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):

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).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

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.
Chapter 1

Introduction

There are two types of information contained in handwritten document images: explicit information, such as characters, words, scripts or text lines, which can be directly read from document images, and implicit information, such as writer, date and geographical location, which can be obtained by analyzing detailed geometric characteristics. An example is shown in Fig. 1.1. Inferring both the explicit and implicit information is the problem of handwritten manuscript understanding, which is a fundamental research problem of a much larger scope than optical character recognition (OCR) alone, addressed by many researchers from different disciplines.

Traditionally, recognizing the explicit information is a optical character recognition problem, which converts images of characters to machine-encoded text for fast research or retrieval. For scholars, it would be interesting if handwritten manuscripts could be processed by OCR methods. However, automatic reading the text context by OCR is not enough to completely understand handwritten manuscripts. Apart from the actual content of the text, the writing style of handwritten characters also contains a lot of additional and useful information, such as the writer’s or script style which is characteristic for the time (date) of document production and reveal the historical context of manuscripts. Automatically extracting this information is very important for historians and paleographers (Stokes, 2015).

In pattern recognition, this process is called feature extraction. Geometric shape features are computed from the scanned images. These features can be very crude, such as raw pixel

<table>
<thead>
<tr>
<th>Explicit information</th>
<th>Implicit information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters</td>
<td>Writer</td>
</tr>
<tr>
<td>Words</td>
<td>Date</td>
</tr>
<tr>
<td>...</td>
<td>Geographical location</td>
</tr>
</tbody>
</table>

Figure 1.1: Illumination of information contained in handwritten document images.
Binary attributes:
- Words contain character 'a'? Yes
- Words contain 5 letters? No
- Printed words? No
- English letters? Yes
- From the same writer? No
...

Relative/Ranking attributes:
- Stroke width: (d) >(c) >(a) >(b)
- Curvature writing: (a) >(b) >(c) ≥ (d)
- Easy to segment: (b) >(a) ≥ (d) ≥ (c)
...

Abstract attributes:
- Who? the writer
- Where? the geographical location
- When? year/time

Figure 1.2: An example of attributes of handwritten words.

Intensities, but also can be advanced geometric structures, such as the Gabor feature, or Zernike moments. Using powerful features can yield very good classification performance under conditions of sparsely labeled data, not requiring complicated model estimations in machine learning. However, classification performance alone is often not enough. There is also a need for methods that are (1) explainable to the user; (2) that allow to build upon available knowledge, given new pattern classes; and (3) the essential information in features is not based on their isolated values, but also on the joint occurrence of feature values. We will now focus on issues (1) and (2), handling a more abstract concept than ‘features’, i.e., the notion of ‘attribute’.

Attribute learning is becoming a hot topic in computer vision and pattern recognition. As mentioned in [Russakovsky and Fei-Fei 2010], the term “attribute” is defined as “an inherent characteristic” of an object (as defined in Webster’s dictionary). More precisely, attributes are linguistically related descriptors of objects with high-level semantically meaningful properties. Generally, attributes can be divided into three categories: binary attributes, relative or ranking attributes and abstract attributes. Fig. 1.2 shows an example of different attributes of handwritten words. The binary attribute is the property that whether a certain object presents or not. The relative or ranking attribute indicates the strength of a property in an object with respect to other object [Parikh and Grauman 2011] and the abstract attribute is the property that describes the property of objects in a high-level, which could not be obtained directly from the object.
1.1 How to identify writers?

The difference between “attribute” and “feature” in this thesis is that feature is the basic description of properties presented in images, such as texture, color, edges or other structure information, while attribute is the semantic description of properties related to images. Features are usually extracted directly from images while attributes are always learned from data sets, based on a feature presentation.

In this thesis, we consider the writer, date and geographical localization as attributes of handwritten documents. Handwriting can be used as human behavioral biometrics measure (Bulacu and Schomaker, 2007) as the individual handwriting style is encoded into handwritten patterns when they were written down. This allows for the analysis of the handwriting style of manuscripts based on handwritten texts using pattern recognition techniques to unlock the important context information and attributes.

