Computer Vision and Image Understanding 000 (2016) 1–9

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Historical manuscript dating based on temporal pattern codebook

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 18 November 2015
Revised 17 August 2016
Accepted 18 August 2016
Available online xxx

Keywords:
Image-based historical document dating
Junction feature
Temporal codebook learning

ABSTRACT

Manuscript dating is an essential part of historical scholarship. This paper proposes a framework for image-based historical manuscript dating based on handwritten pattern analysis in scanned historical manuscript images. We first use a singular structural feature to extract the mid-level handwritten patterns in historical document images and then encode the discovered handwritten patterns based on a codebook which contains the temporal information. We evaluate our method on the Medieval Paleographic Scale (MPS) data set and experimental results demonstrate that the feature representation based on the codebook which contains temporal information is more discriminative and powerful for dating. In addition, our proposed method can also visualize the evolution of handwritten patterns over time.

As a rule, it is virtually impossible to adequately interpret historical documents without knowing when they were written. However, many historical manuscript sources, especially those extant from the Middle Ages, do not carry a date. Dating these undated manuscripts often requires the expert knowledge of paleographers, who try to infer a date or period of origin from the nature of the script a manuscript contains. Usually, such a paleographical dating procedure is based on the scholar's experience, non-discursive intuition and frequently subjective considerations He et al. (2014). It is, moreover, no exception for paleographical experts to arrive at conflicting conclusions when dating the same manuscript. However, in the past decades many historical documents have been scanned and stored in computers to facilitate searching and retrieval, as in the Monk system Van der Zant et al. (2008), which opens up the possibility to automatically estimate the date of origin of historical documents using a large reference corpus.

Any reliable automatic dating method would be patently more efficient than a manually dating procedure. However, thus far such a method has not yet been well defined and addressed. The main challenges of historical document dating based on scanned digital images include: (1) how to extract and represent the handwritten patterns that correlate with temporal information? Temporal information is stored in the patterns of the characters in document images and it is very hard to capture the subtle differences between characters written in different eras. In addition, visualizing the discovered temporal patterns which reveal the date information is necessary for end users, such as paleographers and practicing historians. This requires a computational and visible mid-level descriptor in handwritten images. (2) How to represent handwritten document features suitable to address the dating problem? Many features have been developed to capture the handwriting style of handwritten documents for writer identification in the last decade. However, the existing feature representation methods in writer identification are particularly sensitive to the individual handwriting style, while solving the dating problem requires capturing the general handwriting style in specific time periods instead of the individual handwriting styles. The feature representation related to the dating problem should exhibit synchronic similarity between writers at the same time and diachronic dissimilarity.

In this paper, we propose a novel algorithm to cope with these two problems based on the Medieval Paleographic Scale (MPS) data set He et al. (2014), containing historical documents from the period 1300–1550 CE. First, we use a scale-invariant mid-level Polar Stroke Descriptor (PSD) proposed in He and Schomaker (2015) to extract and describe the meaningful handwritten patterns which contain handwriting style information. The PSD is an efficient and discriminative descriptor which does not depend on line or character segmentation. Furthermore, we apply the Self-Organizing Time Map (SOTM) method proposed in Sarlin (2013) to train a codebook which contains the temporal information of the handwritten patterns. The codebook trained by SOTM is called Temporal Pattern Codebook (TPC) in this paper. Finally, the representation of the input document is computed by the Bag-of-Words (BoW) model based on the trained temporal pattern codebook, and

* Corresponding author.
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.cviu.2016.08.008
1077-3142/

the historical document dating problem becomes a standard pattern recognition problem. The benefits of using the SOTM to train the codebook are that (1) it contains the temporal information; (2) the contourparts of the visual words in different times can be aligned and visualized.

The remainder of this paper is organized as follows. We review related research in Section 1 and introduce our MPS data set in Section 2. Then, we present the proposed method in Section 3, while our experimental results are presented in Section 4. Finally, we give a conclusion and define the tasks ahead in Section 5.

1. Related work

One of the main challenges for historical document dating is concerned with choosing appropriate handwritten document descriptors. In this section, we first give a brief review of feature representation of handwritten documents. Most of the existing features capture the writing style of handwritten texts and are used for writer identification. Secondly, some related aspects of the dating problem using pattern recognition and computer vision techniques are presented.

1.1. Feature representation

Features used for capturing the writing style in handwritten images can roughly be divided into two groups: textural-based and grapheme-based features. The textural-based features extract textural information of handwritten text over the entire image, and their performance is usually better than that of the grapheme-based features. However, the grapheme-based features are intuitive and easy to explain to users in other domains. Therefore, both of the two types of features have been developed in recent years. An edge-based directional distribution of the handwritten text is proposed in Schomaker and Bulacu (2004) and it has been extended to a contour-based directional distribution in Bulacu and Schomaker (2007). These features are termed “Hinge” features because they are the joint probability distribution of the angle combination of two “hinged” edge fragments. The Δ^nHinge He and Schomaker (2014) is an extension of the “Hinge” feature but has a rotation-invariant property. The Quill feature has been proposed in Brink et al. (2012), which combines the ink width with the ink direction to capture the properties of the writing instrument. Filter-based features are also used to extract textural information from handwritten text blocks, such as the Gabor filtering Said et al. (2000), XGabor Helli and Moghaddam (2010), oriented Basic Image Features (oBIF) Newell and Griffin (2014) and chain codes on the text contours Siddiqi and Vincent (2010).

The grapheme-based features extract handwritten patterns and encode these elements into a codebook, inspired by the BoW framework. Most existing research works focus on how to extract the meaningful handwritten patterns. The COnnected-COmponent COntours (CO^3) has been considered as the basic elements in the isolated uppercase western script in Schomaker and Bulacu (2004). This method has been extended to Fraglets in Bulacu and Schomaker (2007) in cursive handwriting. Small parts of handwritten text in Siddiqi and Vincent (2010) and the singular structural regions, such as junction regions in He et al. (2015), have also been considered as the basic graphemes in handwritten text. The synthetic graphemes based on the beta-elliptic model have been proposed in Abdi and Khemakhem (2015) for Arabic handwriting.

