Historical Document Dating using Unsupervised Attribute Learning

Sheng He*, Petros Samara†, Jan Burgers‡, Lambert Schomaker*
*Institute of Artificial Intelligence and Cognitive Engineering, University of Groningen, the Netherlands
†Instituut voor Nederlandse geschiedenis, Den Haag, the Netherlands
‡University of Amsterdam, the Netherlands
Email: heshengxgd@gmail.com, petros.samara@huygens.knaw.nl, J.W.J.Burgers@uva.nl, L.Schomaker@ai.rug.nl

***Abstract*—The date of historical documents is an important metadata for scholars using them, as they need to know the historical context of the documents. This paper presents a novel attribute representation for medieval documents to automatically estimate the date information, which are the years they had been written. Non-semantic attributes are discovered in the low-level feature space using an unsupervised attribute learning method. A negative data set is involved in the attribute learning to make sure that our system rejects the documents which are not from the Middle Ages nor from the same archives. Experimental results on the basis of the Medieval Paleographic Scale (MPS) data set demonstrate that the proposed method achieves the state-of-the-art result.**

I. Introduction

The date is the one of the important metadata of historical documents. Traditionally, scholars who study historical materials estimate the date by using the individual non-verbal intuition, instead of objective criteria. Automatic writer identification techniques have been developed in the last decade, making it possible to automatically date historical documents according to the writing styles of the text. Automatic date estimation provides an efficient tool for scholars who work with a large digitized database. Therefore, it has attracted several researches [1], [2], [3], [4], [5].

Document representations play an important role in automatic date estimation based on the writing style. The basic assumption is that documents written in the same period would have a similar writing style measured by a certain representation. The reason behind this assumption is that most historical documents were written by the professional scribes, using more or less the same style, whose writing careers covered about two to three decades. Therefore, the most important task is to establish appropriate features which can capture the different writing styles of the scribes working in different periods.

Although there are many features used for writer identification, most of them are too sensitive for the individual writers, as a result of which they are not adequate to capture the general styles of the scribes in a given period. For example, if the documents from several writers are available for training, it may be impossible to estimate the date of the undated document from other writers using features for writer identification. Therefore, we consider to use the mid-level representation, which maps the low-level features into the attribute space.

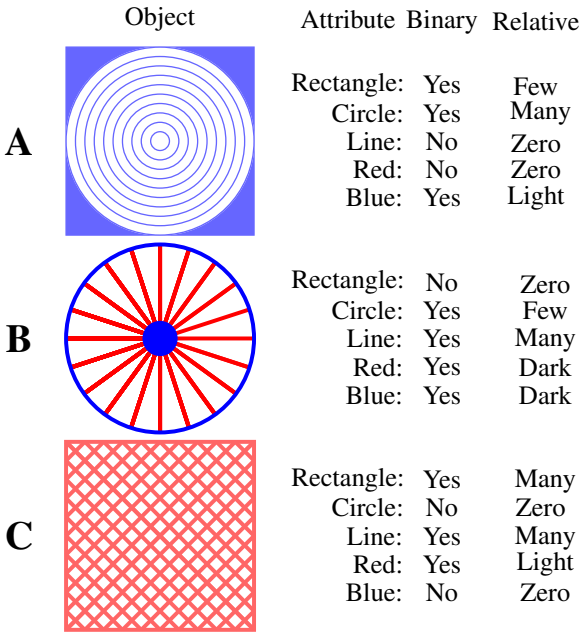

Fig. 1. An example of the binary and relative descriptions of objects.

Attributes [6] are linguistically related descriptors of objects with high-level semantically meaningful properties. Fig. 1 shows examples of objects and attributes. There are two types of attributes: binary and relative attributes. The binary attribute is the property that whether a certain object presents or not and the relative attribute indicates the strength of a property in an object with respect to other object [7]. For example, in Fig. 1 the attribute "Blue" is presented in the objects A and B and the blue color in B is more dark than it in A. Attributes can be shared between objects and the similar objects share large parts of attributes.

