Chapter 1

Introduction

Truth is what stands the test of experience.
Albert Einstein

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

• How can individual handwriting style be characterized using computer algorithms?
• What representations or features are most appropriate and how can they be combined?
• What performance can be achieved using automatic methods?

The current study describes a number of new and very effective techniques that we have developed for automatic writer identification and verification. The goal of our research was to design state-of-the-art automatic methods involving only a reduced number of adjustable parameters and to create a robust writer identification system capable of managing hundreds to thousands of writers.

There are two distinguishing characteristics of our approach: human intervention is minimized in the writer identification process and we encode individual handwriting style using features designed to be independent of the textual content of the handwritten sample. Writer individuality is encoded using probability distribution functions extracted from handwritten text blocks and, in our methods, the computer is completely unaware of what has been written in the samples.

The development of our writer identification techniques takes place at a time when many biometric modalities undergo a transition from research to real full-scale deployment. Our methods also have practical feasibility and hold the promise of concrete applicability.
The writer identification techniques proposed in this thesis have possible impact in forensic science. Our methods are statistically evaluated using large datasets with handwriting samples collected from up to 900 subjects.

1.1 Writer identification as a behavioral biometric

Biometric modalities are classified into two broad categories: physiological biometrics that perform person identification based on measuring a physical property of the human body (e.g. fingerprint, face, iris, retinal blood vessels, hand geometry, DNA) and behavioral biometrics that use individual traits of a person’s behavior for identification (e.g. voice, gait, keystroke dynamics, signature, handwriting). Writer identification therefore pertains to the category of behavioral biometrics. From the physical body property or the individual behavior traits, biometric templates are extracted and used in the identification process. Biometric identification is performed by comparing the biometric template measured at the moment when the identification of an unknown person is needed with templates previously enrolled in a database and linked with certainty to known persons.

Physiological biometrics, like fingerprint (Jain et al. 1997, Moler et al. 1998), iris (Daugman 1993, Daugman 2003) or DNA (Devlin et al. 1992, Benecke 1997), are strong modalities for person identification due to the reduced variability and high complexity of the biometric templates used. However, these physiological modalities are usually more invasive and require cooperating subjects. On the contrary, behavioral biometrics are less invasive, but the achievable performance is less impressive due to the large variability of the behavior-derived biometric templates.

Leading a worrisome life among the harder forms of biometrics, the identification of a person on the basis of handwriting samples still remains a useful biometric modality, mainly due to its applicability in the forensic field.

1.2 Writer identification in forensics

Contrary to other forms of biometric person identification used in forensic labs, automatic writer identification often allows for determining identity in conjunction with the intentional aspects of a crime, such as in the case of threat or ransom letters. This is a fundamental difference from other biometric methods, where the relation between the evidence material and the details of an offense can be quite remote.

The target performance for writer identification systems is less impressive than in the case of DNA or iris-based person identification. In forensic writer identification, as a rule of thumb, one strives for a near-100% recall of the correct writer in a hit list of
1.2. Writer identification in forensics

100 writers, computed from a database in the order of $10^4$ samples, the size of search sets in current European forensic databases. This amount is based on the pragmatic consideration that a number of one hundred suspects is just about manageable in the criminal-investigation process. This target performance still remains an ambitious goal.

Recent advances in image processing, pattern classification and computer technology at large have provided the context in which our research was carried out. The writer identification techniques that we developed accomplished substantial improvements in performance and have potential applicability in forensic practice.

There exist three groups of script-shape features which are derived from scanned handwritten samples in forensic procedures:

1. Fully automatic features computed from a region of interest in the image;

2. Interactively measured features by human experts using a dedicated graphical user-interface tool;

3. Character-based features which are related to the allograph subset which is being generated by each writer.

The complete process of forensic writer identification is never fully automatic. The features pertaining to groups 2 and 3 require some form of intensive human involvement in executing predefined measuring actions on the script image or in isolating and labeling individual characters or words. Two examples of actual forensic writer identification systems are 

Fish (Philipp 1996) and Script (de Jong et al. 1994).

