language identification, remains to this day, a challenging and largely unsolved research topic in the document analysis research community. Nonetheless, a successful breakthrough in this area will help to bridge the existing gap between writer identification of different languages and scripts. Since there is an increasing interest in handwritten documents from different languages and scripts, script and language identification will be able to help improve many pre-processing steps such as document sorting, translation and determining the choice of linguistic resources to attach in character prototype approaches such as the one used in this thesis.

5.4 Future Directions for Writer Identification: Alphabet level

Even though the concept of using the discriminative power of letters has been shown to be feasible, some open issues still remains. One underlying assumption in this chapter is that the letters are independent of one another. This might not necessarily be the case if co-dependencies between letters of the Latin alphabet exist. Such correlations between letters of the alphabet might be imagined in n-gram studies, where we can conceive letters having strong co-dependencies such as ‘th’, ‘he’ or ‘in’ in the English language. For such instances,

n-gram co-dependency cost functions could be further investigated in conjunction with the Alphabet Information Coefficient term to further improve the performance of the writer identification system. Hence, future work could be to take such co-dependencies into account when selecting a subset based on the discriminative power through a study of the various frequent and infrequent n-grams in the design of the Alphabet Information Coefficient.

Another open issue will be how to adapt this discriminative power of letters to specific writers. It would be interesting to have an indication of the probability of how similar or dissimilar the various writing styles are. Emphasis could then be given to letters of the alphabet that have higher discriminatory power for that specific writer. With this writer-specific discriminative power term, we will be able to determine letters which can be ignored for those writers based on the discriminative power of letters for that specific writer.

Another area of improvement is that our Alphabet Information Coefficient currently only considers the total number of instances of that letters of the Latin alphabet in both the reference and test documents. While this is sufficient to achieve a good writer identification rate, it ignores valuable information that can be extracted from the distribution of characters between the reference and test document. For example, if we have a ratio for of 10:10, this will give us the same Alphabet Information Coefficient as when the ratio for is 100:1.The latter ratio is highly skewed and might not sufficiently represent the handwriting style of that writer. This is also the drawback of the Alphabet Information Coefficient presented in this thesis. Thus, it will be of interest to take into consideration the distributions of both the reference and the test document when computing Alphabet Information Coefficient. One such possible evolution would be to consider two independent sigmoid functions. This will allow us to control the saturation region of the Alphabet Information Coefficient for each reference and test writer independently.

As shown earlier in this chapter, we have also established the notion that a complete set of letters of the Latin alphabet system need not always be used as the less discriminative letters of the alphabet do not contribute much towards the identification of the writer. With this idea being established in this thesis, we can extend such a notion to non-Alphabet based writing systems such as the Semanto-Phonetic writing system as discussed in section 5.3. A partial subset of the more discriminative radicals or building blocks of the word may be sufficient to represent the writing style of a writer. The challenge here would therefore to determine a subset of commonly used and highly discriminative set of radical prototypes, which are more consistent and robust than simply using grapheme prototypes. This hypothesis has to be further established in future works through a character prototyping approach for writer identification on Chinese handwritten documents.

On a similar note, it may be interesting to look at how to better adapt and utilize the effect of the Alphabet Information Coefficient on such proposed methodologies for non-Alphabet based writing system. For such adapted character prototyping approaches on non-Alphabet based writing systems, the concept of a radical information coefficient may be more suitable in prescribing a set of information criterion for writer identification. As established in this thesis, the concept of an information criterion coefficient (used in this thesis for Alphabet based writing systems) emphasizes the peculiarities of the grammatical context of the language. The Alphabet Information Coefficient is a measure of reliability of the character prototypes that can better and more pertinently model the writing style of that writer. Infrequently occurring letters of the alphabet such as ‘q’ and ‘z’ in the English language are given less priority in the Alphabet Information Coefficient since they do not model the writing style as reliably.

Extending this notion to non-Alphabet based writing systems, the information criterion coefficient in non-Alphabet based writing systems should take into consideration radicals that are infrequent and do not adequately and reliably model the writing style. At the same time, this information criterion coefficient of a smaller set of finite and exhaustive prototypes should also consider radicals that are able to sufficiently and reliably model the writing style for an effective writer identification.