For handwriting, the possible attributes are: (1) the wide loops and narrow loops created by the letters 'e', 'o' et al. (2) long crosses on 't' and short crosses. (3) large gaps between words and crowded words. (4) slant to right or left or no slant. (5) rounded letters and connected letters. (6) darker, thicker strokes caused by a heavy pen pressure. Although these attributes can be learned from annotations, it is very difficult

for humans to describe the writing styles of handwritten documents, even for the paleographical experts.

In this paper, we use the principle of relative attributes as introduced by [7] to describe the writing styles of historical documents. We assume that there are several documents which can be considered as the anchors that represent the typical writing styles, and other documents can be described as the strength of the relative similarity with respect to the anchors. We use a cluster method to learn the anchors in the given low-level feature space and the Support Vector Machine (SVM) is used to compute the strength of the similarity.

II. Related work

A. Attribute learning

In the recent literature, two main ways have been laid out to design attributes. Most of attributes were manually labelled by picking a set of works (or attributes) to describe the images [6], which can be considered as supervised learning algorithms. Therefore, they have semantic meanings and can describe images or objects with textural descriptions. Other works [8], [9], [10] used semi-supervised methods, which need annotational images or category labels to learn semantic or non-semantic attributes. Augmented attribute representations were proposed in [8], which combined the fixed semantic attributes and the learned non-semantic attributes by augmenting the existed semantic attributes. Binary codes which were considered as attributes in [9] were learned by a data-driven way for image retrieval. Category-level attributes are designed in [10] with a semi-supervised method with category labels for category classification.

Manually designing semantic attributes have several limitations. Firstly, it is hard to develop a large number of semantic attributes [10], even picking a set of works from large natural text corpora [11]. Secondly, the human designed attributes may be intuitive but not discriminative for the visual recognition task [10]. Thirdly, the manually defined attributes are hard to capture the data structure embedded in low-level feature spaces. More importantly, in some specific domain, it is a challenge to define the semantic attributes. As we said before, it is hard even for experts to define the attributes of the writing style in historical documents.

B. Other studies in historical document dating

Several dating methods have been developed recently. In [12], the problem of dating historical color photographs has been studied, using features which capture temporally discriminative information. The dating of medieval manuscripts based on the "Svenskt diplomatariums huvudkartotek" (SDHK) data set has been proposed in [3], with corpus spanning from the years 1050-1523. The date of the document is estimated using regression methods with shape features. Automatically estimating the secure dates of Syriac documents has been proposed in [5] based on a collection of securely dated letter samples. In [4], an approach to estimate the unknown publication date of printed historical documents has been proposed using Convolutional Neural Networks (CNN).

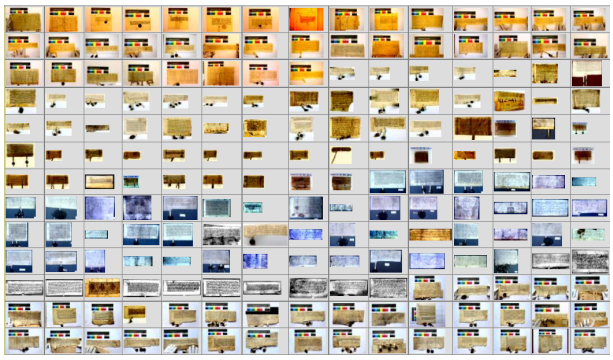

Fig. 2. The thumbnail images of historical documents around 1475 in the MPS data set.

