Automatic Paleographic Exploration of Genizah Manuscripts

Lior Wolf, Nachum Dershowitz, Liza Potikha, Tanya German, Roni Shweka, Yaacov Choueka

Abstract

The Cairo Genizah is a collection containing approximately 250,000 hand-written fragments of mainly Jewish texts discovered in the late 19th century. The fragments are today spread out in some 75 libraries and private collections worldwide, and there is an ongoing effort to document and catalogue all extant fragments.

Paleographic information plays a key role in the study of the Genizah collection. Script style, and – more specifically – handwriting, can be used to identify fragments that might originate from the same original work. Such matched fragments, commonly referred to as “joins”, are currently identified manually by experts, and presumably only a small fraction of existing joins have been discovered to date. In this work, we show that automatic handwriting matching functions, obtained from non-specific features using a corpus of writing samples, can perform this task quite reliably.

In addition, we explore the problem of grouping various Genizah document by script style, without being provided any prior information about the relevant styles. The results show that the automatically obtained grouping agrees, for the most part, with the paleographic taxonomy. In cases where the system fails, it is due to apparent similarities between related scripts.

1 Introduction

Written text is one of the best sources for understanding historical life. The Cairo Genizah is a unique source of preserved middle-eastern texts, accumulated between the 10th and the 19th centuries. These texts are a mix of religious Jewish manuscripts with a smaller proportion of secular texts. As most of the original codices were not found in their original bound state, to make the study of the Genizah more efficient, there is an acute need to group the fragments and try to reconstruct the original codices. Throughout the years, scholars have devoted a great deal of time to manually identify such groups, referred to as *joins*, often visiting numerous libraries for this purpose.

Manual classification is currently the “gold standard” for finding joins. However, it is not scalable and cannot be applied to the entire corpus. We suggest automatically identifying candidate joins to be verified by human experts. To this end, we employ modern image-recognition tools such as local descriptors, bag-of-features representations and discriminative metric learning techniques. These techniques are modified for the problem at hand by applying suitable preprocessing and by using task-specific key-point selection techniques. Lastly, a bag of visual keywords approach is taken in which paleographic samples of various script styles are used. This major step is shown to increase performance considerably.

In addition to the join-finding effort, we also study the problem of automatically deriving the script style of Genizah documents. We choose to do it in an *unsupervised* manner, in which a clustering algorithm groups the various documents, thereby separating the image sets according the script style of each image.

2 Related Work

Genizah research Discovered in 1896 in the attic of a synagogue in an old quarter of Cairo, the Genizah is a large collection of discarded codices, scrolls, and documents, written mainly in the 10th to 15th centuries. The attic was emptied and its contents have found their way to over 75 libraries and collections around the world. The documents, with few exceptions, are of paper and vellum, and the texts are written mainly in Hebrew, Aramaic, and Judeo-Arabic (Arabic language in Hebrew characters), but also in many other languages (including Arabic, Judeo-Spanish, Coptic, Ethiopic, and even one in Chinese). The finds included fragments of lost works (such as the Hebrew original of the apocryphal Book of Ecclesiasticus), fragments of hitherto unknown works (such as the Damascas Document, later found among the Qumran scrolls), and autographs by famous personages, including the Andalusians Yehuda Halevi (1075–1141) and Maimonides (1138–1204).

Genizah documents have had an enormous impact on 20th century scholarship in a multitude of fields, including Bible, rabbinics, liturgy, history, and philology. Genizah research has, for example, transformed our understanding of medieval Mediterranean society and commerce, as evidenced by S. D. Goiten’s monumental five-volume work, *A Mediterranean Society*. See (Reif 2000) for the history of the Genizah and of Genizah research. Most of the material recovered from the Cairo Genizah has been microfilmed and catalogued in the intervening years, but the photographs are of mediocre quality and the data incomplete (thousands of fragments are still not listed in published catalogues).