Author’s Publications

Journals

[1] G. TAN, C. VIARD-GAUDIN, A. KOT
"Automatic Writer Identification Framework for Online Handwritten Documents Using Character Prototypes"
Pattern Recognition, Volume 42, Issue 12, pp. 3313-3323, 2009.

[2] G. TAN, C. VIARD-GAUDIN, A. KOT
"Individuality of alphabet knowledge in online writer identification"
International Journal on Document Analysis and Recognition, Volume 13, Number 2, June 2010, pp. 147-157

[3] C. VIARD-GAUDIN, G. TAN , A. KOT
"Identification de Scripteurs Utilisant les Distributions d'Allographes"
Traitement du Signal, Vol. 26 numéro 5, pp. 365-376, 2009/10.

International Conferences

[1] G. X. TAN, C. VIARD-GAUDIN, A. KOT
"*Online Writer Identification Using Fuzzy C-means Clustering of Character Prototypes*"

International Conference on Frontiers in Handwriting Recognition, ICFHR'2008 (IAPR), pp. 475-480, Montreal, Canada, Aug. 2008.

[2] G. TAN, C. VIARD-GAUDIN, A. KOT
"*Identification de Scripteurs basée sur une Distribution Probabiliste de Prototypes d'Allographes*"

CIFED'08 - Colloque International Francophone sur l'Ecrit et le Document, Rouen, France. pp. 139-144. Oct. 2008.

[3] G. X. TAN, C. VIARD-GAUDIN, A. KOT
"*A Stochastic Nearest Neighbor Character Prototype Approach for Online Writer Identification*"

International Conference on Pattern Recognition, ICPR 2008 (IAPR), pp. 1-4, Tampa, USA, Dec. 2008.

[4] G. X. TAN, C. VIARD-GAUDIN, A. KOT
"Online Writer Identification Using Alphabetic Information Clustering"
Proc. of SPIE-IS&T Electronic Imaging: Document Recognition and Retrieval XVI, pp. 7247 0F-1, 7247 0F-8, San Jose, Jan. 2009.

[5] G. X. TAN, C. VIARD-GAUDIN, A. KOT
"*Impact of Alphabet Knowledge on Online Writer Identification*"

International Conference on Document Analysis and Recognition, ICDAR 2009 (IAPR), pp. 56-60, Barcelona, Spain, July 2009.

[6] G. X. TAN, C. VIARD-GAUDIN, A. KOT
"*Information Retrieval Model for Online Handwritten Script Identification*"

International Conference on Document Analysis and Recognition, ICDAR 2009 (IAPR), pp. 336-340, Barcelona, Spain, July 2009.

Bibliography

Bensefia, A., A. Nosary, T. Paquet and L. Heutte (2001), "Writer Identification by Writer's Invariants." *International Workshop on Frontiers in Handwriting Recognition*, pp. 274-279.

Bensefia, A., T. Paquet and L. Heutte (2003), "Information retrieval based writer identification." *Proceedings. Seventh International Conference on Document Analysis and Recognition*, pp. 946-950.

Bensefia, A., T. Paquet and L. Heutte (2005), "Handwritten Document Analysis for Automatic Writer Recognition." *Electronic Letters on Computer Vision and Image Analysis*, 5, pp. 72-86.

Bensefia, A., T. Paquet and L. Heutte (2005), "A writer identification and verification system." *Pattern Recognition Letters*, 26 (13), pp.2080-2092.

Berlitz (1993). "*Berlitz Russian Phrase Book & Dictionary."* Berlitz, Princeton, New Jersey.

Bulacu, M. and L. Schomaker (2007), "Text-independent writer identification and verification using textural and allographic features." *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 29 (4), pp.701-717.

Busch, A., W. W. Boles and S. Sridharan (2005), "Texture for script identification." *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 27 (11), pp.1720-1732.

Cha, S., S. Yoon and C. C. Tappert (2006), "Handwriting Copybook Style Identification for Questioned Document Examination." *Journal of Forensic Document Examiners*, 17 pp.1-14.