1.1 How to identify writers?

The author of handwriting is an important attribute and recognizing the author is corresponding to writer identification, which is the problem of automatic recognizing the author of a given piece handwritten images and answers the question: “who wrote it or which handwriting style did the author use?” Fig. 1.3 gives an example of the writer identification system, which recognizes the author by analyzing the handwriting style of the given piece of handwriting and comparing it with handwriting styles of handwriting samples of known authorship in a database. The basic assumption is that handwriting styles of handwriting from the same individual are consistent and handwriting styles of handwriting from different writers are distant. People have the character prototype in their brain when they start
Table 1.1: The advantages and disadvantages of text-dependent and text-independent methods for writer identification.

<table>
<thead>
<tr>
<th>Advantage</th>
<th>text-dependent (allograph-based)</th>
<th>text-independent (texture feature)</th>
</tr>
</thead>
<tbody>
<tr>
<td>easy to visualize</td>
<td>explainable for end users</td>
<td>easy to compute more efficient</td>
</tr>
<tr>
<td>need segmentation</td>
<td>characters should be present in training and testing set</td>
<td>users need to know probability and distance function</td>
</tr>
</tbody>
</table>

1. Introduction

To write (Teulings et al., 1986). Therefore, their writing styles of handwriting are relatively stable. The difference of handwriting style between different individuals are from many factors, such as the received education and familiar with the script. More factors can be found in (Huber and Headrick, 1999; Morris and Morris, 2000). These differences can be reflected on their handwriting. The main challenging of writer identification is to design a system and use pattern recognition to eliminate the differences of handwriting from the same writer and highlight the differences between different writers.

Approaches to writer identification can be coarsely divided into text-dependent and text-independent groups, according to the criteria whether the method recognizes the individual writing style based on certain characters or words (text-dependent) or features extracted from the entire image regardless of the semantic content (text-independent). Table 1.1 shows the advantages and disadvantages of the text-dependent and text-independent methods. The text-dependent approaches are limited due to the facts that it requires text segmentation and recognition prior to writer recognition, and the examined characters, such as ‘d’, ‘y’ and ‘f’ in (Pervouchine and Leedham, 2007), should be present in the writing samples to be compared. In addition, those methods are unable to seize the writing styles across different characters. Therefore, many automated writer identification methods fall into the text-independent category, in which statistical features are extracted from the entire image of a text block, and the similarity between two pieces of text is obtained based on those extracted features.

The features used in text-independent approaches have typically been categorized into two classes: statistical features and codebook-based features. Several widely used statistical features have been proposed in the last two decades. In (Bulacu and Schomaker, 2003), the edge-based directional probability distribution and the joint probability distribution of the angle combination of two “hinged” edge fragments are proposed for writer identification, which is termed as the “edge-Hinge” feature. This method has been extended to the contour-Hinge probability distribution (Bulacu and Schomaker, 2007) which computes a Hinge kernel on the contours of texts and Quill-Hinge (Brink et al., 2012) which combines the ink width with the contour-Hinge feature. Some methods use a filtering approach to extract features from text blocks, such as Gabor filtering (Said et al., 2000), Shababi and Rahmati (2009), XGabor filter (Helli and Moghaddam, 2010) and oriented Basic Image Features (oBIF) (Newell and Griffin, 2014). Chain codes and polygon based features on contours have also been used for
1.2. How to estimate date and geographical location?

The task of dating and localizing Medieval manuscripts is of the utmost importance to scholars of various disciplines studying the Middle Ages. Manuscripts that do not carry a date or location make it hard to assess their reliability as a historical source. However, this task is often regarded as the prerogative of a mere handful of specialists capable of correctly evaluating certain handwriting characteristics, but nevertheless sometimes conflicting conclusions. Usually, the dating or localizing of an instance of medieval script is based on the individual non-verbal intuition of the expert rather than on objective criteria. This state of affairs is not surprising, because there is a notorious lack of suitable collections of dated manuscripts that can be used as reference corpus. As the archaeologist has the $^{14}C$ technique to date organic materials, so the medievalist needs a method of dating manuscripts. The reliability of $^{14}C$ method is limited, however, when applied to medieval documents or manuscripts, and is, moreover, destructive because it requires physical samples.

The underlying assumption for historical document dating is that writing styles changed gradually, continuously and in general within a relatively limited time frame (within 25 years) in the ancient time. The rationale behind the assumption of a gradual style evolution comes from the observation that scribes were strictly and formally trained by experienced,
1. Introduction

Figure 1.4: An illustration of the development of the character ‘p’ from the ages 1300 to ages 1525. (Note: top-left is from ages 1300, bottom right is from ages 1525, in reading order.)