III. Medieval Paleographic Scale (MPS)

The Medieval Paleographic Scale (MPS) data set was first used in [2] for historical document dating. The charters in the MPS data set are collected from five cities: Arnhem, Leiden, Groningen and Leuven/Brussels[1] together representing the corners of the Medieval Dutch language area. The MPS covers the period from around 1300 to 1550. The charters are sampled around "key years", like 1300±5, 1325±5, $\cdots$, 1550±5. Up to now there are 3700 documents in the MPS data set. Fig. 2 shows the thumbnails of the historical documents around 1475 in the MPS data set

IV. Method

A. Attribute representation

The attributes can be specificed either on certain characters, certain words, or on certain text zones in handwritten images. In this paper, we consider the attributes of the writing styles on the page level without using any segmentation methods because characters or words segmentation in the degraded historical documents is very difficult.

We use $A = \{a_1, \cdots, a_K\}$ to denote the attributes of the handwriting styles in the considered periods while K is the number of attributes. Each attribute a_i represents certain properties of which a subset of documents are stronger than others. We use the confidence score to measure such relative information and the large score means a stronger property appeared. In order to compute the confidence scores for the given document, an attribute classifier $\mathbf{w}$ is trained for each attribute using a linear SVM. Then a set of attribute classifiers is obtained: $\mathbf{W} = \{\mathbf{w}_1, \cdots, \mathbf{w}_K\}$, and the given document $\mathbf{D}(X) = \{x_1, \cdots, x_N\}$ can be represented by the vector of confidence scores $S = \{s_1, \cdots, s_K\}$ computed from the attribute classifiers:

$$s_i = \Big(1.0 + \exp(-\sum_{j=1}^{N} w_{ij} x_j)\Big)^{-1} \quad (1)$$

where X is the low-level feature vector and N is the number of the dimension of the X. The confidence score vector S

[1]The charters from Brussels were used by kind permission of Erik Kwakkel (Leiden University), who collected them.

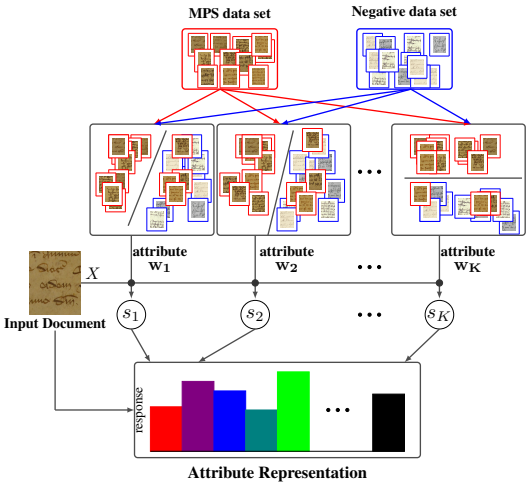

Fig. 3. Illustration of the framework of the proposed attribute learning method. Given the training MPS data set and the negative set, k-means is used to generate K clusters and an attribute classifier is trained for each cluster. For an input document, the final attribute representation is built based on the responses of the attribute classifiers.

is the attribute representation in the K-dimensional attribute space.

B. Unsupervised attribute discovery

Usually, attributes are linguistic descriptors of objects. However, finding a large word set to describe the writing styles of historical documents is very difficult as we mentioned before. In addition, manually labeling attributes is tedious. Therefore, we use the max-margin multiple-instance learning method [13] to discover the non-semantic attributes $A = \{a_1, \cdots, a_K\}$. We assume that documents with the same attribute are closed to each other in the feature space X and there is an anchor document to represent such attribute. Therefore, we choose the k-means cluster method to generate K clusters in the feature space X and we assume that each cluster is an anchor which is corresponding to one single attribute. Then the attribute classifier $\mathbf{w}_i$ is trained using the documents in the i-th cluster as the positive instances and other documents in the training set as the negative instances. Fig. 3 shows the framework of the proposed attribute representation learning method.