Chaabouni, A., H. Boubaker, M. Kherallah, A. M. Alimi and H. E. Abed (2011), "Multi-fractal Modeling for On-line Text-Independent Writer Identification." *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 623-627.

Chan, S. K., C. Viard-Gaudin and Y. H. Tay (2008), "Online writer identification using character prototypes distributions." *SPIE-IS&T Electronic Imaging: Document Recognition and Retrieval XV*, 6815, pp. 1-5.

Chan, S. K., H. T. Yong and C. Viard-Gaudin (2007), "Online text independent writer identification using character prototypes distribution." *6th International Conference on Information, Communications and Signal Processing, ICICS*, pp. 1-5.

Chawki, D. and S. M. Labiba (2010), "A texture based approach for Arabic writer identification and verification." *2010 International Conference on Machine and Web Intelligence (ICMWI)*, pp. 115-120.

Chellapilla, K., P. Simard and A. Abdulkader (2006), "Allograph based writer adaptation for handwritten character recognition." *10th International Workshop on Frontiers in Handwriting Recognition*, pp. 423-428.

Chen, S. M. and Y. J. Horng (1999), "Fuzzy query processing for document retrieval based on extended fuzzy concept networks." *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 29 (1), pp.96-104.

Connell, S. D. and A. K. Jain (2002), "Writer adaptation for online handwriting recognition." *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 24 (3), pp.329-346.

Crosier, M. and L. Griffin (2010), "Using Basic Image Features for Texture Classification." *International Journal of Computer Vision*, 88 (3), pp.447-460.

Ethnologue (2012), "Languages of the World." Retrieved August, 2012, from http://www.ethnologue.com/ethno_docs/distribution.asp?by=family.

Fiel, S. and R. Sablatnig (2012), "Writer Retrieval and Writer Identification Using Local Features." *10th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 145-149.

Foster, C. C. (1982). "*Cryptanalysis for microcomputers."* Hayden Book, New Jersey.

Freitas, C., Oliveira, L.S., Sabourin, R., Bortolozzi, F. (2008), "Brazilian Forensic Letter Database." *11th International Workshop on Frontiers on Handwriting Recognition*, pp. 1-6.

Fujisawa, H. (2008), "Forty years of research in character and document recognition-an industrial perspective." *Pattern Recogn.*, 41 (8), pp.2435-2446.

Garain, U. and T. Paquet (2009), "Off-line multi-script writer identification using AR coefficients." *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 991-995.

Gizmowatch (2012), "Nokia SU-1B Digital Pen." Nokia SU-1B Digital Pen, Retrieved September, 2012, from http://www.gizmowatch.com/entry/nokia-digital-pen-su-1b-saving-100-pages-of-text-just-got-simpler/.

Guyon, I., P. Albrecht, Y. L. Cun, J. Denker and W. Hubbard (1991), "Design of a neural network character recognizer for a touch terminal." *Pattern Recogn.*, 24 (2), pp.105-119.

Han, J. and M. Kamber (2006). "*Data Mining - Concepts and Techniques."* Elsevier, San Francisco.

Hanusiak, R., L. Oliveira, E. Justino and R. Sabourin (2012), "Writer verification using texture-based features." *International Journal on Document Analysis and Recognition*, 15 (3), pp.213-226.

Hassaïne, A. and S. Al-Maadeed (2012), "ICFHR 2012 Competition on Writer Identification - Challenge 1: Arabic Scripts." *ICFHR*, Bari, Italy. pp. 831-836.

Hassaïne, A., S. Al-Maadeed, J. M. Alja'am and A. J. Bouridane (2011), "The ICDAR 2011 Arabic Writer Identification Contest." *ICDAR*, Beijing, China. pp. 1470-1474.

He, Z., X. You and Y. Y. Tang (2008), "Writer identification of Chinese handwriting documents using hidden Markov tree model." *Pattern Recognition*, 41 (4), pp.1295-1307.

Helli, B. and M. E. Moghaddam (2010), "A text-independent Persian writer identification based on feature relation graph (FRG)." *Pattern Recognition*, 43 (6), pp.2199-2209.

Hochberg, J., P. Kelly, T. Thomas and L. Kerns (1997), "Automatic script identification from document images using cluster-based templates." *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 19 (2), pp.176-181.