Figure 1.5: Illumination of the historical document dating problem. Given the query piece of historical document without date information (labeled as “year-?” in this figure), the task is to find the year when it had been written according a reference database with samples of known date or year information.

older teachers. As an example, Fig. 1.4 shows the writing styles of character ‘p’, as it was written in different ways in the period from 1300 to 1525 in the Dutch language area. If one wants to avoid the individual character segmentation and recognition, the question is whether the style evolution in the individual allograph is also reflected in overall page texture features.

Scribes wrote historical documents as a career, who usually lived in a local region for a long time. Therefore, the writing styles of historical manuscripts are quite stable in one city and are different between different cities. This allows us to localize historical documents by their writing styles encoded in the text.

Given an query manuscript without date or location, one possible way to estimate its year or location of origin is to search for similar writing styles in a large reference database consisting of dated documents, or to extract the general trend of writing styles in a certain period from the same database. Fig. 1.5 and Fig. 1.6 show the problems of historical manuscript dating and localization. A dating system such as this should, in other words, contain several steps: 1), a reference database which contains Medieval manuscripts or documents with year label is assembled; 2), several features are used to measure the similarity of writing styles in those documents; 3), machine learning methods are applied to perform the fine-tuned
1.2. How to estimate date and geographical location?

Figure 1.6: Illumination of the historical document geographical localization problem. Given the query piece of historical document without location information (labeled as “city-?” in this figure), the task is to find the year where it had been written according a reference database with samples of known geographical information. The map in the figure is the Dutch language regions in the world.

The differences between writer identification and historical document dating and localization are that (1) the goal of writer identification is to describe the individual’s handwriting in each handwritten document, because it needs to identify the exactly author of a piece of handwriting. Therefore, the data set should contain the corresponding writing samples from the same writer with the query document. In addition, the feature used for writer identification should be sensitive to the differences between different writers; (2) historical document dating and localizing aims to model the general handwriting style in a period or in a local region, among different scribes. The data set does not need to contain the writing samples exactly from the same writer as the query document, but only writing samples from the same period. Moreover, features used for dating and localization should be less sensitive to the individual’s writing style from the same period and discriminative to the individual’s writing style from different periods; (3) we have found that features which achieve a good performance on writer identification are not necessarily suitable for historical document dating and
localization.

The historical document dating problem has been studied recently in (He et al., 2014; Wahlberg et al., 2015; R.Howe et al., 2015; Li et al., 2015). Our previous work in (He et al., 2014) used a combined global and local regression method based on the Hinge and Fraglets features to estimate the year of origin of historical documents from the MPS data set. A similar method was proposed in (Wahlberg et al., 2015) based on the “Svenskt diplomatariums huvudkartotek” collection, consisting of scanned images of charters from the medieval period kept in the Swedish national archive (but not necessary produced in Sweden). A method to date Syriac documents was proposed in (R.Howe et al., 2015), using inkball models on a collection of securely dated letter samples from the period between 500 and 1100 CE. In (Li et al., 2015) a method to infer the date of printed historical documents from their scanned page images was developed, using Convolutional Neural Networks (CNN) on a data set from the Google books corpus (Vincent, 2007).

1.3 Research questions

This thesis focuses on predicting three attributes of handwritten documents, corresponding to three problems: writer identification, historical manuscript dating and localization.

Textural-based feature is a popular method used in writer identification, because it can be extracted from the whole document and used directly to compute the dissimilarity between different documents without any reference codebook or dictionary. Although the existing writer-identification methods have achieved high accuracy based on carefully scanned documents, only few of them has been reported to be rotation-invariant. However, a small rotation angle can be easily introduced into the images of handwriting samples. In the real-world, poor scanning practices result in a small rotation angle, which may have a serious impact on the performance of writer identification system based on the rotation-variant features. This problem raises an important question:

Q1: How to design rotation-invariant features for writer identification?