C. Rejection criteria

In order to reject the documents which are not from the considered period or from the same archives of the MPS data set, we randomly selected 600 scanned pages from the Monk system [14] and 99 modern handwritings from the Firemaker data set [15] as the negative samples for training the attribute classifiers. Fig. 4 shows several examples of the selected negative instances. We give the definition of the average confidence scores acs as:

$$acs = \frac{1}{K}\sum_{j=1}^{K} s_j \tag{2}$$

Fig. 4. An example of selected negative samples of historical documents from Monk system and modern handwritings.

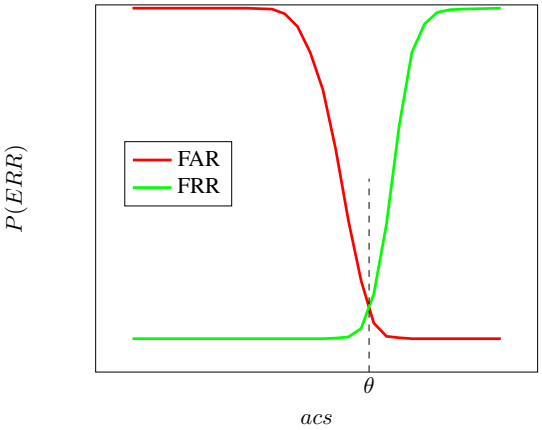

Fig. 5. The distribution of the FAR and FRR distributions. The threshold is chosen as the EER point (where FAR=FRR). This is the result using the Hinge feature with K=100.

which is the average confidence scores of the attribute classifiers. Then reject criterion is given as:

$$acs < \theta \tag{3}$$

where θ is the threshold. Usually, the false accept rate (FAR) and false reject rate (FRR) are used to measure the performance of the system with different threshold values. Fig. 5 shows an example of the distributions using Hinge feature with 100 attributes in terms of the FAR and FRR. The best threshold can be found on the point where FAR=FRR, which is also known as the equal error rate (EER) point.

D. Dating

We assume that all the documents from the same key year form a class. Thus, there are 11 classes in the MPS data set, corresponding to a multi-label classification problem. We choose the classification instead of the regression method because the document distribution in the considered period has a obvious borders between the nearby key years in the MPS data set according to the domain experts (paleographers). We train 11 classifiers using a linear SVM and the query document is assigned to the key year with the maximal SVM score among the 11 trained classifiers.

E. Low-level features

In this section, we give a brief description of the features used for dating. Four features, which are used for writer identification [16], [17], [18] are involved in the experiments, including the textural-based features, such as Hinge and Quill and the grapheme-based features, such as Junclets and Strokelets.

Hinge: The Hinge is a texture level feature, which is a joint probability distribution of the orientations of the two legs of the obtained hinge lied on the ink contours [17]. It captures individual handwriting style in terms of the orientation and curvature information of handwritings. There are two parameters in the Hinge feature: the number of angle bins p and the leg length r. In this paper, we set p=40 and r=20, following the suggestion in [19].

Quill: Inspired by paleographic methodology, the Quill feature [19] captures the relation between the ink direction and the width of the ink trace. The trace-width variations in historical documents were produced by the quill, as a result of the capillary-action of the writing instrument, the "feather", in addition to the pressure that was exercised. There is an additional parameter for the Quill feature: the number of bins q into which the measured stroke width is quantized. We set it to 40 in this manuscript.

Junclets: The junction feature [20] is a local descriptor for the singular structural regions in the text. It is the distribution of the stroke length on the junction point (the fork points and the high curvature points on the skeleton lines of the strokes) in a polar space. The Junclets feature is the probability of the junctions based on a trained codebook with a size of 625.

Strokelets: The connected components of the text are segmented on the fork points into sub-strokes and the Polar Stroke Descriptor [21] is used to describe the sub-stroke. The Strokelets feature is the probability of the sub-strokes based on a trained codebook. The size of the Strokelets is the same as the size of Junclets.