1.2. Dating problem

The dating problem revolves around determining the “age” of images and this has been studied with regard to different types of images in the last decade. Automatically estimating the age of historical color photographs has been proposed in Palermo et al. (2012) based on the temporally discriminative information extracted on color images over time. Discovering the object-specific style-sensitive mid-level features over time has been applied in Lee et al. (2013) to predict geographical or temporal provenance. Human age estimation based face images has been widely studied, such as Geng et al. (2007); Guo et al. (2008). Dating ancient inscriptions and Byzantine codices by means of writer identification has been proposed in Papaodysseus et al. (2014).

Historical document dating has recently received attention in the pattern recognition community. In Wahlberg et al. (2015), the authors used shape context descriptors on gray scale images to date medieval manuscripts from the collection “Svenskt diplomatariums huvudkartotek”. Estimation of the dates of origin of Syriac documents from the period 500–1100 CE has been proposed in Howe et al. (2015), based on the Inkball models of individual character samples. Our previous work in He et al. (2016,2014) estimated the year of origin of historical documents using a part of the MPS data set.

2. MPS data set

We evaluate our proposed method with the MPS data set, which was first proposed in He et al. (2014). The MPS data set contains documents from the period 1300–1550 CE, and includes only so called ‘charters’, a specific type of - explicitly dated - document widely used in the Middle Ages to prove a certain legal or economic action or transaction had taken place. These charters were collected from the archives of four cities: Arnhem, Leiden, Leuven and Groningen, representing so many ‘corners’ of the medieval Dutch language area. Moreover, all charters were established to have been written within one of these four cities. As the evolution of script is a fairly gradual process and fundamental changes in handwriting habits take a few decades to become generally noticeable, not every single year was taken into account. Instead, the MPS set only includes charters written around one of 11 ”key years“ (1300 ± 5, 1325 ± 5, 1350 ± 5, ... , 1525 ± 5, 1550 ± 5) with an interval of five years plus or minus the key year. Currently, there are 2858 documents in the MPS data set. Detailed information about the MPS data set can be found in Samara (2014). The data set and more details on the scanned charter images are available on the Monk system (http://application02.target.rug.nl/monk/Projects/MPS/).

The MPS data set shows a gradual evolution of handwriting styles in the period 1300–1550 CE. The slow rate of change can be explained by the fact that most charters in this data set were written by professional scribes, and the handwriting style of the younger scribes was often affected by the older ones when they worked together. On the other hand, the handwriting of scribes who were active for several decades and attested in the MPS set over a longer period, was usually not completely fixed, but changed somewhat according to the general trend. In addition, the number of different hands who produced the charters was particularly limited. While there are, currently, almost 3000 charters in the MPS data set, the number of hands that produced them is no more than one thousand.

The evolution of the handwriting styles is reflected in the handwritten patterns of the characters. Fig. 1 shows several instances of the character ‘a’ written over the 11 key years, which shows a clearly datable evolution - double ‘a’ being replaced by single ‘a’ from 1375 onwards. Other characters also contain clearly datable information, such as the characters ‘d’, ‘g’, and ‘p’ (see Fig. 2). The appearance of a certain instance of a character written in a specific period therefore contains date information Samara (2014) and the aim of this paper is to automatically discover this information.

Please cite this article as: S. He et al., Historical manuscript dating based on temporal pattern codebook, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.08.008

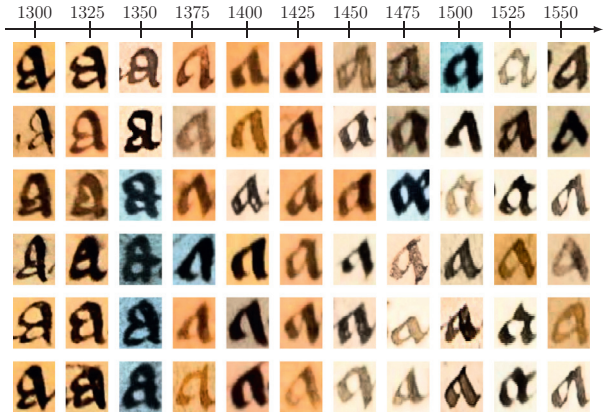


Fig. 1. Six randomly selected instances of the character 'a' in the 11 key years in our MPS data set.

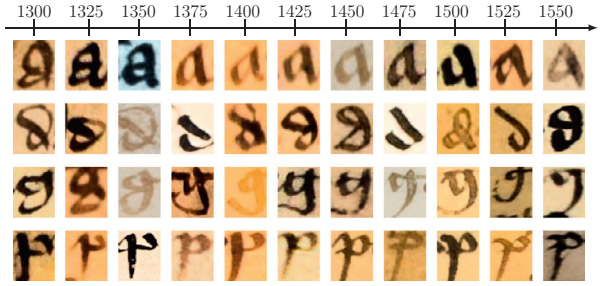


Fig. 2. Four labeled characters ('a','d','g','p' from top to bottom) in the 11 key years.

Historical document dating by means of analyzing the writing style of each character in the document is promising. For example, given a set of training characters which contain most discriminative datable information and the key year labels, we can find the corresponding characters in undated documents and predict the key year based on the similarity of the whole or parts of the characters. However, this procedure requires character segmentation and recognition, which pose challenging problems in itself. In this paper, we propose a compact document representation using a codebook which contains temporal information inspired by the BoW model. A single feature vector is obtained from each document, and the dating is performed on the page level instead of the character level.

3. Historical document representation

3.1. Handwritten pattern extraction

Although there are several grapheme-based methods in existence (mentioned in the related work section), most of them are not suitable for historical document analysis, except for the junction feature He et al. (2015). For example, the CO^3 Schomaker and Bulacu (2004) is designed for isolated characters and the Fraglets Bulacu and Schomaker (2007) requires line segmentation, which poses a very challenging problem in handwritten historical documents. The small parts of text blocks proposed in Siddiqi and Vincent (2010) are scale-sensitive and carry non-meaningful information, and the synthetic graphemes proposed in Abdi and Khemakhem (2015) are only suitable for Arabic handwritten text. Therefore, we use the Polar Stroke Descriptor (PSD) He and Schomaker (2015) to extract handwritten patterns, which is also called junction feature if the key points lie on the junction points He et al. (2015).