In Chapter 2, the rotation-invariant $\Delta^nHinge$ is proposed based on the Hinge feature (Buciu and Schomaker, 2007). In fact, the $\Delta^nHinge$ feature is the extension of the Hinge, which uses the differential operator on several pixels on writing contours. The proposed $\Delta^nHinge$ feature is not only rotation-invariant, but also contains high order derivative information of writing contours and can be used directly to on-line handwriting.

Today, the number of bilingual people is increasing and they often write with not only one script, which requires writer identification in a multi-script environment. Cross-script writer identification is the problem that recognizing the writer of a given piece of handwriting with one script from the samples of the data set written with another different script (Djeddi et al., 2013). Based on above observation or discussion, new research questions are instigated of this thesis:
1.3. Research questions

Q2: How to design efficient texture and grapheme features for writer identification?

Q3: How to perform cross-script writer identification, such as between English and Chinese?

We discuss these questions in Chapter 3 and Chapter 4 by propose novel curvature-free textural-based features and a new junction feature. We have found that handwritten documents wrote by less skilled writers contain a large number of irregular-curvature (curvature-less) strokes. Therefore, two curvature-free features are proposed in Chapter 3 to handle the writer identification problem based on handwritten documents wrote by no-native speakers. Junction feature proposed in Chapter 4, which is the stroke length distribution on every directions around a reference point inside handwritten strokes, is very easy to be detected and described in handwritten documents and can be applied for writer identification cross Chinese and English.

In the previous chapters, we have proposed textural-based and grapheme-based features for writer identification and performance is quite good. However, facing the relatively new historical document dating and localization problems, the research question is:

Q4: How to design an efficient system to automatically date and localize historical manuscripts based on the handwriting style?

This question is addressed in Chapter 5 and Chapter 6. Textural-based features are powerful for writer identification. However, there is only one feature vector from the whole document and most shape or allograph information of the characters are missed. Therefore, in Chapter 5, two fragment descriptors are proposed based on contour and stroke fragments in multiple scale. The extracted contour or stroke fragments are described by rotation and scale invariant descriptors. Combining these two fragment descriptors together achieves very good performance for historical document dating. In Chapter 6, we propose a novel stroke descriptor, which is robust to historical document degradations and a novel multiple-label guided cluster method is proposed to align graphemes in date and location spaces. The proposed cluster method can be used to predict labels directly, such as the date or location. In addition, it can be used to train a codebook, which contains more discriminative information on date and geography.

In addition, many textural features for handwritten document analysis have been proposed in the literature and in this thesis, but there is no general rule to design new features or improve the performance of existed features. This observation instigates the following questions:

Q5: What is the general rule or principle to design new features or increase the discriminative of existed features?

These question is addressed in Chapter 7 by proposing a general joint-feature distribution principle (JFD), which can generate more powerful and discriminative features based on existed features. As mentioned in introduction section, the essential information in features
is not based on their isolated values, but also on joint occurrence of feature values. The proposed JFD principle contains three different groups: the spatial joint feature distribution which can generate new features based on co-occurrence of existed features on different positions; the attribute joint feature distribution which can generate new features based on co-occurrence of different features on the same position and the joint kernel feature distribution which applies a kernel function between features on different positions or different attributes to extract new features. For example, applying a rotate-invariant kernel function can result in rotate-invariant features.

1.4 Material

Although there are several data sets available for writer identification, such as Firemaker \cite{Schomaker2000} and IAM \cite{Marti2002}, both of them contain single script. In order to evaluate the performance for cross-script writer identification between English and Chinese, we collect a new data set, named Chinese-English database of the University of Groningen (CERUG for short). The CERUG data set contains handwritten documents collected from 105 Chinese subjects, predominantly students from China. Some of them live in China and the rest studies in the Netherlands. Every subject is required to write four different A4 pages, following the Firemaker data set. On page 1, they were asked to copy a text of two paragraphs in Chinese. On page 2, the subjects described certain topics they liked in their own words in Chinese. We term the subset containing those two pages as CERUG-CN, in which handwritten documents are written in Chinese. Page 3 contains English text copied from two paragraphs. We split this page into two sub pages, and each sub page contains one paragraph. This forms the subset termed as CERUG-EN. In page 4, the subjects were asked to copy some names of countries and cities both in English and Chinese in two paragraphs. We also split this page into two sub pages to form another subset, which is termed as CERUG-MIXED for short. Note that each sub page in CERUG-MIXED contains both English letters and Chinese characters. In all three subsets, there are two handwritten samples from each writer. All the documents were scanned at 300 dpi, 8 bits/pixel, gray-scale.