Although requiring less human labor, the first group of features has been treated with some skepticism by practitioners within the application domain, given the complexity of the real-life scanned samples of handwriting that are collected in practice. Indeed, automatic foreground/background separation will often fail on the smudged and texture-rich fragments, where the ink trace is often hard to identify. However, there are recent advances in image processing using “soft computing” methods, i.e., combining tools from fuzzy logic and genetic algorithms, which allow for advanced semi-interactive solutions to the foreground/background separation process (Franke and Köppen 1999, Franke and Köppen 2001, Köppen and Franke 1999). Under these conditions, and assuming the presence of sufficient computing power, the use of automatically computed image features (group 1 from above) is becoming feasible.

The current thesis sets out to explore precisely this category of automatic features. It is implicitly assumed that a crisp foreground/background separation has already been realized in a pre-processing phase, yielding a white background with (near-) black ink.
1.3 Writer identification vs. Handwriting recognition

Writer identification is rooted in the older and broader automatic handwriting recognition domain. For automatic handwriting recognition, invariant representations are sought which are capable of eliminating variations between different handwritings in order to classify the shapes of characters and words robustly. The problem of writer identification, on the contrary, requires a specific enhanced representation of these variations, which, per se, are characteristic to a writer’s hand.

Due to its very large applicability, handwriting recognition has always dominated research in handwriting analysis. Writer identification received renewed interest in the last several years, after 9/11 and the anthrax letters (Schomaker and Bulacu 2004, Srihari et al. 2002, Bensefia et al. 2005b, Schlapbach and Bunke 2004, Said et al. 2000, Zois and Anastassopoulos 2000).

The goal in handwriting recognition is to obtain invariance and generalization. For writer identification, one strives for quite the opposite with the aim to maximally expose the specificity of individual handwriting style for writer discrimination.

It is important, however, to mention the idea that writer identification could reduce certain ambiguities in the pattern recognition process if information on the writer’s general writing habits and idiosyncrasies is available to the handwriting recognition system (Maarse 1987, Crettez 1995).


1.4 Writer identification vs. Writer verification

Asserting writer identity based on handwriting images requires three main operational phases after image preprocessing:

- feature extraction
- feature matching / feature combination
- writer identification and verification
1.5 Text-dependent vs. Text-independent methods

Writer identification and verification approaches fall into two broad categories: text-dependent vs. text-independent methods (Plamondon and Lorette 1989).

The text-dependent methods are very similar to signature verification techniques and use the comparison between individual characters or words of known semantic (ASCII) content (see Fig. 1.3). These methods therefore require the prior localization and segmentation of the relevant information. This is usually performed interactively by a
human user.

The \textit{text-independent} methods for writer identification and verification use statistical features extracted from the entire image of a text block. A minimal amount of handwriting (e.g. a paragraph containing a few text lines) is necessary in order to derive stable features insensitive to the text content of the samples. Our approach falls in this latter category. From the application point of view, the notable advantage is that human intervention is minimized.

Typical for the text-independent approaches and therefore a defining property of our approach as well, the features used for writer identification provide a lumped description of the whole region containing handwriting by discarding location information. For this reason, it is questionable to use text-independent methods also in the cases where the textual content of the samples is fixed and known.

\section{1.6 Within-writer variance vs. Between-writer variation}

Writer identification and verification are only possible to the extent that the variation in handwriting style between different writers exceeds the variations intrinsic to every single writer considered in isolation. The results reported in this thesis ultimately represent statistical analyses, on our datasets, of the relationship opposing the \textit{between-writer variation} and the \textit{within-writer variability} in feature space.

The present study assumes that the handwriting was produced using a natural writing attitude. It is important to observe that forged or disguised handwriting is not addressed in our approach. The forger tries to change the handwriting style usually by changing the slant and/or the chosen allographs. Using detailed manual analysis, forensic experts are sometimes able to correctly identify a forged handwritten sample (Huber and Headrick 1999, Morris 2000). On the other hand, our proposed algorithms operate on the scanned handwriting faithfully considering all the graphical shapes encountered in the image under the premise that they are created by the habitual and natural script style of the writer.
1.7 Factors causing variability in handwriting

Figure 1.3: A comparison of handwritten characters (allographs) and handwritten words from three different writers.