As a local descriptor, key points inside the ink stroke should be detected first based on the binarized images. In this paper, the fork points and high curvature points on the skeleton lines of the characters are selected as the key points, because the context of these points contains the structural information of the characters. Choice of the binarization and thinning methods is problem specific and depends on the degradation of the historical document images. Common choices include the Otsu Otsu (1975), Sauvola Sauvola and Pietikäinen (2000) and AdOtsu Moghaddam and Cheriet (2012) for binarization and the method in Zhang and Suen (1984) for thinning. Fig. 3(a) shows examples of high curvature point (green) or fork point (red) in the character 'a'. The high curvature points can be easily computed by the method in He et al. (2015).

Given a point p_i on the skeleton line, the angle on p_i is computed based on two directions ϕ_1 and ϕ_2 with two legs in the preceding and succeeding directions (see Fig. 4) as: $\phi_{2\pi}(p_i) = min(\|\phi_2 - \phi_1\|, 2\pi - \|\phi_2 - \phi_1\|)$. The local minimum of the angles on a skeleton line fragment is selected as the high curvature point. The parameter e in Fig. 4 is set to 7, following the work He et al. (2015).

Given a key point $\mathbf{p} = (x, y)$, the stroke length in the direction θ_m is obtained by searching the background pixel following the ray starting from the key point $\mathbf{p}$ towards the ink boundary in the direction θ_m. The search is stopped when the first background pixel (x_e, y_e) is found, and the stroke length $len(\theta_m)$ is the Euclidean distance between the key point and background pixel Brink et al. (2012):

$$len(\theta_m) = \sqrt{(x - x_e)^2 + (y - y_e)^2} \tag{1}$$

The feature vector is the normalized distribution of the stroke length in the possible directions $\theta_m = 2\pi m/N$:

$$F = \{f_0, f_1, \ldots, f_{N-1}\} \tag{2}$$

where $f_m = len(\theta_m)/\sum_{i=0}^{N-1} len(\theta_i)$, m is from 0 to $N-1$ and N is the number of directions and also the dimension of the feature vector, which is set to 120 in our experiments. The feature vector F can be rotational-invariant if the starting direction f_0 is selected as the maximum value $f_0 = \max_m\{f_m\}$. Fig. 3(b) and Fig. 3(c) show the stroke length distribution on the polar space and the linear coordinate space, respectively. Detailed information of the computation of the PSD can be found in He and Schomaker (2015); He et al. (2015).[1]

The handwritten patterns extracted and described by the PSD descriptor reflect the local geometrical and structural features around the singular and salient regions in the handwritten texts reflecting the handwriting styles. Fig. 5 gives the visual results of the top 12 similar patterns with each query pattern (first column of each row). From the figure we can find that the patterns detected by the PSD are meaningful and might be parts of different characters, which reflect that an individual draws the same or similar patterns the same way irrespective of characters. Another reason that the same patterns are shared between different characters or graphemes is that historical documents in the MPS data set were written by capillary-action writing instruments which produced some frequently occurring patterns resulting from a specific habitual tip angle and individual movement style Brink et al. (2012), and the statistical information of these frequently occurring patterns characterize the writing style of a given handwritten document.

[1] Source code of PSD is available by request.

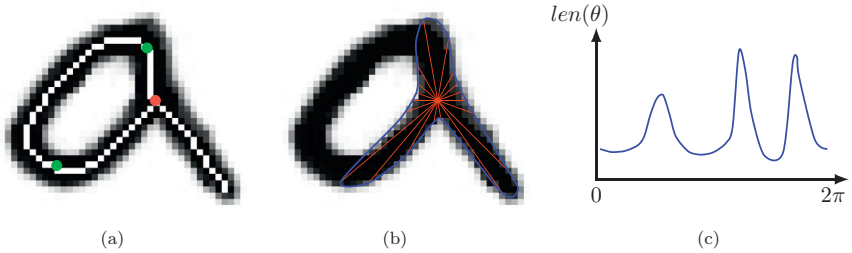


Fig. 3. Figure (a) shows the skeleton line detected in the handwritten character, in which the red point is the fork point and the green points are the high curvature points, which are the candidate points for the PSD descriptor. Figure (b) shows the stroke width distribution around the fork point. Figure (c) shows the normalized distribution in a linear coordinate, from 0 to 2π, which is the feature vector of the fork point.

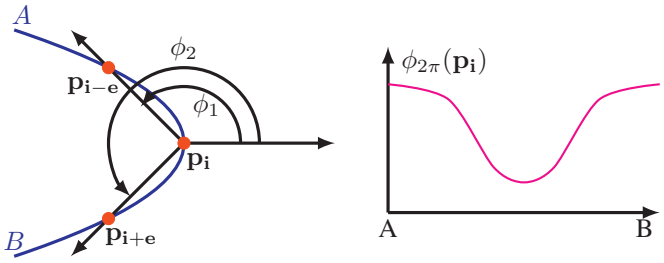


Fig. 4. Computing the tangent of a curve. Figure (a) shows how to compute the angle on point p_i and (b) shows the computed angle on the curve from A to B. (This figure is from the work He et al. (2015)).

3.2. Temporal pattern codebook (TPC)

In order to compute the similarity between documents written in different times, a common space (named codebook) is built, and the extracted patterns from handwritten images are encoded into this space to form the final representation.

Traditionally, unsupervised clustering methods, such as the k-means or the Self-Organizing Map (SOM) Kohonen (1998), are used to train the codebook. However, the unsupervised clustering methods lose the temporal information of the patterns and thus fail to capture the correlations between the patterns and their labels. In this paper, our aim is to capture the temporal information of the extracted handwritten patterns and discover their evolution in the period under consideration (1300–1550 CE). Therefore, we use a weakly supervised clustering method inspired by the Self-Organizing Time Map (SOTM) Sarlin (2013), which discovers the occurrence and explores the properties of temporal structural changes in data based on the Kohonen's standard self-organizing map Kohonen (1998).