V. Experimental Results

The performance of historical document dating can be measured by the mean absolute error (MAE) and the cumulative score (CS), which criteria are also used, for example, in age estimation using face images [22]. The MAE is a Manhattan-type distance, which is typically defined as:

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

where $K(y_i)$ is the ground truth key year of the input document y_i, $\overline{K(y_i)}$ is the estimated key year and N is the number of test documents. The cumulative score CS is typically defined as [22]:

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

where $N_{e\leq\alpha}$ is the number of test images on which the key year estimation makes an absolute error no higher than the acceptable error level: α years. Since, when dating documents, the acceptable error level is rather low (25 or at the most 50 years), the cumulative scores at these lower error levels are most important to evaluate the performance of the system [22].

TABLE I
THE PERFORMANCE OF THE DATING WITH DIFFERENT CONFIGURATIONS. THE "WR.INCL." MEANS THE TEST SETTING INCLUDING WRITER DUPLICATES, THE "WR.EXCL." MEANS THE TEST SETTING EXCLUDING WRITER DUPLICATES AND THE "CITY.SP." MEANS THE TEST SETTING EXCLUDING GEOGRAPHICAL INFORMATION.

	Feature	Hinge	Quill	Junclets	Strokelets
wr.incl.	MAEs	14.8±1.9	23.6±4.6	13.3±0.8	12.8±0.6
	FRRs	2.9±1.3	20.6±2.7	11.1±5.7	15.4±2.9
	CS(α=25)	83.7±5.3	61.2±3.9	78.8±3.5	75.9±3.4
wr.excl.	MAEs	16.7	26.4	14.4	14.1
	FRRs	2.5	3.0	14.6	12.5
	CS(α=25)	79.6	68.6	72.3	74.3
city.sp.	MAEs	47.3±3.1	53.6±6.8	42.5±7.4	41.0±6.7
	FRRs	2.9±1.3	14.4±16.6	4.9±3.5	12.0±6.2
	CS(α=25)	50.3±4.0	39.9±5.1	53.8±5.7	53.3±8.8

A. Dating results when including writer duplicates

In this section, we conduct the experiment of the dating on the MPS data set by randomly splitting it into two parts: a training set (70%) and a testing set (30%). The experiment is repeated 20 times and the average results with the standard deviation are reported. Fig. 6 shows the results of the MAEs, the false rejection rate with different numbers of the attributes and the CSs with 200 attributes. Generally, when the number of attributes K increases, the MAEs go down for all the considered features. The lower MAEs are obtained when K is higher than 100. However, the false rejection rates are also higher when K is large. Among the four features, the performance of the Quill is lower than other three features. The Strokelets achieves the best MAE (12.8) and the Hinge works better than others in terms of CSs. The best performance for each feature is shown in Table I. In terms of MAEs, the performance of the Strokelets is slightly better than Junclets and the Hinge, by 0.5 and 2 respectively. However, the Cumulative Score with α=25 of the Hinge is higher than the Junclets and the Strokelets, by 4.9% and 7.9%, respectively.

B. Dating results when excluding writer duplicates

Because the charters in the MPS data set were written by professional scribes, many of whom kept on producing documents during a long period (20, or even30 years), some documents from different key years may be from the same writer. 2563 documents in the MPS are labeled with the writer (which means that the writing hand is known from more than one document); the writers of the other documents are unknown. In order to avoid effectively practicing writer identification when we are dating, we only use the documents labeled with writers for training, while documents whose writers are unknown are used for testing. Fig. 7 shows the results with different numbers of the attributes. The best MAE is 14.1 using the Strokelets with 200 attributes. From the Fig. 7 and Table I we can see that the performance of the dating using attribute representations is insensitive to the contamination of