1.7 Factors causing variability in handwriting

Figure 1.4 shows four factors causing variability in handwriting (Schomaker 1998).

The first factor concerns the affine transforms (Fig. 1.4a), which are under voluntary control by the writer. Transforms of size, translation, rotation and shear are a nuisance, but not a fundamental obstacle in handwriting recognition or writer identification. In particular, slant (shear) constitutes a habitual parameter determined by pen grip and orientation of the wrist subsystem versus the fingers (Dooijes 1983).

The second factor concerns the neuro-biomechanical variability (Fig. 1.4b) which is sometimes referred to as "sloppiness space": the local context and physiological state determine the amount of effort that is spent on character-shape formation and determine the legibility of the written sample. In realizing the intended shape, a writer must send motor-control patterns which compensate for the low-pass filtering effects of the biomechanical end-effector. This category of variability sources also contains tremors and effects of psychotropic substances on motor-control processes in writing. As such,
a) Affine transforms

\[
\begin{array}{cccc}
\text{alpha} & \text{alpha} & \text{alpha} & \text{alpha} \\
\end{array}
\]

b) Neuro–biomechanical variability

\[
\begin{array}{cccc}
\text{worm} & \text{worm} & \text{worm} & \text{worm} \\
\end{array}
\]

c) Sequencing variability

\[
\begin{array}{cccc}
1 & 2 & 2 & 2 \\
4 & 1 & 1 & 1 \\
\end{array}
\]

d) Allographic variation

\[
\begin{array}{cccc}
\text{ALPHA} & \text{alpha} & \text{alpha} & \text{alpha} \\
\end{array}
\]

**Figure 1.4:** Factors causing handwriting variability: (a) Affine transforms are under voluntary control. However, writing slant constitutes a habitual parameter which may be exploited in writer identification; (b) neuro-biomechanical variability refers to the amount of effort which is spent on overcoming the low-pass characteristics of the biomechanical limb by conscious cognitive motor control; (c) sequencing variability becomes evident from stochastic variations in the production of the strokes in a capital E or of strokes in Chinese characters, as well as stroke variations due to slips of the pen; (d) allographic variation refers to individual use of character shapes. Factors b) and c) represent system state more than system identity. Affine transforms (a) and allographic variation (d) are the most useful sources of information in writer identification and verification.

this factor is more related to system state than system identity.

The third factor is also highly dependent on the instantaneous system state during the handwriting process and is represented by sequencing variability (Fig. 1.4c): the stroke order may vary stochastically, as in the production of a capital E. A four-stroked E can be produced in \(4! \times 2^4 = 384\) permutations. In the production of some Asian scripts, such as Hanzi, stochastic stroke-order permutations are a well-known problem in handwriting recognition (even though the training of stroke order at schools is rather strict). Finally, spelling errors may occur and lead to post-hoc editing strokes in the writing sequence. Although sequencing variability is generally assumed to pose a problem only for handwriting recognition based on temporal (on-line) signals, the example of post-hoc editing
(Fig. 1.4c) shows that static, optical effects are also a possible consequence of this form of variation.

The fourth factor, *allographic variation* (Fig. 1.4d and Fig. 1.3), refers to the phenomenon of writer-specific character shapes, which produces most of the problems in automatic script recognition, but at the same time provides essential information for automatic writer identification.

The handwriting of a person also changes with age and this constitutes another important variability factor. As a child grows, his handwriting becomes more comfortable, rapid, smooth, continuous, rhythmic and without hesitation. The amount of time a person spends writing may determine his general skill level and speed of writing. At older age, handwriting may become impaired due to chronic conditions that affect hand strength and dexterity.

### 1.8 Factors determining individuality of handwriting

As the writer matures, he departs from the copybook style learned in the classroom and progressively incorporates into his writing his own individuality. Especially nowadays, when there is less emphasis on penmanship in school.