We use $y(t)$ to denote the key year in this paper, where $y(t) = \{1300, 1325, \ldots, 1550\}$ with an ascending order of the index t. That is $y(1) = 1300, y(2) = 1325, \ldots, y(11) = 1550$. The learned codebook $D = \{D_1, D_2, \ldots, D_t, \ldots, D_{11}\}$ is composed by the sub-codebook D_t trained by the handwritten patterns $\Omega(t)$ from the key year $y(t)$. In order to capture the temporal changes of the visual words, the sub-codebooks are trained in an ascending order of key years, starting at $y(1) = 1300$. Since we do not have a priori knowledge of the handwritten patterns before key year 1300, we randomly initialize and train the sub-codebook D_1 with the input patterns $\Omega(1)$ from the key year $y(1)$ by the standard SOM method. The patterns of each cluster of the sub-codebook D_1 can be considered to represent a certain writing style, and our aim is to find their counterparts in the following key years in order to model their evolution in style over the 11 key years. Therefore, given an initial cluster in the year $y(1)$, the counterpart in the following key year $y(2)$ can be found by fine tuning the cluster using the patterns from the year $y(2)$. This can be done by the SOTM method, which works as follows. The sub-codebook D_t is initialized by the trained D_{t-1}, in order to retain information about past patterns and preserve the connection between consecutive patterns in the key year series, and is then fine tuned by the input patterns $\Omega(t)$ from the key year $y(t)$. The training procedure is summarized in Table 1 (more details of the training processing can be found in Sarlin (2013); Sarlin and Yao (2013)).

Table 1
The procedure of the SOTM method.

$t = 1$
randomly initialize D_t
train D_t using input patterns $\Omega(t)$ by the standard SOM
while $t <= 11$
$t = t + 1$
initialize D_t using D_{t-1}
train D_t using $\Omega(t)$ by the standard SOM
end
output $D = \{D_1, D_2, \ldots, D_t, \ldots, D_{11}\}$

We call the codebook D trained by the SOTM the Temporal Pattern Codebook (TPC). As mentioned before, the patterns of each

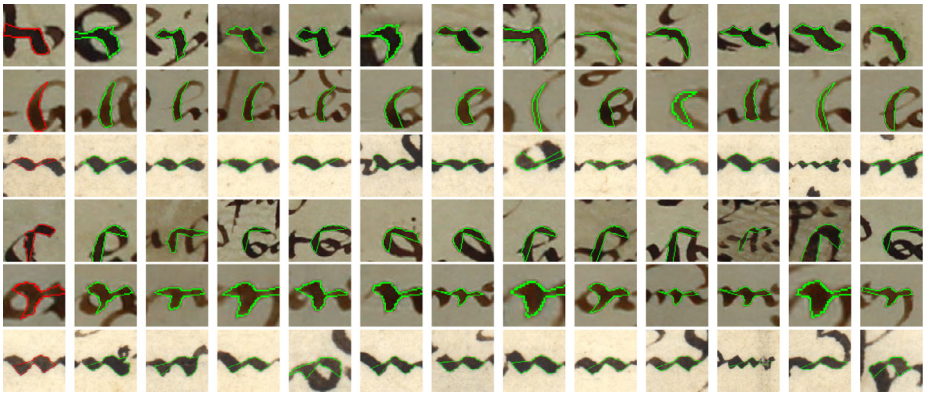
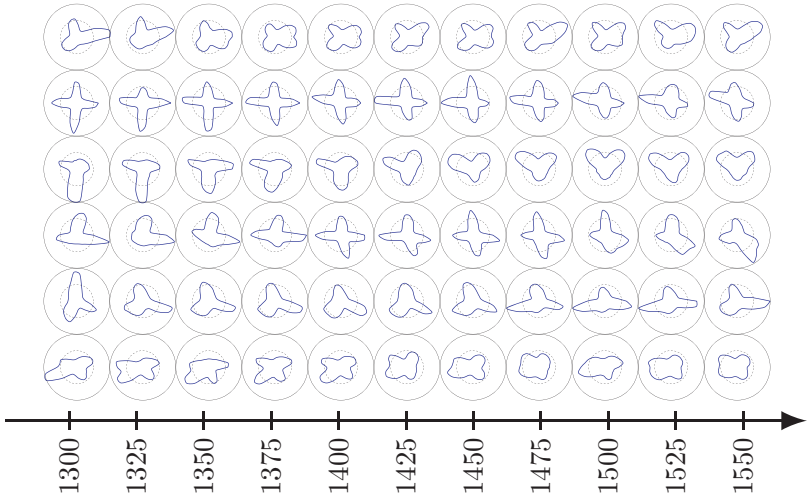
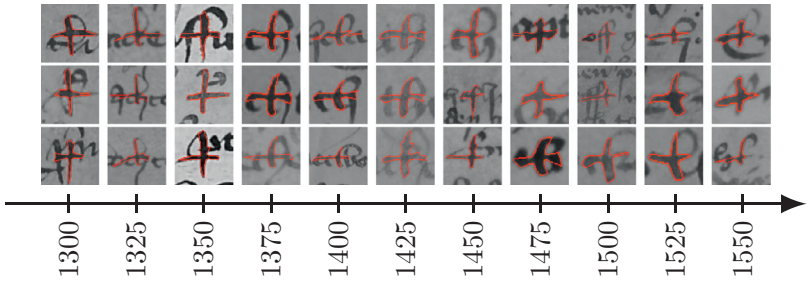


Fig. 5. Each row shows the first 12 instances in a hit list of the query patterns of the first column. Note that each patch is normalized into a fixed size.