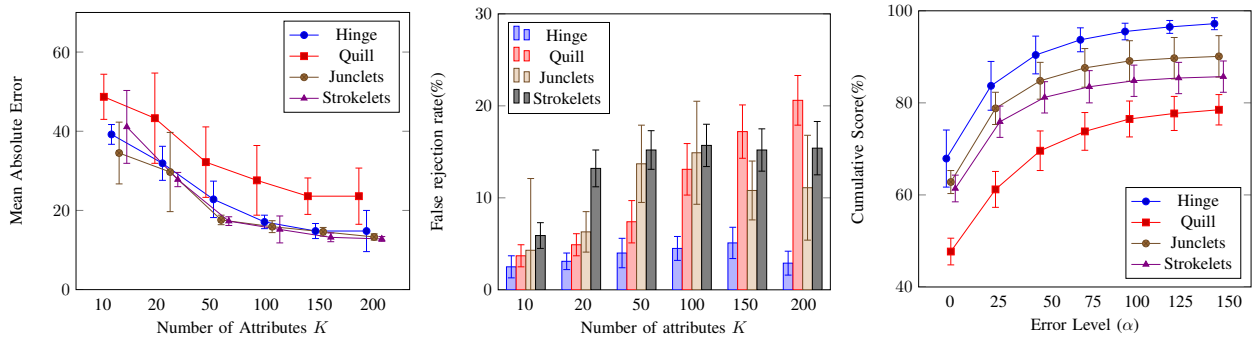

Fig. 6. Writer included results. The left figure shows the Mean Absolute Error with different number of attributes. The middle figure shows the False rejection rate and the right figure shows the cumulative scores when the attribute number K=200.

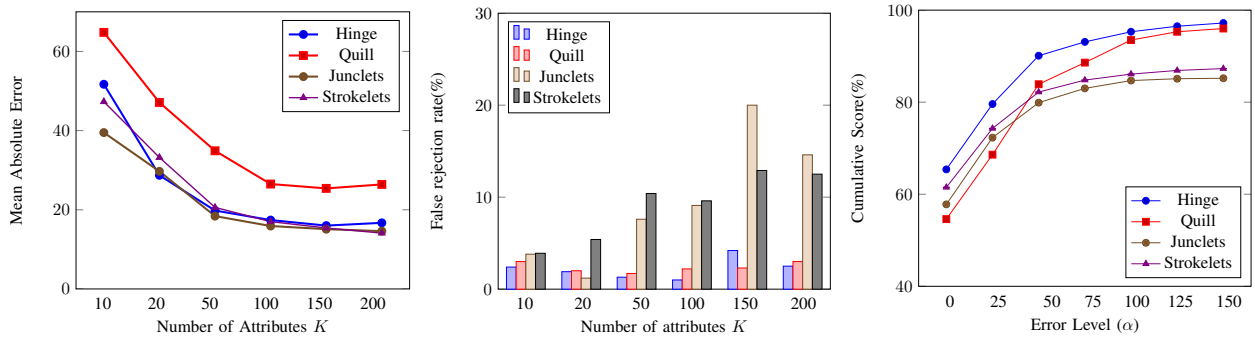

Fig. 7. Writer excluded results. The left figure shows the Mean Absolute Error with different number of attributes. The middle figure shows the False rejection rate and the right figure shows the cumulative scores when the attribute number K=200.