The philanthropically-funded Friedberg Genizah Project is in the midst of a multi-year process of digitally photographing (in full color, at 600dpi) most – if not all – of the extant manuscripts. The entire Genizah collections of the Jewish Theological Seminary in New York (ENA), the Alliance Israélite Universelle in Paris (AIU), The Jewish National and University Library in Jerusalem (JNUL), the recently rediscovered collection in Geneva, and many smaller collections have already been digitized, and comprise about 90,000 images (recto and verso of each fragment). The digital preservation of another 140,000 fragments of the Taylor-Schechter Genizah Collection at The Cambridge University Library is now underway. The images are being made available to researchers online at www.genizah.org.

Unfortunately, most of the leaves that were found were not found bound together. Worse, many are fragmentary, whether torn or otherwise mutilated. Pages and fragments from the same work (book, collection, letter, etc.) may have found their way to disparate collections around the world. Some fragments are very difficult to read, as the ink has faded or the page discolored. Accordingly, 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, usually as part of their research in a particular topic or literary work. Despite the several thousands of such joins that have been identified by researchers, much more remains to be done (Lerner and Jerchower 2006).

Writer identification A related task to that of join finding is the task of scribe identification, in which the goal is to identify the writer by morphological characteristics of a writer's handwriting. Since historical documents are often incomplete and noisy, preprocessing is often applied to separate the background and to remove noise. See, for instance, (Bres, Eglin, and Auger 2006; Leedham, Varma, Patankar, and Govindarayu 2002). Latin letters are typically connected, unlike Hebrew ones which are usually only sporadically connected. Efforts were thus expended on designing segmentation algorithms to disconnect letters and facilitate identification. See (Casey and Lecolinet 1996) for a survey of the subject. The identification itself is done either by means of local features or by global statistics. Most recent approaches are of the first type and identify the writer using letter- or grapheme-based methods, which use textual feature matching (Panagopoulos, Papaodysseus, Rousopoulos, Dafi, and Tracy 2009; Bensefia, Paquet, and Heutte 2003). The work of (Bres, Eglin, and Auger 2006) uses text independent statistical features, and (Bulacu and Schomaker 2007a; Dinstein and Shapira 1982a) combine both local and global statistics.

Interestingly, there is a specialization to individual languages, employing language-specific letter structure and morphological characteristics (Bulacu and Schomaker 2007a; Panagopoulos, Papaodysseus, Rousopoulos, Dafi, and Tracy 2009;

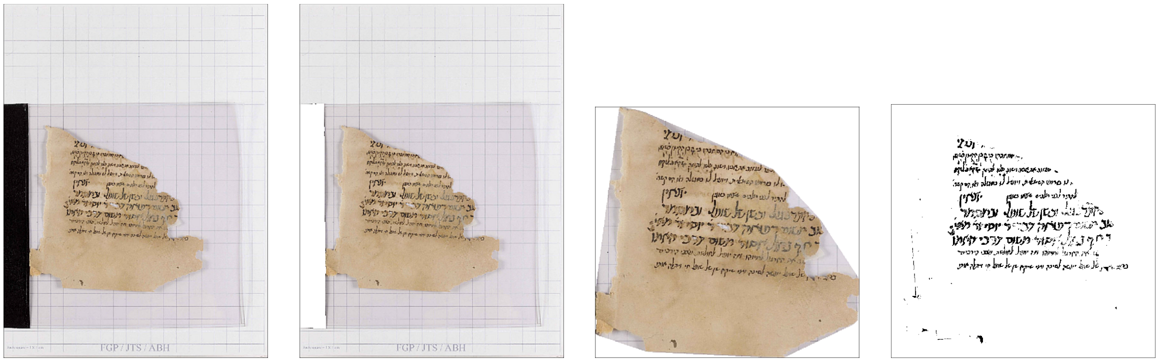
Dinstein and Shapira 1982a). In our work, we rely on the separation of Hebrew characters by employing a keypoint detection method that relies on connected components in the thresholded images.