There exist two fundamental factors contributing to the individuality of script: *genetic* (biological) and *memetic* (cultural) factors.

The first fundamental factor consists of the *genetic* make up of the writer. Genetic factors are known or may be hypothesized to contribute to handwriting style individuality:

- The biomechanical structure of the hand, i.e., the relative sizes of the carpal bones of wrist and fingers and their influence on pen grip;
- The left or right handedness (Francks et al. 2003);
- Muscular strength, fatigability, peripheral motor disorders (Gulcher et al. 1997);
- Central-nervous system (CNS) properties, i.e., aptitude for fine motor control and the CNS stability in motor-task execution (Van Galen et al. 1993).

The second factor consists of *memetic* or culturally transferred influences (Moritz 1990) on pen-grip style and the character shapes (allographs) which are trained during education or are learned from observation of the writings of other persons. Although the term *memetic* is often used to describe the evolution of ideas and knowledge, there does not seem to be a fundamental objection to view the evolution and spreading of character shapes as a memetic process: the fitness function of a character shape depends
on the conflicting influences of (a) legibility and (b) ease of production with the writing tools (Jean 1997) which are available within a culture and society. The distribution of allographs over a writer population is heavily influenced by writing methods taught at school, which in turn depend on factors such as geographic distribution, religion and school types.

Together, the genetic and memetic factors determine a habitual writing process, with recognizable shape elements at the local level in the writing trace, at the level of the character shape as a whole and at the level of character placement and page layout. In this thesis, we will focus on the local level in the handwritten trace and on the character level.

Handwriting can be described as a hierarchical psychomotor process: at a high level, an abstract motor program is recovered from long-term memory; parameters are then specified for this motor program, such as size, shape, timing; finally, at a peripheral level, commands are generated for the biophysical muscle-joint systems (Maarse 1987). Writing consists of rapid movements of the fingers and the hand, and superimposed on this a slow progressive horizontal movement of the lower arm. In experiments performed by fixing the lower arm (Maarse 1987) (see Fig. 1.5), Maarse has studied on-line handwriting produced by using only two biophysical systems: one consisting of the
1.8. Factors determining individuality of handwriting

Figure 1.6: The experimental set-up from Fig. 1.5 was used to record on-line signals of simple movements and complete handwriting. a) Recorded movements: the top traces show hand rotating around the wrist (X′ direction) and finger (Y′ direction) movements reflecting the biomechanical geometry and predominant writing direction. b) Recorded handwriting: writing slant changes are considerably smaller than the orientation changes of the Y′ system. Handwriting slant is held constant by an intricate interaction between the X′ and Y′ subsystems. Reprinted from (Maarse 1987), with kind permission from the author.

thumb and fingers (Y′), and the other consisting of the entire hand rotating around the wrist (X′), from radial abduction to ulnar abduction. Maarse shows (see Fig. 1.6) that the changes observed in writing directions are less than the changes in the orientation of the effector subsystems, with the conclusion that “this unexplained slant constancy may be caused by a setting of writing slant in a motor program at a higher level” (Maarse 1987). The writer therefore tries to maintain his / her preferred slant and letter shapes over the complete range of motion in the biomechanical systems thumb-fingers and hand-wrist.

Additional evidence regarding the constancy of individual writing habits is provided in another study (Maarse and Thomassen 1983) observing that changes in the horizontal progression motion affect predominantly the up strokes, while the down strokes maintain their direction correlated with the perceived slant of the manuscript. Up stokes contain practically all the connecting strokes between letters, whereas down strokes appear relatively much more as parts of actual letters or graphemes. Maarse affirms that down strokes “might change less if there is a tendency to keep the grapheme features unchanged, either because the visual appearance of the product is preferably
held constant, or because the motor program for the characteristic of the graphemes remains more or less constant”, as a result of writing education (Maarse 1987).

