COMPUTERIZED PALEOGRAPHY: TOOLS FOR HISTORICAL MANUSCRIPTS

Lior Wolf, Liza Potikha, Nachum Dershowitz

The Blavatnik School of Computer Science
Tel Aviv University

Roni Shweka, Yaacov Choueka

The Friedberg Genizah Project
Jerusalem, Israel

ABSTRACT

The Digital Age has brought with it large-scale digitization of historical records. The modern scholar of history or of other disciplines is often faced today with hundreds of thousands of readily-available and potentially-relevant full or fragmentary documents, but without computer aids that would make it possible to find the sought-after needles in the proverbial haystack of online images. The problems are even more acute when documents are handwritten, since optical character recognition does not provide quality results.

We consider two tools: (1) a handwriting matching tool that is used to join together fragments of the same scribe, and (2) a paleographic classification tool that matches a given document to a large set of paleographic samples. Both tools are carefully designed not only to provide a high level of accuracy, but also to provide a clean and concise justification of the inferred results. This last requirement engenders challenges, such as sparsity of the representation, for which existing solutions are inappropriate for document analysis.

1. INTRODUCTION

Paleography, the study of old handwriting, is the discipline that engages in the decipherment of ancient texts. In addition, it tries to ascertain when and where a given manuscript was written, and—if possible—by whom. Paleographers bring many skills and tools to bear on these questions, in what is often a complicated and laborious task, requiring reference to paleographic, linguistic and archaeological data. Because it is difficult to quantify the degree of certainty in the final readings and assessments, experts have begun to develop computer-based methods for paleographic research. So far, such methods have only been applied to small cases due to the high degree of labor involved. Moreover, these efforts have focused almost exclusively on scribal identity, and tend to use the computer as a "black box" that receives images of manuscripts and replies with a classification of the handwriting, which scholars may be reluctant to accept.

In this work, we therefore study what we term "computerized paleography", that is, digital tools that furnish the analysis of a human paleographer with large-scale capabilities and assist with evidence-based inference. We explore two tasks: handwriting matching and paleographic classification, focusing mainly on the Cairo Genizah. Due to the scattering of the Genizah fragments in over 75 libraries, handwriting matching is a fundamental task for Genizah scholars. These scholars have expended a great deal of time and effort on manually rejoining leaves of the same original book or pamphlet, and on piecing together smaller fragments, often visiting numerous libraries for this purpose.

Recently, a system was proposed [1] that automatically identifies such potential *joins*, so that they may be verified by human experts. While successful in finding new joins, that system is a ranking algorithm that provides the expert with a simple numeric matching score for every pair of documents. The expert is left with the task of validating the join, without being provided any insight as to the basis for the score. We suggest that by using sparse representations and by avoiding metric learning, the results we obtain are easily interpretable. Sparse representations in the literature are shown to be ineffective for this specific task, and a new scheme is proposed.

The second task we consider is paleographic classification. We construct a paleographic tool that, given a fragment, provides suitable candidates for matching writing styles and dates. Such a tool can expedite the paleographic classification of fragments within the Genizah, and have long-reaching implications beyond Hebrew texts.

Background. Recent approaches to writer identification, such as letter- or grapheme-based methods, tend to employ local features and employ textual feature matching [2, 3]. Early uses of image analysis and processing for paleographic research are surveyed in [4]. Quantitative aspects can be measured by automated means and the results can be subjected to computer analysis and to automated clustering techniques [5].

2. BASELINE IMAGE REPRESENTATION

The baseline method follows previous work [1, 6] and employs a general framework for image representation that has been shown to excel in domains far removed from document processing, based on a bag of visual keywords. The signature of a leaf is based on descriptors collected from local patches in its fragments, centered around key visual locations, called *keypoints*. Such methods follow the following pipeline: first,

keypoints around the image are localized by examining the image locations that contain the most visual information. In our case, the pixels of the letters themselves are good candidates for keypoints, while the background pixels are less informative. Next, the local appearance at each such location is encoded as a vector. The entire image is represented by the obtained set of vectors, which, in turn, is represented as a single vector. This last encoding is based on obtaining a *dictionary*, containing representative prototypes of visual keywords, and counting, per image, the frequency of visual keywords that resemble each prototype.