Most of the abovementioned works identify the writer of the document from a list of known authors. Here, we focus on finding join candidates, and do not assume a labeled training set for each join. Note, however, that the techniques we use are not entirely suitable for distinguishing between different works of the same writer. Still, since writers are usually unknown (in the absence of a colophon or signature), and since joins are the common way to catalog Genizah documents, we focus on this task. Additional data such as text or topic identification, page size and number of lines can be used to help distinguish different works of the same writer.

Digital Paleography Paleographers traditionally use a mix of qualitative and quantitative features to distinguish hands (Mallon 1952). Early uses of image analysis and processing for paleographic research include (Fournier and Vienot 1971; Sirat 1981; Dinstein and Shapira 1982b); see (Plamondon and Lorette 1989) for a survey of other early work. Quantitative aspects can be measured by automated means and the results can be subjected to computer analysis and to automated clustering techniques (Ciula 2005; Aussems 2006; Aiolli and Ciula 2009). Features amenable to automatization, including texture (Said, Tan, and Baker 2000; Bulacu and Schomaker 2007b), angularities (Bulacu, Schomaker, and Vuurpijl 2003), and others (Brink, Smit, Bulacu, and Schomaker ; Aussems and Brink 2009) have been suggested. Concavity, moments and other features have been used to correctly classify selected Hebrew letters by writer (Bar-Yosef, Beckman, Kedem, and Dinstein 2007; Bar Yosef, Kedem, Dinstein, Beit-Arie, and Engel 2004). What distinguishes our work is that we are using generic image features for this purpose.

3 Image Processing and Physical Analysis

The images we obtained from the Friedberg Genizah Project are given as 300–600 dpi JPEGs, of arbitrarily aligned fragments placed on different backgrounds. An example, which is relatively artifact free, is shown in Figure 1(a). Many of the images, however, contain superfluous parts, such as paper tags, rulers, color tables, etc. (as in Figure 5). Therefore, a necessary step in our pipeline is preprocessing of the images to separate fragments from the background and to align fragments so the rows of text are horizontal. Then the physical properties of the fragments and of the text lines are measured. Both stages are described in detail in (Wolf, Littman, Mayer, Dershowitz, Shweka, and Choueka 2009).



(a) (b) (c) (d)

Figure 1: Example of a document from the Cairo Genizah (ENA collection). (a) The original image. (b) After removing the black folder. (c) After segmentation (using the convex hull). (d) After binarization and alignment.

3.1 Preprocessing

The goal of the preprocessing stage is to eliminate parts of the images that are irrelevant or may bias the join finding process, and to prepare the images for the representation stage.

Coarse manual alignment In the manual stage, the written sides of each fragment were manually identified. All the images were then manually rotated as necessary in multiples of $90°$, resulting in alignment in the range of $[-45°, 45°]$ from upright. This initial rotation prevents the auto-alignment step, explained below, from rotating documents upside-down. Note that both the identification of the written side and the coarse alignment stages could have been automated; however, the manual effort expended was not large, even for the entire Genizah collection.

Foreground segmentation The process of separating fragments from their background depends on the way the image was captured. At first, a classifier is used to identify foreground pixels based on RGB values or HSV values. To create a region-based segmentation of the fragments, we mark the connected components of the detected foreground pixels, and we calculate the convex hull of each component. Those steps retain almost all of the relevant parts of the images, while excluding most of the background.

Detection and removal of non-relevant components Labels, ruler, color swatches and any other non-relevant components that fall in separated regions are manually removed. In some images, especially large documents, a ruler is adjacent to the actual fragments and is not separated by the region-segmentation process. The ruler used in the images is of a known type, and we locate it by a detector, based on correspondence to a reference image of this ruler. The correspondence is done by RANSAC-based (Fischler and Bolles 1981) SIFT (Lowe 2004) keypoint matching. The region of the detected ruler is segmented by color and removed.