The writer produces a pen-tip trajectory on the writing surface in two dimensions (x,y), modulating the height of the pen tip above the surface by vertical movement (z). Displacement control is replaced by force control (F) at the moment of landing. The pen-tip trajectory in the air between two pen-down components contains valuable writerspecific information, but its shape is not known in the case of off-line scanned handwritten samples. Similarly, pen-force information is highly informative of a writer’s identity, but is not directly known from off-line scans (Schomaker and Plamondon 1990).

An important theoretical basis for the usage of handwritten shapes for writer identification is the fact that handwriting is not a feedback process which is largely governed by peripheral factors in the environment. Due to neural and neuromechanical propagation delays, a handwriting process based upon a continuous feedback mechanism alone would evolve too slowly (Schomaker 1991). Hence, the brain is continuously planning series of ballistic movements ahead in time, i.e., in a feed-forward manner. A character is assumed to be produced by a ”motor program” (Schmidt 1975), i.e., a configurable movement-pattern generator which requires a number of parameter values to be specified before being triggered to produce a pen-tip movement yielding the character shape (Schomaker et al. 1989, Plamondon and Maarse 1989, Plamondon and Guerfali 1998) by means of the ink deposits (Doermann and Rosenfeld 1992, Franke and Grube 1998, Franke 2005). The final resulting shape on paper represents a variation around the ”master pattern” stored centrally, in the motor memory of the writer. Although the process described thus far is concerned with continuous variables such as displacement, velocity and force control, the linguistic basis of handwriting allows for postulating a discrete symbol from an alphabet to which a given character shape refers.

This thesis will show that very effective writer identification and verification is achievable by combining local directional features informative about habitual pen grip and slant with allograph shape features informative about the character forms engrained in the motor memory of the writer.

1.9 A survey of recent research in the field

In this section, we present a review of the recent papers published on the topic of automatic writer identification in order to provide a general literature background for our own research work contained in this thesis. A comprehensive review covering the period until 1989 is given in (Plamondon and Lorette 1989) and we provide a number of references in the bibliography section: (Arazi 1977, Kuckuck et al. 1979, Klement et al. 1980, Kuckuck 1980, Steinke 1981, Klement 1981, Dinstein and Shapira 1982, Naske

Here we will survey the approaches proposed in the last several years, as a result of the renewed interest in the scientific community for this research topic. Throughout the survey, we will make clear the distinction between text-dependent versus text-independent approaches.

(Said et al. 2000, Said et al. 1998) propose a text-independent approach and derive writer-specific texture features using multichannel Gabor filtering and gray-scale co-occurrence matrices. The method requires uniform blocks of text that are generated by word deskewing, setting a predefined distance between text lines / words and text padding. Two sets of 20 writers, 25 samples per writer are used in the evaluation. Nearest-centroid classification using weighted Euclidean distance and Gabor features achieved 96% writer identification accuracy. A similar approach has also been used on machine-print documents for script (Tan 1998) and font (Zhu et al. 2001) identification.

(Zois and Anastassopoulos 2000) perform writer identification and verification using single words. Experiments are performed on a dataset containing 50 writers. The word ‘characteristic’ was written 45 times by each writer, both in English and in Greek. After image thresholding and curve thinning, the horizontal projection profiles are resampled, divided into 10 segments and processed using morphological operators at two scales to obtain 20-dimensional feature vectors. Classification is performed using either a Bayesian classifier or a multilayer perceptron. Accuracies around 95% are obtained both for English and Greek words.

(Srihari et al. 2002, Srihari et al. 2005) propose a large number of features divided into two categories. Macro-features operating at document / paragraph / word level: gray-level entropy and threshold, number of ink pixels, number of interior / exterior contours, number of 4-direction slope components, average height / slant, paragraph aspect ratio and indentation, word length and upper / lower zone ratio. Micro-features operating at word / character level: gradient, structural and concavity (GSC) attributes, used originally for handwritten digit recognition (Favata and Srikanthan 1996). Text-dependent statistical evaluations are performed on a dataset containing 1000 writers who copied 3 times a fixed text of 156 words (the CEDAR letter). This is the largest dataset used up to the present in writer identification studies. Micro-features are better than macro-features in identification tests with a performance exceeding 80%. A multilayer perceptron or parametric distributions are used for writer verification with an accuracy of about 96%. Writer discrimination was also evaluated using individual characters (Zhang et al. 2003, Srihari et al. 2003) and words (Zhang and Srihari 2003, Tomai et al. 2004).

