Pattern Recognition 58 (2016) 159–171

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Image-based historical manuscript dating using contour and stroke fragments

Sheng He [a,*], Petros Samara [b], Jan Burgers [c], Lambert Schomaker [a]

[a] Institute of Artificial Intelligence and Cognitive Engineering, University of Groningen, PO Box 407, 9700 AK Groningen, The Netherlands
[b] Department of History, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam, The Netherlands
[c] Huygens Instituut voor Nederlandse geschiedenis, PO Box 90754, 2509 LT, The Hague, The Netherlands

ARTICLE INFO

Article history:
Received 23 November 2015
Received in revised form
2 March 2016
Accepted 25 March 2016
Available online 7 April 2016

Keywords:
Historical manuscript dating
Writer identification
Contour fragment
Stroke fragment
Handwriting style

ABSTRACT

Historical manuscript dating has always been an important challenge for historians but since countless manuscripts have become digitally available recently, the pattern recognition community has started addressing the dating problem as well. In this paper, we present a family of local contour fragments (*k*CF) and stroke fragments (*k*SF) features and study their application to historical document dating. *k*CF are formed by a number of *k* primary contour fragments segmented from the connected component contours of handwritten texts and *k*SF are formed by a segment of length *k* of a stroke fragment graph. The *k*CF and *k*SF are described by scale and rotation invariant descriptors and encoded into trained codebooks inspired by classical bag of words model. We evaluate our methods on the Medieval Paleographical Scale (MPS) data set and perform dating by writer identification and classification. As far as dating by writer identification is concerned, we arrive at the conclusion that features which perform well for writer identification are not necessarily suitable for historical document dating. Experimental results of dating by classification demonstrate that a combination of *k*CF and *k*SF achieves optimal results, with a mean absolute error of 14.9 years when excluding writer duplicates in training and 7.9 years when including writer duplicates in training.

1. Introduction

Handwritten historical documents are the most important sources of information about the past, especially where the more distant past is concerned, before the wide spread dissemination of printing and semi-mechanical text production. Increasing numbers of such documents are currently being digitized and stored in the computer, as in the Monk system [1], which contains more than 100K scanned page images. Thanks to this development, pattern recognition techniques can now be applied to solve historical document problems, which has already been attempted at length in the case of writer identification [2–5], word spotting [6,7] and character recognition [8,9]. These methods aim to provide efficient tools for scholars in the humanities to discover informative patterns in large digital collections. The Monk system [1], providing a web-based search engine for characters and words annotation, recognition and retrieval, can serve as an example.

Historical manuscripts loose much of their usability as sources if they cannot be dated with some accuracy. However, the fact is that most of them, especially those from the Middle Ages, do not carry any explicit date information. Often the only way to date these manuscripts is by inferring the year or period of origin from the characteristics of the handwriting they contain. Traditionally, this type of historical document dating has been the prerogative of paleographical specialists, basing themselves on years of experience and the non-verbal intuition acquired from it, rather than on objective criteria. Manual script dating is not efficient, as paleographical expertise is comparatively rare and, moreover, it is no exception for experts to arrive at conflicting conclusions when dating the same manuscript. Therefore, automatic script dating offers great promise for countless scholars working with undated handwritten historical sources.

The main motivation of using the computer to date historical manuscripts is to exploit patterns of handwritten texts that correlate with temporal information. This problem is similar to the "visual dating" problem in computer vision, such as historical color image dating [10], estimating the date of historical cars [11] and human age estimation based on face images [12,13]. The aim of visual dating is to mine the visual patterns that are specific for a

* Corresponding author. Tel.: +31 50 363 7410.
E-mail addresses: heshengxgd@gmail.com (S. He), petros.samara@huygens.knaw.nl (P. Samara), Jan.burgers@huygens.knaw.nl (J. Burgers), L.Schomaker@ai.rug.nl (L. Schomaker).

http://dx.doi.org/10.1016/j.patcog.2016.03.032
0031-3203/

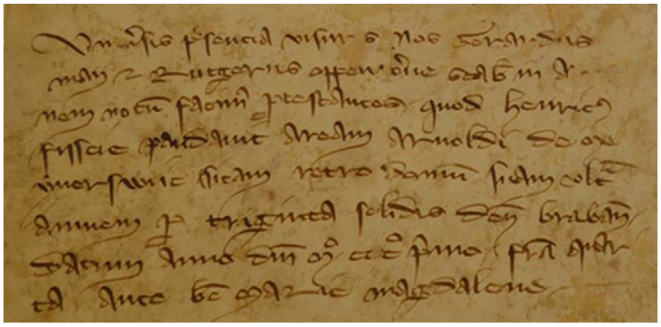

Fig. 1. An example of a charter in the MPS data set.

certain period in time [10] and to track and trace visual styles that change gradually over time [11].