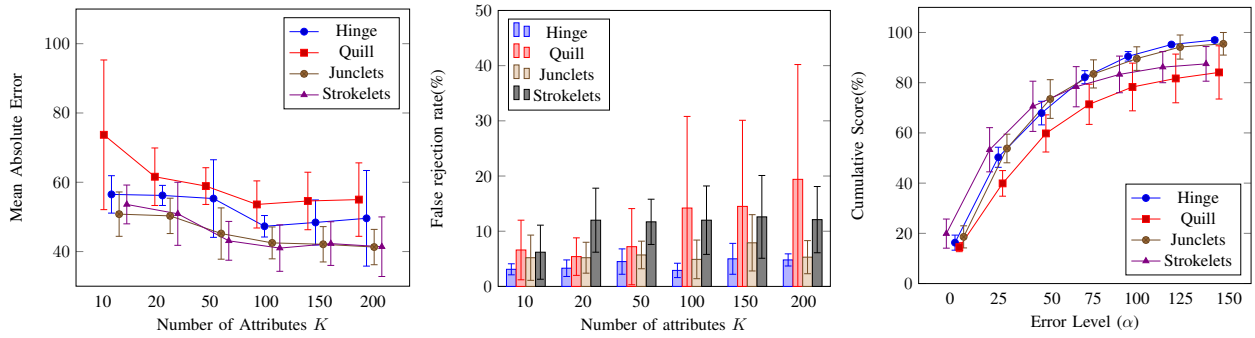

Fig. 8. City-specific results. The left figure shows the Mean Absolute Error with different number of attributes. The middle figure shows the False rejection rate and the right figure shows the cumulative scores when the attribute number K=100.

the writers. Therefore, the attribute representation captures the general writing style of each single period.

C. Dating results with geographical differentiation

It is very interesting to evaluate whether the writing styles in one city can be used to date the documents from other locations. In this section, we leave all the documents from one city out for testing and use the remaining documents for training. We do not consider the documents from Brussels because in our data set we have these only from the years 1300-1400. Fig. 8 shows the results with geographical differentiation. The best performance is achieved when the number of attributes K is 100 for all the features. From the Fig. 8 and Table I (the third row) we can conclude that the writing styles developed in one location are not useful to date the documents from other cities (MAE $\geq$ 40). There clearly existed specific local writing traditions.

D. MAEs in each year

Fig. 9 shows the MAEs of each key year including writer duplicates with 200 attributes, as well as the number of documents in each key year. Due to the low performance, the Quill feature was omitted from the Fig. 9. The same conclusion is reached that Strokelets works slightly better than Junclets and Hinge. The lowest MAE is on the key year 1450, in which the number of documents is largest. The number of documents

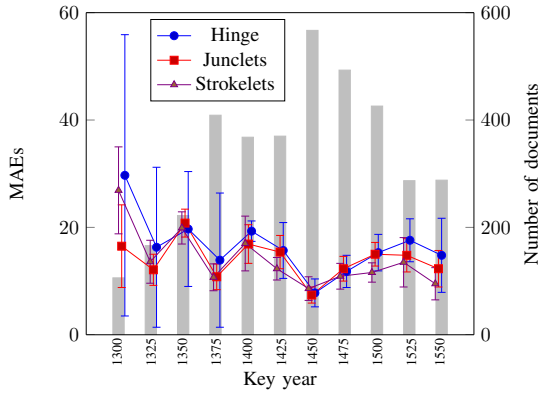

Fig. 9. The MAEs of each key year including writer duplicates with 200 attributes. The gray bars show the number of documents in each key year in the MPS data set.

TABLE II
THE PERFORMANCE OF DIFFERENT METHODS FOR HISTORICAL DOCUMENT DATING IN THE MPS DATA SET.

Methods	MAE	CS(α=25)
random guess	85.3	25.7%
study[2]	35.4	63.5%
study[21]	20.9	77.5%
writer included	12.8	75.9%
writer excluded	14.1	74.3%
geographical differentiation	41.0	53.3%

in the key year 1300 is smaller than in other key years and the MAEs are correspondingly higher with Hinge, and Strokelets. However, the highest MAE with Junclets is on the key year 1350, not on the 1300, and the change of MAEs with Junclets is more flat than Hinge, and Strokelets. Therefore, we can reach the conclusion: The MAE of each key year depends on its number of documents. If more document images are available, a lower MAE is achieved. The Pearson correlations between the number of document samples and the MAEs in each key year are: -0.826, -0.56 and -0.73 for Hinge, Junclets and Strokelets, respectively. The Junclets is less disturbed by the unbalanced data set.