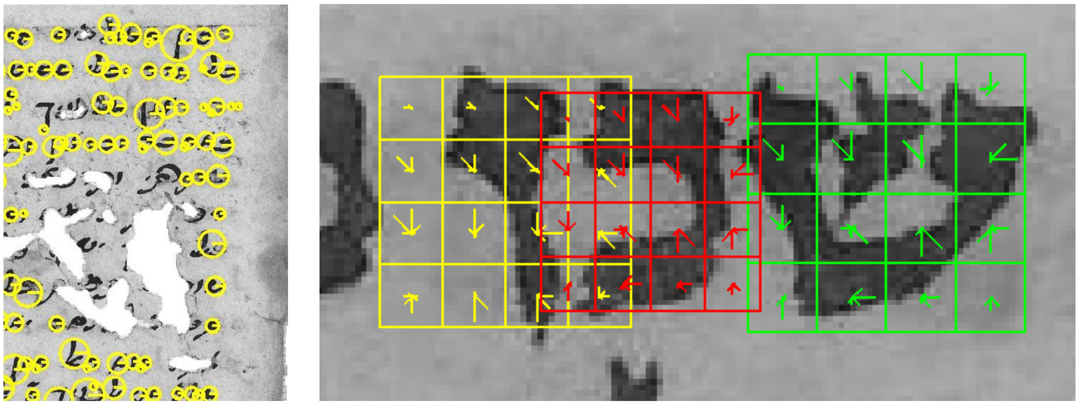
Binarization The regions detected in the foreground segmentation process are then binarized using the autobinarization tool of the ImageXpress 9.0 package by Accusoft Pegasus. To cope with failures of the Pegasus binarization, we also binarized the images using the local threshold set at 0.9 of the local average of the 50×50 patch around each pixel. The final binarization is the pixel-wise AND of the two binarization techniques. Pixels nearby the fragment boundary are set to zero. An example result is shown in Figure 1(b).

Auto-alignment Each region is rotated so the rows (lines of text) are in the horizontal direction. This is done using a simple method, which is similar to (Baird 1992; Srihari and Govindaraju 1989). For each rotation angle, we rotate the image at all angles and sum the pixels along horizontal lines. We then normalize each profile by dividing by the foreground region length along each such horizontal line. We then calculate the variance of the projection for each angle, and select the angle for which the variance is the largest.

Physical measurements The measurements that are used in the classification are characteristics of the text rows, and dimensions of the text bounding box. The number of text rows, height of the rows, and the spaces between the rows are calculated using the projection profile of the fragment. The text rows are localized at the maxima points of these projections. In addition, the minimal area bounding box of each fragment is also computed. Note that this bounding box does not need be axis-aligned.

4 Image Handwriting Representation

We decided to employ a general framework for image representation that was shown to excel in domains far removed from document processing, namely, a method based on a bag of visual keywords (Dance, Willamowski, Fan, Bray, and Csurka 2004; Lazebnik, Schmid, and Ponce 2006). The signature of a leaf is based on descriptors collected from local patches in its fragments, centered around keypoints. Such methods follow the following pipeline: first, key visual locations (keypoints) around the image are localized. The local appearance at each such location is then encoded as a vector. The entire image is



(a) (b)

Figure 2: (a) Keypoint detection methods using the proposed CC method. (b) SIFT descriptors of three nearby detected keypoints

represented by the obtained set of vectors, which in turn, is represented as a vector. This last encoding is based on obtaining a “dictionary” containing representative prototypes of visual keywords, and counting, for each image, the frequency of visual keywords that resemble each prototype.

4.1 Keypoint Detection

We detect the image keypoints using the fact that, in Hebrew writing, letters are usually separated. We start by calculating the connected components (CC) of the binarized images. (By *connected component* we mean a contiguous region of black pixels.) To filter out broken letter parts and dark patches arising from stains and border artifacts, we compare the size of the CC to the height of the lines which is estimated in a similar fashion to the alignment stage above. For encoding, each keypoint requires a scale, which is taken as the maximum dimension of the associated CC. Figure 2(a) shows the keypoints found using the SIFT and CC detectors.

4.2 Local Descriptors

Each keypoint is described by a descriptor vector. After experimenting with a large variety of such descriptors, the most popular descriptor, SIFT, was chosen for its accuracy. SIFT (Lowe 2004) encodes histograms of gradients in the image. Figure 2(b) illustrates the application of SIFT to one fragment.