In [1], it was suggested to detect the image keypoints using connected components, since, in Hebrew writing, letters are usually separated. Each keypoint is described by the popular SIFT descriptor [7]. To construct a dictionary, keypoints are detected in a set of documents set aside, and a large collection of 100,000 descriptors is subsampled. These are then clustered by the k-means algorithm to obtain a dictionary of varying sizes. Given a dictionary, a histogram-based method is employed to encode each manuscript leaf as a vector: for each cluster-center in the dictionary, the number of leaf descriptors (in the encoded image) closest to it are counted. The result is a histogram of the descriptors in the encoded leaf with as many bins as the size of the dictionary. Finally, the histogram vector is normalized to sum to 1.

3. SPARSE CODING OF HANDWRITINGS

We wish to present evidence to the user regarding the similarity of two document in a concise and intuitive way. This evidence should therefore be much more compact than the size of the dictionary, which typically contains hundreds, if not thousands, of prototypes. This suggests the use of a sparse image representation.

Recently, several advancements have been obtained for the bag-of-visual keywords approach. An efficient sparse coding technique [8] is obtained by requiring that the solutions be local, i.e. that coefficients are nearly 0 for all dictionary items that are not in the vicinity of the descriptor. This is enforced, for example, by adding weights in an L2-minimization setting.

While such sparse-coding methods outperform the standard bag-of-feature techniques for object recognition tasks, they do not seem to perform well on Genizah fragments, neither for join finding nor for paleographic classification. We hypothesize that the reason is that these methods were designed as hashing schemes, where recognition is obtained by detecting the footprint of unique and distinctive descriptors in the vector representing the image. In historical documents, however, the distinguishing descriptors repeat multiple times with small, yet significant, variations between documents.

Moreover, our interest in sparseness arises from the need to have a succinct representation for the entire document (or pair of documents). In both sparse and locality coding, individual keypoints are represented by sparse vectors; however, the entire image is represented by a combination of these vectors, in which the zero-elements are only those elements that are zero in all of the image keypoints.

We propose a novel per-document compacting process that is more suitable for historical documents. In this process, which we use both for handwriting matching and for paleographic classification, each document is represented by a limited set of descriptors. The size of this set depends on the diversity of the document, which is estimated by the number of prototypes required to represent the documents.

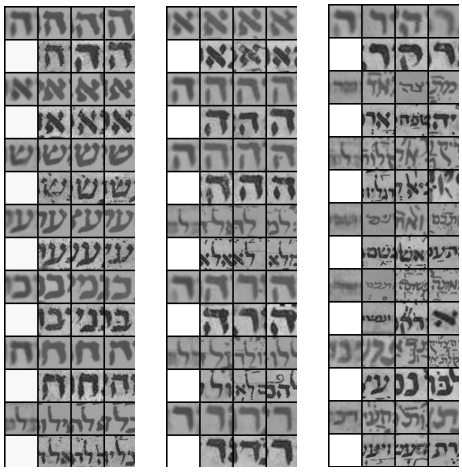
Given a document, we extract the set of keypoints and cluster them in the following manner. First, the correlation between every pair of keypoint descriptors is computed, and a graph is constructed in which the nodes correspond to the keypoints and edges exist between nodes of keypoints for which the correlation between the descriptors is above 0.9. The connected components of this graph are then computed. If there are fewer than 100 connected components, the process is repeated with a slightly lower threshold. Otherwise, 100 random components are selected.

At the next step, a k-means clustering algorithm is initialized with the centers of these 100 components. Upon convergence, up to 100 clusters of varying sizes are obtained for a given document D. We filter the clusters such that all clusters whose cardinality is smaller than half of the size of the largest cluster are discarded. The resulting clusters serve for two purposes. First, the number of remaining clusters C_D is taken as the diversity measure of the document, and second, the cluster centers themselves sometimes serve as training examples for training the prototype dictionary (see below). Typical values of C_D are between 10 and 40.

The sparsity of the fragment representation is enforced by a postprocessing step that nullifies every coefficient that is not among the highest C_D coefficients. In other words, the unnormalized representation is a dictionary-based histogram representation in which all coefficients other than the highest C_D are zeroed. The representation of a leaf is the sum of the individual representations of the fragment images (both recto and verso), normalized to sum to 1.

When comparing a pair of documents, we employ dot-product. The contributing coefficients are only those coefficients that are non-zero in both histograms. Therefore, the score is composed out of a handful of contributing factors. Each contributing factor is easily interpreted as the product of frequencies of a specific prototype in both documents.