(Bensefia et al. 2005b, Bensefia et al. 2005a, Bensefia et al. 2002, Bensefia et al. 2003) use graphemes generated by a handwriting segmentation method to encode the individual characteristics of handwriting independent of the text content. Our allograph-
level approach is similar to the work reported in these studies. Grapheme clustering is used to define a feature space common for all documents in the dataset. Experimental results are reported on three datasets containing 88 writers, 39 writers (historical documents) and 150 writers, with 2 samples (text blocks) per writer. Writer identification is performed in an information retrieval framework, while writer verification is based on the mutual information between the grapheme distributions in the two handwritings that are compared. Concatenations of graphemes are also analyzed in the mentioned papers. Writer identification rates around 90% are reported on the different test datasets.

(Marti et al. 2001) and (Hertel and Bunke 2003) take text lines as the basic input unit from which text-independent features are computed using the height of the three main writing zones, slant and character width, the distances between connected components, the blobs enclosed inside ink loops, the upper / lower contours and the thinned trace processed using dilation operations. A feature selection study is also performed in (Schlapbach et al. 2005). Using a k-nearest-neighbor classifier, identification rates exceeding 92% are obtained in tests on a subset of the IAM database (Marti and Bunke 2002) with 50 writers, 5 handwritten pages per writer. The IAM dataset will also be used in our experiments described in Chapter 5 of the thesis.

(Schlapbach and Bunke 2004) use HMM-based handwriting recognizers (Marti and Bunke 2001) for writer identification and verification. The recognizers are specialized for a single writer by training using only handwriting originating from the chosen person. This method uses the output log-likelihood scores of the HMMs to identify the writer on separate text lines of variable content. Results of 96% identification with 2.5% error in verification are reported on a subset of the IAM database containing 100 writers, 5 handwritten pages per writer.

In (Bulacu et al. 2003), we proposed a texture-level approach using edge-based directional PDFs as features for text-independent writer identification. The joint PDF of “hinged” edge-angle combinations outperformed all the other evaluated features. Further improvements are obtained through incorporating also location information by extracting separate PDFs for the upper and lower halves of text lines and then adjoining the feature vectors (Bulacu and Schomaker 2003). Our allograph-level approach (Schomaker and Bulacu 2004, Schomaker et al. 2004) assumes that every writer acts as a stochastic generator of ink-blob shapes, or graphemes. The grapheme occurrence PDF is a discriminatory feature between different writers and it is computed on the basis of a common shape codebook obtained by clustering (Bulacu and Schomaker 2005a). An independent confirmation of our early experimental results is given in (van der Maaten and Postma 2005). In this thesis, we collect our published work in a coherent overall scene. We provide full details regarding our features, together with their extensive experimental evaluation. We also provide a comprehensive analysis of feature combi-
nations. On a large dataset containing 900 writers with 2 samples per writer, our best performing feature combinations yield writer identification rates of Top-1 85-87% and Top-10 96% with an error rate around 3% in verification.

An interactive approach involving character retracing and DTW matching is proposed in (van Erp et al. 2003). A layered architecture for forensic handwriting analysis systems is proposed in (Franke and Köppen 2001). The relevance of biometrics in the area of document analysis and recognition is discussed in (Fairhurst 2003).

From the studies reviewed in this section, two main conclusions can be drawn. Firstly, in the text-dependent approach, high performance is achievable even with very small amounts of available handwritten material (in the order of a few words). However, serious drawbacks are the limited applicability due to the assumption of a fixed text or the need for human intervention in localizing the objects of interest. The text-independent approach involves less human work and has broader applicability, but it requires larger amounts of handwriting in order to derive stable statistical features. Secondly, training writer-specific parametric models leads to significant improvements in performance, under the assumption, however, that sufficiently large amounts of handwriting are available for every writer.