Please cite this article as: S. He et al., Historical manuscript dating based on temporal pattern codebook, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.08.008



(a) Visual words of the TPC



(b) Handwritten patterns

Fig. 6. Fig. (a) shows an example of selected six visual words of the trained temporal pattern codebooks; Fig. (b) shows the corresponding instances of the visual words in the second row of Fig. (a). Given the initial patterns, our algorithm can find the corresponding patterns in the following key years, which shows how the patterns evolve in the period 1300–1550. It should be noted that in order to obtain optimal results, *all* codebook elements need to be included in the computations.

cluster of the sub-codebook D_1 can be considered to represent a certain writing style and their counterparts in the following key years can be maintained in the TPC. For example, Fig. 6(a) shows six visual words in the codebook in the 11 key years and the corresponding instances of the visual word of the second row are shown in Fig. 6(b). From the figure we can see that the same handwritten patterns can look different across the 11 key years, but such differences might be very subtle. However, such subtle differences containing datable information are lost in the traditional unsupervised codebook learning methods, but are retained in our TPC. Therefore, our temporal pattern codebooks are more discriminative. In addition, the similar counterparts in different time-periods in the TPC provide visual evidence to end users and allow us to model the changes in style of the handwritten patterns over the 11 key years.

All the handwritten patterns extracted from the given historical document are mapped to the codebook to build a histogram by searching the nearest neighbor using the χ^2 distance function. Finally, the normalized histogram is considered as the feature representation.

4. Experiments

4.1. Historical document dating

The dating problem can be considered as either a classification problem Palermo et al. (2012) or a regression problem Lee et al. (2013). Since the documents in the MPS data set have obvious borders between the nearby key years, we regard the historical document dating as a classification problem. There are 11 key years in our data set, corresponding to 11 classes spanning from 1295 to 1555 CE. We train 11 corresponding classifiers $w_i, i = 1, 2, \ldots, 11$ using a linear SVM with the one-versus-all strategy. The query document with the feature vector f is predicted by

$$\bar{y} = \arg\max_i \{w_i^T f\}. \quad (3)$$

The writers of some documents in the MPS data set are known while others are not. Therefore, we consider two different evaluation scenarios for historical document dating. In the first one, we carefully split the data set into training and test sub sets to make sure that the same writer never appears in both the training and test sets, which means that all documents from the same hand should be only in the training set or only in the test set. We call this scenario *excluding writer duplicates* (*wr.excl.* for short), which avoids the risk of dating by writer identification. In the second scenario, we randomly split the data set into training and test sets without considering the writer labels. We call this scenario *including writer duplicates* (*wr.incl.* for short), in which the system performs the dating based on the general handwriting style built by other writers. However, in the *wr.incl.* scenario, the processing of writer identification is probably involved in the dating. This is reasonable because if we know the writer, the date of the document can be directly obtained. In fact, most of the charters were written by scribes for whom writing was a profession and

Please cite this article as: S. He et al., Historical manuscript dating based on temporal pattern codebook, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.08.008

their working careers could cover several decades. If the writers of documents are identified correctly, then their career periods can be obtained and thus the documents can be dated.

We compare our proposed method with Hinge Bulacu and Schomaker (2007), and Quill Brink et al. (2012), which are textural-based features for writer identification. We choose these two features because in practice we have found that they are very efficient and their performance for historical document dating is better than others. The parameters of these features are set following their original papers.

4.2. Experimental settings

We employed two widely used measures for performance evaluation: the Mean Absolute Error (MAE) and Cumulative Score (CS). The MAE is a Manhattan-type distance, which is typically defined as:

$$MAE = \sum_{i=1}^{N} |\overline{K(y_i)} - K(y_i)|/N \quad (4)$$

where $K(y_i)$ is the ground-truth of the input document y_i and $\overline{K(y_i)}$ is the estimated key year, N is the number of test documents. The Cumulative Score(CS) is typically defined as Geng et al. (2007):

$$CS(\alpha) = N_{e \leq \alpha}/N \times 100\% \quad (5)$$

where $N_{e \leq \alpha}$ is the number of test images on which the key year estimation makes an absolute error e no higher than the acceptable error level: α years. For historians, an error less than or equal to ±25 years is, more often than not, acceptable when dating historical documents. Therefore, we report the Cumulative Score with error level $\alpha = 0$ and $\alpha = 25$ years in the experiments.

We split the data set into training (70%) and test (30%) sets. The experiment is repeated 20 times and the average results with the standard deviation are reported in the following experiments.

4.3. The effect of codebook size

In this section, we evaluate the performance of historical document dating using different sizes of the TPC. We use N_{sub} to denote the size of the sub-codebook of each key year and finally the size of the TPC is $N_{sub} \times 11$. The performance in terms of MAE with different sizes of the TPC is shown in Fig. 7. From the figure we can conclude that the performance when including writer duplicates is better than when excluding writer duplicates. Generally, the performance improves as the size of codebook increases. However, it gets saturated when sub-codebook size N_{sub} is larger than 500. In addition, a large size of the TPC needs a large memory and a long computational time. When $N_{sub} = 50$, the MAE in the *wr.excl.* scenario is 19.9 and it decreases to 15.1 when $N_{sub} = 650$. The same trend can be found in the *wr.incl.* scenario. In the following experiments, we report the best results with the size $N_{sub} = 650$. All experiments are performed on a cluster [2] and generally it takes less than 5 min with the codebook size $N_{sub} = 650$.

4.4. Performance when excluding writer duplicates

This section provides the experimental results in the *excluding writer duplicates* scenario. Table 2 shows the results in terms of the MAEs and Cumulative Score with $\alpha = 0$ and $\alpha = 25$. We can see that the best performance is achieved by the proposed method, with MAE of 15.1 years, which improves the performance of Quill and Hinge by 8.6 and 7.0 years, respectively. 65.3% of documents in the test set are correctly dated by the proposed method. Our proposed method also improves the scores when the error level is equal to 25 years. Fig. 8 shows the Cumulative Scores of the three methods. As can be seen, our proposed method achieves the best performance.