Evidence charts. Our goal is to present the human expert with concise evidence justifying the matching score of every two document vectors. To this end, we make use of the sparsity property and provide the expert with a limited number of blocks, each containing sample keypoints from both documents that were assigned to the same dictionary prototype. There is one such block for each dictionary prototype whose associated coefficients are positive for both vectors, and the blocks are sorted by the contribution of the associated coeffi-



(a) (b) (c)

Fig. 1. Example of charts produced for pairs of matching documents (a,b) and a non-matching "false-positive" pair (c). Every two rows constitute a block that corresponds to one specific dictionary prototype. Each block segment depicts the cluster center (top-left corner), one row of keypoint examples from one document of the pair, and one row of examples from the second document. Pairs of matching keypoints (one per document) are placed one on top of the other, with the best matching pair on the left. Shown are the top 7 contributing blocks for each document, sorted from top to bottom. The total number of blocks is 25, 22, and 18 for the examples shown in (a), (b), and (c) respectively.

cient to the similarity score. See Figure 1 for examples.

In each block, we present examples from the first document above matching examples from the second. The leftmost pair is the best matching pair. The next pair is the next best matching pair after the "influence" of the first pair is "subtracted". More formally, we consider the set of keypoints from the first document that were assigned to a particular dictionary prototype, and the analogous set from the second document. All pairwise similarities between these two sets are computed. The most similar pair is selected, and the descriptors of all other keypoints are projected onto the subspace perpendicular to the average of the two selected descriptors. The process then repeats, a second pair is selected and so on.

4. OBTAINING SUITABLE DICTIONARIES

In our experiments, we observed a crucial difference in both accuracy and interpretability of the results, depending on the method used to construct the dictionary. This method must be tailored to the application at hand and the properties of the training data used to construct the dictionary.

Handwriting matching. The dictionary for handwriting matching is constructed from the 500 documents set aside in the "Genizah Benchmark" [1] for this purpose. First, each document separately goes through the process detailed in the previous section, in which the keypoints are clustered and smaller clusters are discarded. Then, the centers of the relatively large clusters from all 500 documents are clustered into 600 prototypes using k-means. This process ensures that only prominent templates are considered while constructing the dictionary, which is important in order to avoid dictionary prototypes that are based on visual patterns that arise as a result of binarization errors due to stains and other artifacts.

Following the clustering process, each prototype is weighed by the so called idf term (the logarithm of the quotient obtained by dividing the total number of documents by the number of documents containing the prototype) computed over the training set at each run. The weighed histogram count of each document therefore becomes a tf-idf vector [9].

Paleographic classification. For the paleographic classification task, we use sample documents and letters for each of 365 script types. These samples are extracted from the pages of the medieval Hebrew script specimen volumes, [10], which contain many example manuscripts whose provenance are known, and serve as an important tool in recent Hebrew paleography.

The dictionaries for the paleographic classification tool are constructed out of the extracted sample letters. As a result, the prototypes are constructed from letters that are much cleaner than the connected components that are extracted from the documents.The dictionary construction in this case follows two steps. First, k-means is applied to all professionally-drawn sample letters from the specimen volumes resulting in 600 clusters. The centers of each cluster are selected as prototypes. Second, the keypoints of each manuscript page from the specimen volumes are localized by extracting connected components of the binary images. The keypoints are then assigned to the prototypes by means of descriptor similarity. Lastly, prototypes for which the assigned keypoints (from the manuscripts) greatly differ from the clustered samples of the professionally-drawn letters are discarded.

The numerical criterion is as follows: first the center of all assigned manuscript keypoints is computed by taking the mean vector of all associated descriptors. Next, this new center is compared to the original cluster center. Clusters for which the two vectors differ considerably are discarded. Figure 2 shows the initial dictionary obtained and the results of the filtering step. As can be seen ambiguous clusters that do not correspond to actual letters are discarded by this process.

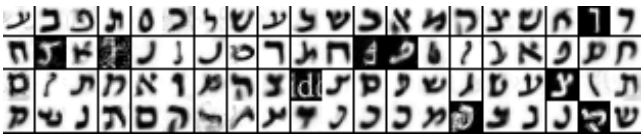


Fig. 2. A part of the dictionary constructed for the paleographic classification tool. The dictionary prototypes that were removed, because they were found to attract keypoints that are different than the ones clustered during training, are shown as negatives.