The current thesis proposes text-independent methods for writer identification and verification. Our approach is sparse-parametric, it involves minimal training and the testing conditions are relevant to the forensic application domain. In our experimental datasets there are only two samples per writer containing usually an amount of handwriting in the order of one paragraph of text.

1.10 Main assumptions underlying the methods proposed in the thesis

There are three fundamental assumptions at the basis of the research work reported in this thesis. We make them explicit here.

- **Natural writing attitude**: our proposed statistical features capture the general distinctive aspects of the scanned script as it visually appears in the dataset samples. These global features can only be linked to a certain person under the assumption that the collected handwriting is genuine and no attempt has been made by the writer to disguise or forge his / her natural writing.

- **Foreground / background separation**: the input to the feature extraction algorithms described in this thesis are images containing handwritten text blocks with (near-) black ink on white background. We assume that the handwritten trace was separated from the document background and from all other graphical material that
may be present in the scanned images. It was possible to perform this separation automatically on the documents contained in our experimental datasets. In more concrete situations, this may actually require some limited form of human involvement.

- **Sufficient amount of ink:** in order to derive stable and text-independent estimates for the probability distributions used as writer identification features, a sufficient amount of handwritten material, in the order of a few text lines, must be present in the samples. The majority of the samples used in our experiments contain more than three handwritten text lines.

Under these assumptions, the general computer vision task of identifying a person on the basis of scanned images of handwriting becomes a tractable pattern recognition problem (Duda et al. 2001, Jain et al. 2000). The present thesis describes novel and effective statistical methods to automatically solve this interesting biometric identification problem.

In our view, the three assumptions underlying our pattern recognition methods are reasonable and not too restrictive. To a first approximation, handwriting may be considered as a natural binary image. Therefore, the techniques proposed in this thesis have practical applicability. Nevertheless, further research may be directed at eliminating part (ideally all) of these assumptions.

### 1.11 Overview of the thesis

The writer identification and verification methods presented in this thesis operate at two levels of analysis: the texture level and the character-shape (allograph) level. The body of the thesis is therefore divided into two main parts treating these two important aspects.

Chapter 2 and Chapter 3 cover the texture level features. Chapter 4 and Chapter 5 present the allograph level features and the method used to combined multiple features for improving the final performance of our writer identification and verification system. The substance of this thesis resides in the design of new and effective statistical features. An important characteristic that distinguishes our approach is that the proposed features are text-independent: the handwriting is merely seen as a texture characterized by some directional probability distributions or as a simple stochastic shape-emission process characterized by a grapheme occurrence probability.

Chapter 2 introduces the idea of using the directionality of script as a fundamental source of information for writer identification. We show that using edge-angles, and especially edge-angle combinations, to build directional probability distributions is an
1.11. Overview of the thesis

effective way to capture individual handwriting style with good performance in the task of writer identification.

Chapter 3 shows that further improvements in performance are obtained by capturing, besides orientation, also location information in the computation of our joint directional probability distributions. A comparison in terms of writer identification performance is carried out between lowercase and uppercase handwriting.

Chapter 4 presents our allograph-level method for automatic writer identification and verification. This theoretically founded method assumes that each writer is characterized by a stable probability of occurrence of some simple ink-trace shapes. We use the term graphemes for these sub- or supra-allographic ink fragments resulting from a handwriting segmentation procedure. Three clustering algorithms are compared on the task of generating the common shape codebook needed for estimating the writer-specific grapheme occurrence probability.

Chapter 5 considers the problem of fusing multiple features for improving the combined performance. Algorithmic refinements are described for the directional texture-level features and the experiments are extended to larger datasets. Our largest test set contains 900 writers and it is comparable in size to the largest dataset used in writer identification studies until the present.

Chapter 6 summarizes the research results presented in this thesis, draws the final general conclusions and sketches the future research directions opened by the work reported here. In the closing appendix, we use an HTML-based visualization tool to shows some representative results generated by our software, named GRAWIS, an acronym from Groningen Automatic Writer Identification System.