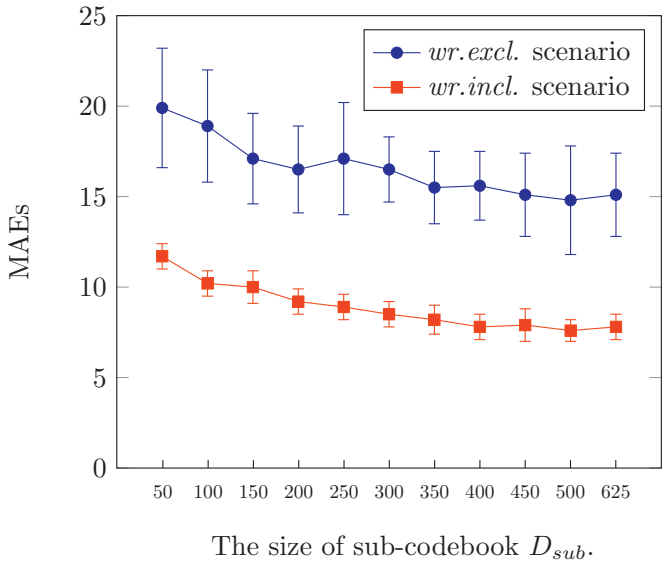


Fig. 7. The MAEs of the dating on the MPS data set using different sizes of the TPC codebook changing from 50 to 625.

Table 2
The dating performance of different methods when excluding writer duplicates.

Feature	MAEs	CS($\alpha = 0$)	CS($\alpha = 25$)
Quill Brink et al. (2012)	23.7 ± 2.9	50.2 ± 5.2%	80.6 ± 3.0%
Hinge Bulacu and Schomaker (2007)	22.1 ± 2.9	53.8 ± 5.9%	80.6 ± 3.1%
Proposed	**15.1** ± 2.3	**65.3** ± 5.9%	**87.4** ± 2.7%

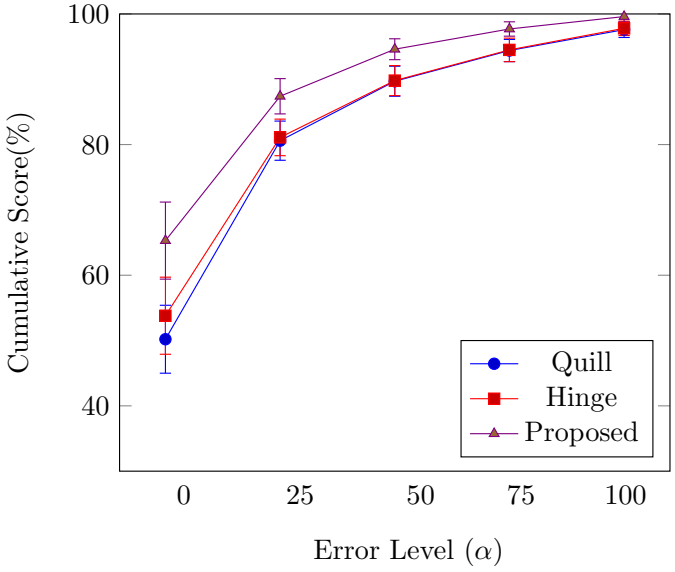


Fig. 8. Comparisons between our proposed methods and other methods in terms of the cumulative scores. Writer duplicates were excluded in the training data set.

4.5. Performance when including writer duplicates

This section provides the experimental results in the *including writer duplicates* scenario. Table 3 shows the results in terms of the MAEs and Cumulative Score (CS) with $\alpha = 0$ and $\alpha = 25$. From the

[2] https://redmine.hpc.rug.nl/redmine/projects/peregrine/wiki.

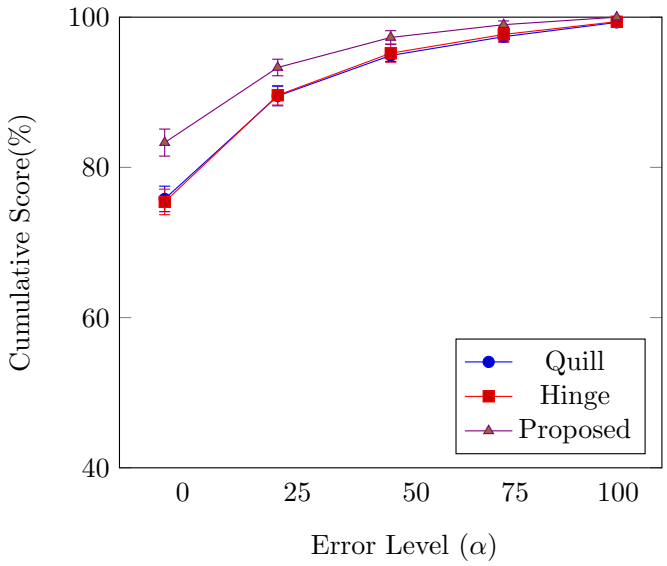


Fig. 9. Comparisons between our proposed methods and other methods in terms of the cumulative scores when including writer duplicates.

Table 3
The dating performance of different features when including writer duplicates.

Feature	MAEs	CS($\alpha = 0$)	CS($\alpha = 25$)
Quill Brink et al. (2012)	12.1 ± 0.9	75.8 ± 1.7%	89.5% ± 1.3
Hinge Bulacu and Schomaker (2007)	12.2 ± 0.9	75.4 ± 1.7%	89.6% ± 1.3
Proposed	**7.8** ± 0.7	**83.3** ± 1.8%	**93.3** ± 1.1%

table we can find that our proposed method also achieves the best performance. The MAE of the proposed method is 7.8 years, which is lower than the MAEs of Quill and Hinge by 4.3 years and 4.4 years, respectively. Our proposed method also improves both the CS($\alpha = 0$) and CS($\alpha = 25$), which achieves 83.3% and 93.3%. Fig. 9 shows the Cumulative Scores of the three methods. Our proposed method achieves much better results, especially when the error level is low. For example, when $\alpha = 0$, the CS scores are 75.8%, 75.4% and 83.3% for Quill, Hinge and the proposed method, respectively. Compared to the excluding writer duplicates scenario, the MAEs and CSs of the three methods are lower in the including writer duplicates scenario.

