Beyond OCR: Handwritten manuscript attribute understanding
He, Sheng

IMPORTANT NOTE: You are advised to consult the publisher’s version (publisher’s PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher’s PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
He, S. (2017). Beyond OCR: Handwritten manuscript attribute understanding [Groningen]: University of Groningen

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Summary

Three fundamental problems have been studied in this thesis for handwritten document understanding based on handwriting style analysis: Writer identification, historical document dating and geographical localization, characterized by three questions: Who wrote it, and when and where was it written? Features of handwritten patterns which characterize writing style take a very important role in handwritten document understanding. Therefore, this thesis focuses on how to design discriminative and powerful features for different problems or applications.

Problems of writer identification based on rotated images and cross-script handwriting identification have been investigated in Chapter 2, Chapter 3, and Chapter 4. For historical document dating and geographical localization, three mid-level descriptors have been introduced to represent the handwriting style of historical documents. The contour fragments and stroke fragments are studied in Chapter 5 and Histogram of Orientations of Handwritten Strokes (H2OS) is studied in Chapter 6, as well as the Multi-Label Self-Organizing Map (MLSOM) clustering method. A comprehensive study of proposed features on writer identification, historical document dating and localization is presented in Chapter 7.

Chapter 2 provides a new rotation-invariant $\Delta n$Hinge feature for writer identification, which is an extension of the Hinge feature. When computing the angle distribution along the writing contours of handwritten texts, several points are involved to compute the relative angles, instead of the absolute ones, to construct the final feature vector, which works similar to the derivative of pen coordinates in on-line handwriting.

Chapter 3 proposes two curvature-free features for writer identification: run-lengths of local binary pattern (LBPruns) and cloud of line distribution (COLD) features. These features are based on the fact that the joint feature distribution of two properties can improve the performance because the joint distribution makes the feature relationships explicit instead of hoping that a trained classifier picks up a non-linear relation present in the data. The LBPruns feature is an extension of the LBP and run-lengths methods, which computes the run-lengths of the LBP codes, instead of the simple ‘0/1’ pattern. Therefore, it can capture the spatial neighboring relationship between the simple ‘0/1’ pattern over the neighbor scanning lines. The COLD feature is the joint distribution of orientation and length of line segments obtained by approximating writing contours using a polygon estimation method. These two proposed curvature-free features work very good on the proposed irregular-stroke handwritten CERUG data set.

Chapter 4 introduces a new mid-level feature based on the fact that the junction regions in handwritten texts are informative elements in visual patterns and characters. Junctions are prevalent in different scripts and also in both historical and modern handwritten documents and it is important to detect these informative junctions. Given the candidate junction points, which are the fork points as
well as the high curvature points on the skeleton lines, the junction strength is defined in every direction by the stroke length from the center point to the boundary points. This method is very easy to implement, independent on character or word segmentation, and can detect junctions in any kind of handwritten manuscripts. The procedure of junction detection yields a junction feature in a natural manner when considering the normalized stroke length in every direction as the feature vector. Our basic assumption is that junction regions are different when generated by different writers. For example, the number or directions of junction branches are different from different writers. Therefore, we consider the detected junctions as the graphemes and apply it for writer identification based on a codebook trained by a cluster method.

Chapter 5 provides the method of historical manuscript dating using a family of local contour fragments ($k$CF) and stroke fragments ($k$SF) based on the MPS data set. Contour and stroke fragments can be considered as the basic graphemes which encapsulate the handwriting style of historical manuscripts. $k$CF are formed by a number of $k$ primary contour fragments and $k$SF are formed by segments of length $k$ of a stroke fragment graph. The classical bag of words model is used to compute the feature representation of historical documents. They are described by the scale- and rotation-invariant descriptors and different codebooks are trained with different $k$.

The historical document dating problem is considered as a typical classification problem. When dating by general handwriting style identification, we get the conclusion that features which achieve a good performance on writer identification are not necessarily suitable for historical document dating via writer identification when there exists no sample for a target writer in the training set. When dating by classification, the experiments show that the combination of the contour and stroke fragments with multiple scale gives the optimal results.

In Chapter 6 introduces the new scale-invariant Histogram of Orientations of Handwritten Stroke ($H_2$OS) descriptor, which is a gradient-based feature and is very useful to describe the primary visual elements in handwritten document images. Experimental results show that on historical documents with a good quality (with less noise and easy to binarize), the junction feature and stroke fragments give good performance. However, on historical documents with low quality (contain degradations or noise and hard to binarize), the gradient-based descriptor provides most stable results.

In order to perform the dating and localization of historical documents, a Multi-Label Self-Organizing Map (MLSOM) is trained to discover the correlations between the low-level visual elements and their multiple labels. The MLSOM can be used to predict labels directly because it contains labels and it can also be used to train the codebook which contains more subtle information related to labels. The experiments on the MPS data set show that using the multi-label guided clustering to train the codebook provides much better results for both dating and localization.

Chapter 7 presents the pervasive common theme in this thesis: The joint-feature distribution principle (JFD) to design more powerful and discriminative features based on existing textural features. It consists three sub-principles: the spatial joint-feature distribution (JFD-S), the attribute joint-feature distribution (JFD-A) and the joint-kernel distribution principle (JFD-K). Recursively using these three principles with proper local features and kernel functions can generate new and more abstract features that might have specific meanings. A comprehensive study of the existing and proposed features are evaluated for historical document understanding, including writer and script identification, historical document dating and localization.

Three novel features are given based on the Hinge feature [Bulacu and Schomaker, 2007] for writer identification: Co-occurrence Hinge (CoHinge), Quadruple Hinge (QuadHinge) and $\Delta^2$Hinge, follow-
ing the JFD-S, JFD-A and JFD-K principles, respectively. CoHinge is the spatial joint Hinge kernel on different positions and QuadHinge is the attribute joint Hinge kernel with curvature information. The COLD feature is also presented in this chapter, which is the joint distribution of the relation between orientations and lengths of a set of line segments from the ink trace contours. Experimental results on five benchmark datasets for writer identification and writer retrieval show that the CoHinge and QuadHinge provide much better results than the original Hinge feature. \( \Delta^n \)Hinge follows the JFD-K principle and uses the differential operator kernel between two different Hinge kernels on different positions, and it is rotation invariant. In addition, the \( \Delta^n \)Hinge and COLD are less sensitive to the stroke length and they give the best performance on English handwriting written by Chinese people with long stroke lines.

The studies of this thesis show that designing hand-crafted features is not merely an ad hoc approach: powerful features can be constructed and used following certain principles, such as the proposed joint-feature distribution principle. The proposed features describe the handwriting style of manuscripts in different aspects, such as curvature or structure information of the handwritten strokes. They have possible impact in forensic science or digital humanities: they can be used to search documents by the similarity of their handwriting styles and thus can be used for not only writer identification, but also dating and geographical localization. Since most machine learning methods are "black boxes" for end users, the communication of the evidence and the result is very important. Our methods, such as junction feature, contour and stroke fragments, are very easy to be visualized and used for interface design. Therefore, our methods can produce the computational results that end users, such as paleographers or historians, may understand.