Chapter 6

Concluding the thesis

It is your work in life that is the ultimate seduction.

Pablo Picasso

6.1 Summary and Contributions

There are two fundamental dogmas underpinning handwriting identification. Their clear-cut statements are as follows:

- No two people write exactly alike.
- No one person writes exactly the same way twice.

These two principles, albeit oversimplified and disputable, unequivocally highlight the two natural factors that are in direct conflict in the attempt to identify a person based on samples of handwriting: between-writer variation as opposed to within-writer variability.

Our goal in this thesis was to automate the process of writer identification using scanned images of handwriting and thereby to provide a computer analysis of handwriting individuality. In this endeavor, a third computational factor takes center stage: the design and use of appropriate representations, computable features capturing the writing style of a person from the scanned handwritten samples. The power of such a representation or feature relies in its ability to maximize the separation between different writers, while remaining stable over samples produced by the same writer. We present in this thesis novel and very effective features for automatic writer identification on the basis of scanned images of handwriting. The similarity in handwriting style between any two samples is computed by using appropriate distance measures between their corresponding feature vectors. Our features and writer classification operate in the general framework of statistical pattern recognition (Duda et al. 2001, Jain et al. 2000).

Two fundamental sources of information regarding the individuality of handwriting are exploited by our methods functioning at two levels of analysis. First, handwriting slant, curvature and roundness, as determined by habitual pen grip, are captured by
joint directional probability distributions operating at the texture level. Second, the personalized set of letter shapes, called allographs, that a writer has learned to use under educational, cultural and memetic influences is captured by a grapheme-emission probability distribution operating at the character level. Combining texture-level and allograph-level features provides a very intimate and comprehensive characterization of the individual handwriting style of a person. Our methods achieved very high writer identification and verification performance in extensive tests carried out using large datasets with handwriting samples collected from up to 900 subjects.

In our methods, writer individuality is robustly encoded using probability distribution functions extracted from handwritten text blocks. 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 samples. In our methods the computer is completely agnostic of the actual text written in the samples. 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 distribution. Our techniques have practical feasibility and hold the potential of concrete use in real applications.

Chapter 1 of the thesis introduces writer identification as a behavioral biometric modality and presents the fundamental genetic and cultural factors causing the individuality of handwriting. The task of writer identification is equivalent to answering the question: “Who wrote this sample?” A writer identification system performs a one-to-many search in a large database with samples of known authorship and returns a likely list of candidates containing the handwritings most similar to the questioned one. The hit list is further scrutinized by a human expert. The task of writer verification is equivalent to answering the question: “Were these two samples written by the same person?” A one-to-one comparison is performed and an automatic yes / no decision is taken. In the introductory chapter, a connection is also drawn between writer identification and the related, but much broader, field of handwriting recognition. In handwriting recognition, the variations between different handwritings must be eliminated to obtain invariance and generalization. In writer identification, on the contrary, these same variations must be enhanced to obtain writer specificity and discrimination. Further in Chapter 1, a survey of recent publications in the field makes clear the distinction between text-dependent versus text-independent approaches and provides the necessary context in which to place our own research work.

The thesis then shows the progression of our writer identification research from low level textural features to higher level allographic features. The thesis is divided into two main parts. Chapter 2 and Chapter 3 describe our texture-level approach. Chapter 4 and Chapter 5 present our allograph-level approach and the fusion method used to combine
textural and allographic features for improved writer identification performance.

Chapter 2 shows that using the orientation of short fragments of edges along the written trace provides the basis for building several directional probability distributions that are very effective features for writer identification. The first angular feature constructed using oriented edge fragments is the edge-direction distribution, a classically known descriptor for writer identification. The mode of this distribution, i.e. the dominant direction in the script, corresponds to the slant of handwriting, which is a stable personal trait and a discriminatory characteristic between different writers. We propose further a new and potent method that considers the angle combinations of two ”hinged” edge fragments and builds a joint directional probability distribution that simultaneously encodes both orientation and curvature information. This novel ”edge-hinge” feature is a bivariate probability function that delivers a very significant improvement in writer identification and performance over the simple edge-angle distribution. The edge-based directional distributions, as a group of related features, outperform a number of non-angular features (run-length distributions, autocorrelation, entropy). Reducing the amount of ink in the test samples leads to an overall decrease in performance for all features, but the performance standings of the different features with respect to each other remain the same.

Chapter 3 carries on the idea of using the directionality of the script as an effective source of information for text-independent writer identification. And another new and strong feature is designed that considers the edge-angle combinations co-occurring at the extremities of run-lengths. Further performance improvements are obtained by incorporating also location information into the basic features. This is achieved by extracting two probability distributions separately from the top and bottom halves of text lines and then adjoining the two feature vectors. The asymmetry between to top and bottom distributions provides extra information regarding writer identity. The experimental study is performed as a comparison between lowercase and uppercase handwriting on test samples containing controlled amounts of ink. We obtain similar writer identification performance for lowercase and uppercase handwriting for the battery of features considered in the analysis.

Chapter 4 introduces our allograph-level method for writer identification and verification. This theoretically founded approach assumes that each writer is characterized by the occurrence probability of elementary shapes from a common shape codebook. These elementary shapes, or graphemes, are obtained by applying a heuristic segmentation procedure on the written ink. The common shape codebook is generated by clustering the set of graphemes extracted from the handwritings of a sufficiently large number of writers, kept separate from those used in identification and verification tests. The graphemes resulting from handwriting segmentation may, but usually will not, overlap a complete character. This is a fundamental problem for handwriting recognition.
Nevertheless, the ensemble of these sub- or supra-allographic shapes is very descriptive about the identity of the writer who generated them. And therefore is very effective in writer identification. In large scale computational experiments, we compare three clustering algorithms used for generating the common grapheme codebook: k-means, Kohonen Self-Organizing Maps 1D and 2D. The results prove the robustness of the proposed allograph-level writer identification method: similar good performance is obtained for all three clustering algorithms over a large range of codebook sizes.