4.6. Comparison with other studies

In this section, we compare the proposed method with our previous work as well as a random guess. In the Monk system, characters from the historical documents in the MPS data set were manually labeled and used for dating. The results from the Monk system can be found on our website.[3] The study in He et al. (2014) used a three-stage global/local regression method to estimate the year of origin of historical documents using the combined Hinge and Contour features. The sub-strokes described by the PSD feature have been proposed in He and Schomaker (2015) for dating using the k Nearest Neighbors method. Table 4 shows the MAEs and CS($\alpha = 25$) of different methods. From the table we can see that our proposed method achieves state-of-the-art results on the MPS data set.

[3] http://application02.target.rug.nl/monk/Overslag/date-histogram-MPS.html.

Table 4
The MAEs and CSs($\alpha = 25$) of different methods.

Method	MAE	CS(α=25)
Random Guess	85.3 ± 58.5	25.7%
Monk Van der Zant et al. (2008)	36.0 ± 20.6	–
Study He et al. (2014)	35.4	63.5%
Study He and Schomaker (2015)	20.9	–
Proposed (wr.excl. scenario)	15.1 ± 2.3	87.4%
Proposed (wr.incl. scenario)	7.8 ± 0.7	93.3%

Table 5
Dating performance obtained with different methods.

Method	*wr.excl.* scenario		*wr.incl.* scenario	
	MAE	CS($\alpha = 25$)	MAE	CS($\alpha = 25$)
Quill Brink et al. (2012)	23.7 ± 2.9	80.6% ± 3.0	12.1 ± 0.9	89.5% ± 1.3
Hinge Bulacu and Schomaker (2007)	22.1 ± 2.9	80.6% ± 3.1	12.2 ± 0.9	89.6% ± 1.3
SOM He and Schomaker (2015)	21.5 ± 3.3	81.9% ± 3.9	12.0 ± 0.7	89.2% ± 1.4
TPC	15.1 ± 2.3	87.4% ± 2.7	7.8 ± 0.7	93.3% ± 1.1

4.7. Comparison with the codebook trained by the unsupervised SOM

In this section, we compare the TPC with the codebook trained by the traditional unsupervised SOM method with a size of 625, denoted by SOM in this paper.

The performance of different methods is shown in Table 5, from which we can see that the performance of the Quill, Hinge and SOM are comparable, indicating that textural-based features (Quill and Hinge) and a grapheme-based feature (SOM) are equal to represent the handwriting style of historical documents. However, the TPC achieves much better results than the SOM. The proposed method based on the TPC improves the MAE by 6.4 and 4.2 in the *wr.excl.* and *wr.incl.* scenarios, respectively. The experimental results show that the codebook trained by the TPC method contains more discriminative information than the one trained by the traditional unsupervised SOM method. Therefore, the feature representation of historical documents based on a codebook which contains temporal information is more discriminative and powerful for dating.

4.8. Discussion

As we have reported in this paper indicate that our approach with graphemes detected and described according to the PSD feature achieves valuable results in historical document dating on the MPS data set. This success can be ascribed to two factors: the used graphemes and the TPC. The handwritten patterns detected by the PSD can very effectively capture the local characteristics of the structural pattern, which reflect the handwriting writing style of the writers He et al. (2015). Most importantly, it is a segmentation-free method. The PSD descriptor is also scale-invariant, which is very important since the font size in documents of the MPS data set is variant. This can be seen in Fig. 10, in which there are four patches from different images with a fixed image size of 1000 × 500. In the image space, the sizes of the font are different in these four patch images, which are determined by the font size of the original charters and the resolution of the scanned images.

In Computer Vision, the widely used patch-based descriptors are SIFT Lowe (2004), SURF Bay et al. (2008) and the Histogram of Oriented Gradients (HOG) Dalal and Triggs (2005). However, these features are not suitable for historical document images. The key points detected by SIFT and SURF are located not only on the strokes but also on the background or near the stroke contours, and the corresponding features capture more the textural information of the text outline than the handwriting style of the strokes

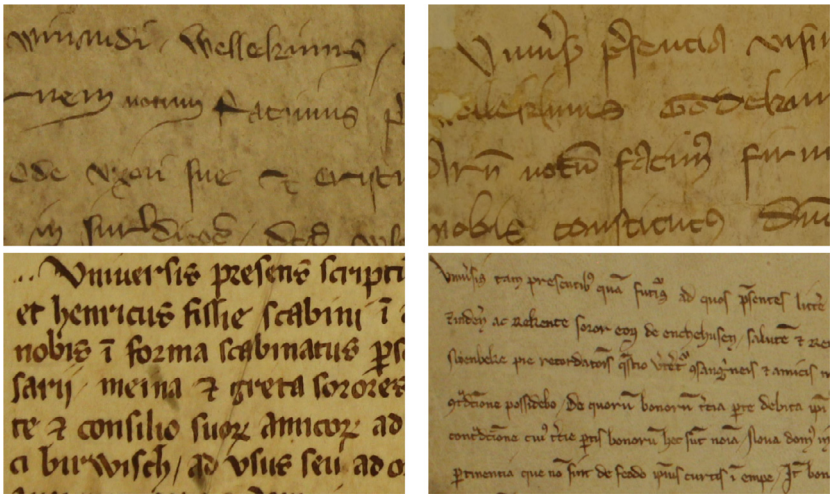


Fig. 10. The patches of historical document images in the MPS data set with fixed image size 1000 × 500. The junction-contours approach is largely scale invariant.

or characters. The main problem with the HOG feature is that it is sensitive to the scale of the images. Applying HOG with a multi-scale strategy is computationally inefficient when document images have a high resolution (300 dpi).

The TPC codebook trained by the SOTM contains the temporally determined structure of the handwritten patterns, which can capture the subtle differences between the same or similar patterns in different periods. This can be inferred from the visual example in Fig. 6 and the compared results in Table 5.