Hoppner, F., F. Klawonn and R. K. Runkler (1999). "*Fuzzy Cluster Analysis - Methods for Classification, Data Analysis and Image Recognition."* Wiley.

Huber, R. A. and A. M. Headrick (1999). "*Handwriting Identification - Facts and Fundamentals."* CRC Press, Boca Raton.

Jain, A. K. and A. M. Namboodiri (2003), "Indexing and retrieval of on-line handwritten documents." *ICDAR*, pp.655-659.

Jin, L., Y. Gao, G. Liu, Y. Li and K. Ding (2011), "SCUT-COUCH2009—a comprehensive online unconstrained Chinese handwriting database and benchmark evaluation." *International Journal on Document Analysis and Recognition*, 14 (1), pp.53-64.

Kelly, J. S. and B. S. Lindblom (2006). "*Scientific examination of questioned documents."* CRC Press, Boca Raton.

Kneser, R. and H. Ney (1995), "Improved backing-off for M-gram language modeling." *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 1, pp. 181-184.

Levinson, J. (2001). "*Questioned Documents: A Lawyer's Handbook."* Academic Press.

Louloudis, G., B. Gatos and N. Stamatopoulos (2012), "ICFHR 2012 Competition on Writer Identification Contest Challenge 1: Latin/Greek Documents." *ICFHR*, Bari, Italy. pp. 825-830.

Louloudis, G., N. Stamatopoulos and B. Gatos (2011), "ICDAR 2011 Writer Identification Contest." *ICDAR*, Beijing, China. pp. 1475-1479.

Lyne, A. A. (1985). "*The Vocabulary of French Business Correspondence: Word Frequencies, Collocations, and Problems of Lexicometric Method."* Slatkine.

Marcelli, A., A. Parziale and A. Santoro (2012), "Modeling Handwriting Style: a Preliminary Investigation." *International Conference on Frontiers in Handwriting Recognition*, pp. 409-413.

Marti, U. V. and H. Bunke (1999), "A full English sentence database for off-line handwriting recognition." *Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on*, pp. 705-708.

Morris, R. N. (2000). "*Forensic handwriting identification - Fundamental concepts and principles."* Academic Press, London.

MyScript (2012), "Vision Objects Industrial Text Recogniser SDK MyScript Builder Help." SDK documentation, Retrieved August, 2012, from http://www.visionobjects.com/about-us/download-center/_263/myscript-products-datasheets.html.

Namboodiri, A. M. and S. Gupta (2006), "Text independent writer identification from online handwriting." *Proceedings of Int'l workshop of Frontier in Handwriting Recognition*, pp. 23-26.

Namboodiri, A. M. and A. K. Jain (2004), "Online Handwritten Script Recognition." *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 26 (1), pp.124-130.

Niels, R., F. Gootjen and L. Vuurpijl (2008), "Writer Identification through Information Retrieval: The Allograph Weight Vector." *International Conference on Frontiers in Handwriting Recognition*, pp. 481-486.

Niels, R., L. Vuurpijl and L. Schomaker (2007), "Automatic allograph matching in forensic writer identification." *International Journal of Pattern Recognition and Artificial Intelligence*, 21 (1), pp.61-81.

Omniglot (2012), "Types of writig system." Retrieved August, 2012, from http://www.omniglot.com/writing/types.htm.

Oviatt, S., P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson and D. Ferro (2000), "Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions." *Human-Computer Interaction*, 15 (4), pp.263-322.

PCMag (2012), "Anoto Digital Paper and Nokia Digital Pen." Anoto Digital Paper and Nokia Digital Pen, Retrieved August, 2012, from http://www.pcmag.com/slideshow_viewer/0,3253,l%253D147579%2526a%253D147579%2526po%253D1,00.asp?p=n.

Pervouchine, V. and G. Leedham (2007), "Extraction and analysis of forensic document examiner features used for writer identification." *Pattern Recognition*, 40 (3), pp.1004-1013.

Peter Allen Miller, J. L. S., Cullen I. K. Story (2002). "*Greek to Me."* Xulon Press.