E. Comparison with other studies

Table II shows the performance of the proposed method compared with the approaches proposed in [2], [21] and the random guess. The method in [2] used two layers of regression methods using the Hinge and the Fraglets features and the method in [21] used a simple k nearest neighbor classifier with the Strokelets feature. From the table we can find that our proposed method achieves the best results in term of MAE (12.8) using the Strokelets with 200 attributes. However, the CS of method [21] is higher than proposed method, by 1.6%.

VI. DISCUSSION AND CONCLUSION

In this paper, we have proposed an unsupervised learning method to discover the attributes that describe the writing styles of historical documents in the MPS data set. The proposed method is simple and easy to implement and achieves the state-of-the-art results on the MPS data set. In a future work, we will try to use other unsupervised learning methods which can deal with the relationships or correlations between the attributes and hence improve the discriminative of the attribute representation. In addition, the domain experts will attempt to label the semantic attributes of handwritings and build a full list of attributes.

ACKNOWLEDGMENTS

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

REFERENCES

[1] M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, and S. Tracy, "Automatic writer identification of ancient greek inscriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1404–1414, 2009.
[2] S. He, P. Samara, J. Burgers, and L. Schomaker, "Towards style-based dating of historical documents," in *ICFHR*, 2014, pp. 265–270.
[3] F. Wahlberg, L. Martensson, and A. Brun, "Large scale style based dating of medieval manuscripts," in *HIP*, 2015, pp. 107–114.
[4] Y. Li, D. Genzel, Y. Fujii, and A. C.Popat, "Publication date estimation for printed historical documents using convolutional neural networks," in *HIP*, 2015, pp. 99–106.
[5] N. R.Howe, A. Yang, and M. Penn, "A character style library for Syriac manuscripts," in *HIP*, 2015, pp. 123–128.
[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009, pp. 1778–1785.
[7] D. Parikh and K. Grauman, "Relative attributes," in *ICCV*, 2011, pp. 503–510.
[8] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Augmented attribute representations," in *ECCV*, 2012, pp. 242–255.
[9] M. Rastegari, A. Farhadi, and D. Forsyth, "Attribute discovery via predictable discriminative binary codes," in *ECCV*, 2012, pp. 876–889.
[10] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *CVPR*, 2013, pp. 771–778.
[11] L. Marchesotti, F. Perronnin, and F. Meylan, "Learning beautiful (and ugly) attributes." BMVC, 2013.
[12] F. Palermo, J. Hays, and A. A. Efros, "Dating historical color images," in *ECCV*, 2012, pp. 499–512.
[13] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple-instance dictionary learning," in *ICML*, 2013, pp. 846–854.
[14] T. Van der Zant, L. Schomaker, and K. Haak, "Handwritten-word spotting using biologically inspired features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1945–1957, 2008.
[15] L. Schomaker and L. Vuurpijl, "Forensic writer identificaiton: a benchmark data set and a comparison of two systems," *Technical Report, Nijmegen: NICI*, 2000.
[16] L. Schomaker and M. Bulacu, "Automatic writer identification using connected-component contours and edge-based features of uppercase Western script," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 4, pp. 787–798, 2004.
[17] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 701–717, 2007.
[18] S. He and L. Schomaker, "Delta-n Hinge: rotation-invariant features for writer identification," in *ICPR*, 2014, pp. 2023–2028.
[19] A. Brink, J. Smit, M. Bulacu, and L. Schomaker, "Writer identification using directional ink-trace width measurements," *Pattern Recognition*, vol. 45, no. 1, pp. 162–171, 2012.
[20] S. He, M. Wiering, and L. Schomaker, "Junction detection in handwritten documents and its application to writer identification," *Pattern Recognition*, vol. 48, no. 12, pp. 4036–4048, 2015.
[21] S. He and L. Schomaker, "A polar stroke descriptor for classification of historical documents," in *ICDAR*, 2015, pp. 6–10.
[22] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, 2007.