Our proposed method has more advantages than the textural-based methods, such as Quill and Hinge. First, our proposed method is a grapheme-based method, which can be visualized for end users, while the textural-based method extracts the feature in the whole document, and the visualization has no meaning for humans. More importantly, the performance of the proposed method is better than Quill and Hinge. The Quill and Hinge features are somewhat too sensitive to the scale of the document images.

5. Conclusion

In this paper we explored the task of historical document dating based on scanned images of charters from four cities in the period 1300–1550 CE. The PSD feature, which is a scale-invariant and discriminative local descriptor, was used to detect and describe the handwritten patterns. Then the handwritten patterns were mapped into a temporal pattern codebook containing the temporal structure information of the training set. Our method achieved state-of-the-art results on the MPS data set. An important future task would be to date the many undated book scripts extant from the same period, using the collection of the MPS data set as a reference corpus.

Acknowledgments

This work has been supported by the Dutch Organization for Scientific Research NWO (project No. 380-50-006).

References

Abdi, M.N., Khemakhem, M., 2015. A model-based approach to offline text-independent Arabic writer identification and verification. Pattern Recognit. 48 (5), 1890–1903.
Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Computer Vision Image Understanding 110 (3), 346–359.
Brink, A., Smit, J., Bulacu, M., Schomaker, L., 2012. Writer identification using directional ink-trace width measurements. Pattern Recognit. 45 (1), 162–171.
Bulacu, M., Schomaker, L., 2007. Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Anal. Mach.Intell. 29 (4), 701–717.
Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition (CVPR), 1, pp. 886–893.
Geng, X., Zhou, Z.-H., Smith-Miles, K., 2007. Automatic age estimation based on facial aging patterns. IEEE Trans. Pattern Anal. Mach.Intell. 29 (12), 2234–2240.
Guo, G., Fu, Y., Dyer, C.R., Huang, T.S., 2008. Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Trans. Image Process. 17 (7), 1178–1188.
He, S., Samara, P., Burgers, J., Schomaker, L., 2016. Image-based historical manuscript dating using contour and stroke fragments. Pattern Recognit. 58, 159–171.
He, S., Samara, P., Burgers, J., Schomaker, L., 2014. Towards style-based dating of historical documents. In: International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 265–270.
He, S., Schomaker, L., 2014. Delta-n hinge: rotation-invariant features for writer identification. In: International Conference on Pattern Recognition (ICPR), pp. 2023–2028.
He, S., Schomaker, L., 2015. A polar stroke descriptor for classification of historical documents. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 6–10.
He, S., Wiering, M., Schomaker, L., 2015. Junction detection in handwritten documents and its application to writer identification. Pattern Recognit. 48 (12), 4036–4048.
Helli, B., Moghaddam, M.E., 2010. A text-independent persian writer identification based on feature relation graph (FRG). Pattern Recognit. 43 (6), 2199–2209.
Kohonen, T., 1998. The self-organizing map. Neurocomputing 21 (1), 1–6.
Lee, Y.J., Efros, A., Hebert, M., et al., 2013. Style-aware mid-level representation for discovering visual connections in space and time. In: IEEE International Conference on Computer Vision (ICCV), pp. 1857–1864.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int.J.Comput.Vision 60 (2), 91–110.
Moghaddam, R.F., Cheriet, M., 2012. AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization. Pattern Recognit. 45 (6), 2419–2431.
Newell, A.J., Griffin, L.D., 2014. Writer identification using oriented basic image features and the delta encoding. Pattern Recognit. 47 (6), 2255–2265.
Otsu, N., 1975. A threshold selection method from gray-level histograms. Automatica 11 (285–296), 23–27.
Palermo, F., Hays, J., Efros, A.A., 2012. Dating historical color images. In: Computer Vision–ECCV 2012. Springer, pp. 499–512.
Papaodysseus, C., Rousopoulos, P., Giannopoulos, F., Zannos, S., Arabadjis, D., Panagopoulos, M., Kalfa, E., Blackwell, C., Tracy, S., 2014. Identifying the writer of ancient inscriptions and Byzantine codices. a novel approach. Comput. Vision Image Understanding 121, 57–73.
Howe, N.R., Yang, A., Penn, M., 2015. A character style library for Syriac manuscripts. In: Workshop on Historical Document Image and Processing (HIP), pp. 123–128.
Said, H.E., Tan, T.N., Baker, K.D., 2000. Personal identification based on handwriting. Pattern Recognit. 33 (1), 149–160.
Samara, P., 2014. Towards a medieval palaeographical scale. In: Fees, I. (Ed.), Papsturkundenforschung zwischen internationaler Vernetzung und Digitalisierung, Munich. http://rep.adw-goe.de/bitstream/handle/11858/00-001S-0000-0023-9A13-A/5_Samara.pdf?sequence=69
Sarlin, P., 2013. Self-organizing time map: an abstraction of temporal multivariate patterns. Neurocomputing 99, 496–508.
Sarlin, P., Yao, Z., 2013. Clustering of the self-organizing time map. Neurocomputing 121, 317–327.

Sauvola, J., Pietikäinen, M., 2000. Adaptive document image binarization. Pattern Recognit. 33 (2), 225–236.
Schomaker, L., Bulacu, M., 2004. Automatic writer identification using connected–component contours and edge-based features of uppercase western script. IEEE Trans. Pattern Anal. Mach.Intell. 26 (6), 787–798.
Siddiqi, I., Vincent, N., 2010. Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. Pattern Recognit. 43 (11), 3853–3865.
Wahlberg, F., Martensson, L., Brun, A., 2015. Large scale style based dating of medieval manuscripts. In: Workshop on Historical Document Image and Processing (HIP), pp. 107–114.
Van der Zant, T., Schomaker, L., Haak, K., 2008. Handwritten-word spotting using biologically inspired features. IEEE Trans. Pattern Anal. Mach.Intell. 30 (11), 1945–1957.
Zhang, T., Suen, C.Y., 1984. A fast parallel algorithm for thinning digital patterns. Commun. ACM 27 (3), 236–239.

Please cite this article as: S. He et al., Historical manuscript dating based on temporal pattern codebook, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.08.008