Pitak, T. and T. Matsuura (2004), "On-line writer recognition for Thai based on velocity of barycenter of pen-point movement." *International Conference on Image Processing*, 2, pp. 889-892.

Quang Anh, B., M. Visani, S. Prum and J. Ogier (2011), "Writer Identification Using TF-IDF for Cursive Handwritten Word Recognition." *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 844-848.

Radecki, T. (1983), "Generalized Boolean methods of information retrieval." *International Journal of Man-Machine Studies*, 18 (5), pp.407-439.

Rosenbaum, R. and M. Fleischmann (2003), "Character Frequency in Multilingual Corpus 1 - Part 2." *Journal of Quantitative Linguistics*, 10 (1), pp.1 - 39.

S. Mozaffari, H. E. A., V. Margner, K. Faez, A. and Amirshahi (2008), "IfN/Farsi-Database: A Database of Farsi Handwritten City Names." *ICFHR*, pp. 1-6.

Said, H. E. S., T. N. Tan and K. D. Baker (2000), "Personal identification based on handwriting." *Pattern Recognition*, 33 (1), pp.149-160.

Saldarriaga, S. P., C. Viard-Gaudin and E. Morin (2009), "On-line Handwritten Text Categorization." *SPIE-IS&T Electronic Imaging: Document Recognition and Retrieval XVI*, pp. 724709-724701-724711.

Salton, G. and C. Buckley (1988), "Term-weighting approaches in automatic text retrieval." *INFORMATION PROCESSING AND MANAGEMENT*, 24, pp. 513-523.

Salton, G., A. Wong and C. S. Yang (1975), "A vector space model for automatic indexing." *Commun. ACM*, 18 (11), pp.613-620.

Schlapbach, A. and H. Bunke (2004), "Using HMM based recognizers for writer identification and verification." *Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR*, Tokyo. pp. 167-172.

Schlapbach, A., M. Liwicki and H. Bunke (2008), "A writer identification system for on-line whiteboard data." *Pattern Recognition*, 41 (7), pp.2381-2397.

Schomaker, L. and M. Bulacu (2004), "Automatic writer identification using connected-component contours and edge-based features of uppercase western script." *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 26 (6), pp.787-798.

Sesa-Nogueras, E. and M. Faundez-Zanuy (2012), "Writer recognition enhancement by means of synthetically generated handwritten text." *Engineering Applications of Artificial Intelligence*, 26 (1), pp.609-624.

Shivram, A., C. Ramaiah, U. Porwal and V. Govindaraju (2012), "Modeling Writing Styles for Online Writer Identification: A Hierarchical Bayesian Approach." *International Conference on Frontiers in Handwriting Recognition*, pp. 385-390.

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

Srihari, S. and G. Leedham (2003), "A survey of computer methods in forensic document examination." *Proceedings of 11th Conference International on Graphonomics Society (IGS2003)*, pp. 278-281.

Srihari, S. N., M. J. Beal, K. Bandi, V. Shah and P. Krishnamurthy (2005), "A statistical model for writer verification." *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2005, pp. 1105-1109.

Srihari, S. N., S.-H. Cha and S. Lee (2001), "Establishing Handwriting Individuality Using Pattern Recognition Techniques." *Proceedings of the Sixth*

International Conference on Document Analysis and Recognition, pp. 1195-1204.

Srihari, S. N., S. Cha, H. Arora and S. Lee (2001), "Individuality of handwriting: a validation study." *Sixth International Conference on Document Analysis and Recognition*, pp. 106-109.

Srihari, S. N., S. H. Cha, H. Arora and S. Lee (2002), "Individuality of handwriting." *Journal of Forensic Sciences*, 47 (4), pp.856-872.

Srinivasan, H., S. Kabra, H. Chen and S. Srihari (2007), "On Computing Strength of Evidence for Writer Verification." *Ninth International Conference on Document Analysis and Recognition*, 2, pp. 844-848.

Sungsoo, Y., C. Seungseok, C. Sung-Hyuk and C. C. Tappert (2005), "Writer profiling using handwriting copybook styles." *Eighth International Conference on Document Analysis and Recognition*, pp. 600-604.