Chapter 5 performs an extensive analysis of feature combinations. It is natural to try to combine the proposed features for improving the performance and robustness of our writer identification and verification system: while not totally orthogonal, the different features do offer different points of view on a handwritten sample and operate at different levels of analysis and also at different scales. In our fusion scheme, the final unique distance between two handwritten samples is computed as the average of the distances due to the individual features participating in the combination. In this chapter, more efficient algorithms are proposed for computing the directional features using contours, rather than edges. The functioning of the considered features is also put in an overall Fourier perspective that better explains also their relative performance merits. The evaluation experiments are extended to bigger datasets. The largest dataset comprises 900 writers and is comparable in size to the largest dataset used in writer identification studies until the present. The experimental results, consistent across the different test datasets, show that fusing multiple features yields increased writer identification and verification performance. The best performing feature combinations fuse directional, grapheme and run-length information yielding, on the large dataset containing 900 subjects, writer identification rates of Top-1 85-87% and Top-10 96% with an error rate around 3% in verification.

Chapter 6 concludes the thesis and Appendix A presents an HTML-based visualization tool developed with the purpose of visually assessing our writer identification and verification system called GRAWIS, an acronym from Groningen Automatic Writer Identification System.

The present thesis analyzes in depth the algorithmic aspects of automatic writer identification and verification. The proposed text-independent methods have possible impact in forensic science: they allow the search in a large dataset with handwritten samples with the retrieval of only those documents that pictorially look similar to the query in terms of handwriting style. In this way, the hit list containing the likely candidates is reduced to a size than can be analyzed in detail by the forensic expert to finally establish the writer identity for the questioned document. Part of the texture-level methods described in this thesis have already been used in a concrete industrial setting. Nevertheless, the wider application beyond the realm of academic research of our writer identification and verification techniques still remains a challenge for the future.
6.2 Further research directions

Considered in the general context of biometrics, automatic writer identification and verification is presently a thriving research topic. It is also a very engaging one. Here we sketch a number of further research ideas.

The writer identification methods presented in this thesis require the separation of the ink from the background of the document (image binarization). They also require the separation of handwriting from other graphical objects that might be present also in the document (layout analysis). Our academic datasets did not exhibit the full range of problems that must be solved in a complete document analysis and recognition system. A more extended examination is therefore needed of the document processing steps preceding feature extraction for writer identification and verification.

It is important to observe that the full variability of a person’s handwriting (within-writer variability) is not completely exposed in our datasets. For example, the long term changes occurring over years in the handwriting of an individual would require longitudinal studies.

While the writer identification and verification techniques presented in this thesis make extensive use of probabilities, our approach is not manifestly Bayesian. Nevertheless, our methods can be cast into a Bayesian framework and a more extensive analysis along this line is needed. Regarding the adoption of a Bayesian approach to writer identification and verification in the forensic application domain, a word of caution must be said about using prior probabilities: a maximum likelihood (ML) solution that ignores priors and rests on the shape evidence alone might be preferable to a full maximum a posteriori (MAP) solution. Weighting the shape evidence results with priors is considered to take place outside the scope of the current research. In this thesis, within the context of feature combinations presented in Chapter 5, we discussed the underlying Bayesian feature fusion model. Throughout most of our work however, we used vectorial representations and distances, rather than probability multiplication. During our research, the methods were developed using explanations and encodings that were close to the actual physical meaning of the features and the intuitive interpretation of the information they convey about the specific handwriting style of an individual (Bulacu and Schomaker 2007).

The experimental studies presented in this thesis were performed on Western script. Considering that our techniques are fairly generic and text-independent, their applicability to other scripts, e.g. Chinese, Arabic, Indian, is a pertinent and interesting research question.

In this thesis, we have designed and evaluated a number of writer identification features belonging to the category of fully automatic features computed from a region of interest (i.e. a handwritten text block) in the image. In forensic praxis, two other cat-
6. Concluding the thesis

egories of features derived from scanned samples are additionally used: *interactively measured features* by human experts using a dedicated graphical user-interface tool and *character-based features* related to the allograph subset that is being generated by each writer and requiring human work to isolate and label individual handwritten characters. Further exploration is required of the text-dependent methods, applicable for samples containing very limited amounts of handwriting, where probability estimation becomes unreliable. A performance comparison between the automatic features and the features requiring human involvement is still a fundamental open problem. This represents, in fact, the main topic of a new project, called Trigraph and financed by NWO, that will continue our research in the area of writer identification and verification.

Automatic writer identification can be applied to historical documents (Schomaker et al. 2007). While recognizing the actual text content of the documents is clearly more worthwhile, the identification of the writer can nevertheless have a degree of relevance in historical studies of paleography and codicology. Our methods are equally applicable to handwriting and machine print, writer identification versus font identification. Automatic script identification on historical documents would, in principle, open a number of interesting possibilities. It would allow to identify the scribe in case of handwritten documents or identify the printing house in case of machine printed documents. This would allow for automatic manuscript dating and/or authentication. Also, manuscript indexing and retrieval based on script style (graphical, rather than textual content) would become possible. Different types of calligraphy with their corresponding historical period could also be identified in a collection of documents. Because it provides some form of content enrichment, automatic script identification on historical documents might become a useful tool for the historian. This topic, placed at the confluence between computer science and humanities, can be a rewarding future research direction.