Method	AUC	EER	Accuracy	tpr@ 10^{-4}fpr
Baseline method w/ OSS learning [1]	95.6	9.2	93.7	76.0
LLC [8]	86.8	17.3	86.7	40.9
Sparse document coding (ours)	96.2	8.4	93.3	64.7
Multiple + physical + subject [6]	98.9	4.3	96.8	84.5

Table 1. Results obtained (percent) for the single dictionary baseline method, the LLC method, our sparse document coding method, and the state-of-the art method that combines non-handwriting data.

5. EXPERIMENTS

To evaluate the quality of our join-finding efforts, we use the comprehensive benchmark consisting of 31,315 leaves [1]. The benchmark comprises 10 equally sized sets, each containing 1000 positive pairs of images taken from the same joins and 2000 negative (non-join) pairs. Care is taken so that no known join appears in more than one set, and so that the number of positive pairs taken from one join does not exceed 20. To report results, one repeats the classification process 10 times. In each iteration, 9 sets are taken as training, and the results are evaluated on the 10th set. Results are reported by constructing an ROC curve for all splits together by computing statistics of the ROC curve (area under curve, equal error rate, and true positive rate at a low false positive rate of 0.001) and by recording average recognition rates for the 10 splits.

Table 1 summarizes the obtained benchmark results. As can be seen our sparse document coding method that is aimed at providing transparent (justifiable) scores performs similarly to the best baseline method, which employs the OSS metric-learning technique. It also considerably outperform the LLC [8] sparse-coding approach. Moreover, LLC vectors are not sparse (see Section 3), and 60% of the coefficients in each vector are non-zeros.

Our method, which is based solely on handwriting, is, as can be expected, not as good as the state-of-the-art method, which combines multiple dictionaries and physical measurements together with catalogic subject classification information. Note, however, that these auxiliary sources of information could be combined with our method as well.

Paleographic classification. To evaluate the performance of our paleographic classification tools, we collect a sample of 500 Genizah documents of varying script types and apply the tool to them. It was shown in [6], that unsupervised clustering is able to group such documents into (eighteen) clusters that are relatively pure with regard to a *coarse* paleographic classification. Here, the focus is on the supervised task of matching documents to an annotated gallery of examples.

For each Genizah document of the above-mentioned sample, we retrieve the most similar documents among those collected in [10]. We use both the baseline representation from [6] and our new sparse coding method, which relies on a dictionary extracted from these volumes. The results are then verified by a human who relies on both the evidence charts and on the original documents themselves. We mark whether the correct match is the first, second, third, or none of those. Table 2 summarizes the results, comparing the two methods. As can be seen, the proposed method does better than the baseline for the top-ranked document; however, the baseline method "catches-up" with the document ranked second.

Method/ ranking	Baseline method	Sparse document coding
1st match	68%	79%
2nd match	27%	16%
3rd match	3%	3%
other	2%	2%

Table 2. Results obtained for the baseline method and for the proposed sparse document coding method. The five highest ranking results are obtained by each method, and the average percent of correct matches per rank is recorded.

6. REFERENCES

[1] Lior Wolf, Rotem Littman, Naama Mayer, Tanya German, Nachum Dershowitz, Roni Shweka, and Yaacov Choueka, "Identifying join candidates in the Cairo Genizah," *IJCV*, 2010.

[2] M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, and S. Tracy, "Automatic writer identification of ancient Greek inscriptions," *PAMI*, 2009.

[3] A. Bensefia, T. Paquet, and L. Heutte, "Information retrieval based writer identification," in *Int. Conf. Document Analysis and Recognition*, 2003.

[4] Rjean Plamondon and Guy Lorette, "Automatic signature verification and writer identification – the state of the art," *Pattern Recognition*, vol. 22, no. 2, pp. 107–131, 1989.

[5] Arianna Ciula, "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis," in *Digital Medievalist*, 2005.

[6] Lior Wolf, Nachum Dershowitz, Liza Potikha, Tanya German, Roni Shweka, and Yaacov Choueka, "Automatic paleographic exploration of Genizah manuscripts.," in *Codicology and Palaeography in the Digital Age II*. 2011, Norderstedt: Books on Demand, Germany.

[7] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, 2004.

[8] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[9] K.S. Jones et al., "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[10] Malachi Beit-Arie, Edna Engel, and Ada Yardeni, *Specimens of Mediaeval Hebrew Scripts, Volume 1: Oriental and Yemenite Scripts (in Hebrew)*, The Israel Academy of Sciences and Humanities, Jerusalem, 1987.