Viard-Gaudin, C., P. M. Lallican, S. Knerr and P. Binter (1999), "The ireste on/off (ironoff) dual handwriting database." *IEEE Int. Conf. Document Analysis and Recognition*, pp.455-458.

Vuurpijl, L. and L. Schomaker (1997), "Finding structure in diversity: a hierarchical clustering method for the categorization of allographs in handwriting." *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, 1, pp. 387-393.

Witten, I. H. and T. C. Bell (1991), "The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression." *IEEE Transactions on Information Theory*, 37 (4), pp.1085-1094.

Xu, L., X. Ding, L. Peng and X. Li (2011), "An Improved Method Based on Weighted Grid Micro-structure Feature for Text-Independent Writer Recognition."

Proceedings of the 2011 International Conference on Document Analysis and Recognition, pp. 638-642.

Yasuda, H., K. Takahashi and T. Matsumoto (2000), "A discrete HMM for online handwriting recognition." *International Journal of Pattern Recognition and Artificial Intelligence*, 14 (5), pp.675-688.

Zois, E. N. and V. Anastassopoulos (2000), "Morphological waveform coding for writer identification." *Pattern Recognition*, 33 (3), pp.385-398.

l'université nantes angers le mans

PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

UNIVERSITÉ DE NANTES

Thèse de Doctorat

GuoXian TAN

Writing Style Modelling Based on Grapheme Distributions - Application to On-Line Writer Identification

Modélisation des Styles d'écriture Basée Distributions de Graphèmes – Application à l'Identification de Scripteurs

Abstract

In this work, we propose to model the unique handwritten style of a person by computing the grapheme distribution produced by this writer. This distribution is computed from any text with a few lines. First, representative grapheme prototypes are automatically defined at the character level with a clustering algorithm. These prototypes should cover the variability of the different writing styles. Next, the modeling distribution of two writing styles can be compared, which allows to proceed to writer identification from a reference dataset of documents and a questioned document. The query, unknown writer document, is matched again all the reference documents. For this purpose, the proposed framework segments and recognizes the text at the character level and then performs a fuzzy function assignment to the corresponding prototypes of all the letters.

Some important issues are raised. They concern the number of prototypes for representing a letter, the choice of a metric to compare two distributions, the discriminative power of the alphabet letters, the effect of the length of the texts.

Two datasets with different complexities have been used to assess the performances of the proposed system. A writer identification rate of 99.2% has been reached with a set of 120 French writers, whereas with a bigger set of 200 English writers, the identification rate is of 87%.

This method has been applied on on-line handwriting where the available signal which defines the sampled trajectory of the writing tool is available as a sequence of points.

Key Words
Writer identification, information retrieval, on-line handwriting, grapheme, k-nearest neighbor.

Résumé

Dans cette thèse nous proposons de modéliser le style d'écriture manuscrite spécifique d'une personne en utilisant pour un scripteur donné la distribution de l'usage de prototypes de graphèmes. Cette distribution est calculée à partir d'un texte quelconque de quelques lignes. Les prototypes résultent d'un clustering préalable et indépendant permettant de recenser la variabilité des styles d'écriture. Cette modélisation permet de comparer deux styles d'écriture et de procéder à l'identification d'un scripteur à partir d'une base de documents de référence. La comparaison est basée sur une mesure de mise en correspondance des distributions obtenues. Pour cela, un système automatique segmente le texte en lettres, puis classe chaque lettre de manière probabiliste parmi les prototypes disponibles pour cette lettre.

Une analyse portant sur le choix du nombre de prototypes, la nature de la distance utilisée pour la comparaison, le caractère discriminant des différentes lettres de l'alphabet, et la longueur du texte disponible est proposée.

Deux bases de complexité différentes sont utilisées pour évaluer ce système. Un taux d'identification de 99,2 % est obtenu sur une base de recherche de 120 textes écrits en français, tandis qu'il se situe à 87 % sur une base de recherche de 200 textes écrits en anglais.

Cette méthode est développée sur de l'écriture en ligne où le signal d'écriture représentant la trajectoire de l'instrument d'écriture est disponible sous la forme d'une séquence de points.

Mots clés
Identification de scripteur, recherche d'information, écriture manuscrite en-ligne, allographe, k-plus-proches-voisins.