For historical document dating and geographical localization, we introduce the Medieval Paleographical Scale (MPS) data set. The MPS data set consists of images of charters produced between 1300 and 1550 CE in four cities in the Low Countries: Arnhem, Leiden, Leuven and Groningen. Geographically, these four cities can be regarded as a cross section of the Medieval Dutch language area, and the development of writing styles visible within this data set therefore as approximating the development of writing within this area in general. Fig. \ref{fig:example} shows examples of charters from different cities in the MPS data set.

As the evolution of writing is a rather slow process, not every year in the period under consideration (1300-1550 CE) needed to be taken into account. The charters were therefore collected according to a sampling interval method. “key years” were set at every quarter
1.4. Material

century such as 1300, 1325, 1350, ⋯, 1550. Only explicitly dated charters produced in these key years and within a period of five years before or after them that were determined to have been written in one of the four cities mentioned before were included. There are currently 2858 charter images in the MPS data set, grouped around 11 key years. Table 1.2 shows the numbers of documents over the key years and the four cities. The frequencies are the natural counts of appearance in archives which have an underlying (historical) cause.
1. Introduction

Figure 1.8: Examples of charters for different cities in the MPS data set.

Figure 1.9: The left figure shows four labeled characters (‘a’, ‘d’, ‘g’, ‘p’ from top to bottom) in different key years in our MPS data set and the right figure shows their models, defined as the average shapes of manually labeled characters in the Monk system [Van der Zant et al. 2008].
1.5 Organization of this thesis

This thesis deals with understanding of handwritten documents from two perspectives and can be divided into three main parts. Chapter 2, Chapter 3, and Chapter 4 cover the writer identification problem. Chapter 5 and Chapter 6 cover the historical document dating and localization problem. Chapter 7 provides a general feature designing principle and a comprehensive study about the proposed features for four different applications.

Chapter 2 introduces an extension of the Hinge feature, called $\Delta^n$Hinge feature, which is a rotation-invariant feature. The experimental results on two widely used benchmark data sets show that the proposed method is promising and comparable to state-of-the-art methods.

Chapter 3 shows two novel curvature-free features: LBPruns and COLD, for writer identification based on handwritten documents wrote by less-skilled writers. Run-length of local binary pattern (LBPruns) is the joint distribution of the traditional run-length and local binary pattern methods and cloud of line distribution (COLD) is the joint distribution of the relation between orientation and length of line segments obtained from writing contours. Experimental results on the CERUG data set show that the combination of the LBPruns and COLD features provides a significant improvement.
Chapter 4 provides a novel junction detector and descriptor based on the fact that junction regions of handwritten strokes are informative elements and contain handwriting styles of writers. The junction descriptor is the stroke length distribution in every direction around a reference point inside the ink and it does not rely on any segmentation. The performance of cross-script writer identification between Chinese and English on the CERUG data set indicates that junctions are important atomic elements to characterize the writing styles.

Chapter 5 presents a family of local contour fragments ($kCF$) and stroke fragments ($kSF$) features and applies them for historical manuscript dating based on the MPS data set. $kCF$ captures the contour curvature information and $kSF$ captures the stroke structure information and their combination provides better results.

Chapter 6 proposes a novel descriptor built on a scale-invariant log-polar space, called Histogram of Orientations of Handwritten Strokes (HOHS or $H^2OS$), to extract and describe the visual elements in historical documents. In order to predict multi-labels, such as the date and location of the historical manuscript, the Multi-Label Self-Organizing Map (MLSOM) is proposed to discover the correlations between the low-level visual elements described by $H^2OS$ and their labels. The proposed method is evaluated on the MPS data set for historical manuscript dating and localization.

Chapter 7 presents a joint feature distribution principle, which allows the researchers to generate more efficient features based on existing textural features. All features proposed in this thesis follow this principle. Seventeen features including twelve textural-based and five grapheme-based features are studied for four applications: writer and script identification, historical document dating and localization.

Chapter 8 concludes the thesis by presenting several discussions and the answers to the research questions. In addition, future directions are also provided by answering several questions.
Part I

Writer Identification