We have proposed a number of features [2,14–16] to capture handwriting styles. However, there is one aspect of the visual appearance of handwritten samples that has not been addressed yet. In Fig. 1, a sample is shown. As we can see, the visual appearance is dominated by long curved stroke elements crossing other ink stroke traces in an irregular manner. Such a complicated thread structure was not covered by the junction feature [16,17] nor by other methods [2,14,15]. In addition, the existing methods concern low-level features, which cannot capture the properties of mid-level graphemes or stroke information. The research questions then are as follows: (1) How to define a feature that addresses the aspect of style at intermediate scale? (2) Which type of properties of handwritten strokes in historical documents contain the temporal information that can be used for dating? (3) What degree of feature complexity is required to obtain the optimal year estimation performance?

In this paper, we propose a family of local contour and stroke features and their application to historical document image dating. These features are small fragments of contours and strokes, called k contour fragments (kCF) and k stroke fragments (kSF), respectively. The fragments in kCF are the contour fragments resulting from a combination of a number of k consecutive primary fragments generated by the discrete contour evolution (DCE) [18] and the fragments in kSF form a segment of length k of a stroke fragment graph (SFG). The larger the number k of contour and stroke fragments in kCF and kSF, the more complex the contour and stroke fragment structures it can capture. We use the relative coordinates of the fragment points of kCF as the feature vector and use the polar stroke descriptor (PSD) proposed in [17] to describe the kSF.

The proposed kCF and kSF can be considered as grapheme-based representations and have several attractive properties: (1) kCF and kSF cover short contour and stroke fragments of the connected components in handwritten documents, which are probably shared between different characters and allographes. The statistical distribution of these small fragments can capture the handwriting style of historical documents. (2) For a certain range of k, both kCF and kSF can discover the meaningful and intermediate complexity patterns in a large connected component which may span several lines due to touching ascenders and descenders in cursive handwriting. (3) The descriptors of the kCF and kSF are insensitive to the scale and rotation of document images, which are very important properties in historical document analysis because historical documents are often digitized with different resolutions and font sizes in different documents are also different, making them sensitive to scale and rotation.

Inspired by the bag-of-words model [19], we construct codebooks of kCF and kSF with different complexity degrees k, each of which capture statistical information with different degrees of complexity of local fragments. All the kCF and kSF detected from handwritten images are mapped into the trained corresponding codebooks to form statistical histograms, the normalizations of which are the final representations of handwritten documents. We demonstrate the flexibility and power of kCF and kSF by applying them to historical document dating using the MPS data set [20].

We organize the rest of the paper as follows. Section 2 provides a review of related work on features used in writer identification and historical document dating. We introduce our MPS data set in Section 3. The details of the proposed kCF and kSF are outlined in Section 4 and Section 5, respectively. We evaluate the kCF and kSF on the MPS data set in Section 6. Finally, we conclude this paper in Section 7.

2. Related work

Various features have been proposed for handwritten document analysis in the previous studies. In this section, we first provide a brief review of the features used for writer identification. Previous studies on historical document dating are summarized in the second part.

2.1. Features used in writer identification

Features used in writer identification can be typically divided into two groups: textural-based and grapheme-based features. Textural-based features extract the texture, curvature or slant information from the entire document image, while grapheme-based features are the normalized histograms of individual graphemes based on trained codebooks, following the bag-of-words framework.

2.1.1. Textural-based features

Several types of textural-based features have been proposed in the literature, which can be roughly categorized into contour-based texture methods and filter-based texture methods.

The Hinge kernel on edges of the text can reflect the writing style [21] and the corresponding Hinge feature which is a distribution of the Hinge kernel on the entire document image has been used for writer identification in [14,15]. The Hinge feature has been extended to Δ^nHinge [22] to achieve the rotation-invariant property. In order to capture the width of ink traces, the Quill feature has been proposed in [2], which is a probability distribution of the relation between the ink direction and the ink width.

Spatial filtering techniques have been used to extract texture features from a handwritten text block. In [23], the Gabor filters and gray-scale co-occurrence matrices have been applied to writer identification. XGabor filters [24] which are obtained by modulating a centered sinusoid with a Gaussian have been used in Persian language writer identification. The oriented Basic Image Features (oBIFs) at two scales have been proposed in [25], using a bank of six Derivative-of-Gaussian filters.

2.1.2. Grapheme-based features

The COnnected-COmponent COntours (CO^3) has been proposed in [14] and applied to isolated uppercase handwritten documents with clear character segmentation. This was extended to lowercase handwriting in [15] by splitting cursive handwriting at the minima in the lower contour that are proximal to the upper contour, called Fraglets. Redundant small patterns of handwritten text were proposed in [26]. Recently, synthetic graphemes based on the beta-elliptic model were used for Arabic writer identification [27]. Singular structural regions in handwriting texts, such as junction regions, were extracted and a junction feature was