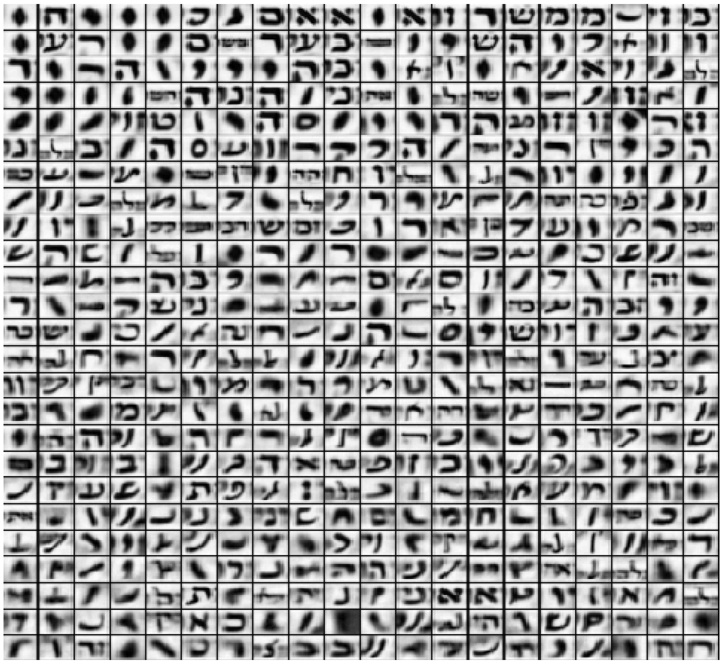


Figure 3: By clustering (grouping into homogenous groups) a large collection of visual descriptors obtained from random images, a set of distinctive visual keywords or prototypes is formed. This collection is referred to as a “dictionary”.

4.3 Dictionary Creation and Vectorization

Bag-of-visual-keyword techniques (Dance, Willamowski, Fan, Bray, and Csurka 2004) rely on a dictionary that contains a representative selection of descriptors obtained on various interest points. To this end, we set aside a small dataset of 500 documents. We detect keypoints in those documents and subsample a large collection of 100,000 descriptors. These are then clustered by the k-means algorithm to obtain a dictionary of varying sizes. The result is a set of prominent prototypes or “visual keywords”; see Figure 3.

Given a dictionary, we employ either histogram-based methods or distance-based methods to encode each leaf as a vector. In histogram-type vectorization methods (Dance, Willamowski, Fan, Bray, and Csurka 2004), one counts, for each cluster-center in the dictionary, the number of leaf descriptors (in the encoded image) closest to it. The result is a histogram of the descriptors in the encoded leaf with as many bins as the size of the dictionary. To account for the variability in fragment sizes, we normalize the histogram vector to sum to 1, that is, we divide the histogram vector by its sum. Another alternative is to normalize each vector so that its Euclidean norm is 1.

Distance-based representation techniques (Serre, Wolf, and Poggio 2005) are based on computing the minimum distance to all descriptors of the given leaf, for each prototype in the dictionary.

4.4 Employing Paleographic Data in the Representation

The resulting representation transforms each image to a vector based on a global dictionary, in which prototypes from various script types are mixed together. Since most Genizah joins contain just one script style, having multiple dictionaries, each corresponding to one script style, might be helpful.

To obtain such dictionaries, we apply the procedure described in Sec 4.3 multiple times, each time on documents of the same script. Such documents are extracted from the pages of the medieval Hebrew script specimen volumes (Beit-Arie, Engel, and Yardeni 1987; Beit-Arie, Engel, and Yardeni 2002). These volumes contain many examples of medieval Hebrew manuscripts, whose provenance are known, and serves as an important tool in Hebrew paleography. High-quality sample pages are placed side-by-side with professionally-drawn sample letters of the alphabet, based on exemplars from the manuscript. Note that we use the images of the original documents and not the individually extracted letter samples.

The groups of script styles obtained from this collection, and the corresponding pages of the collection are reported in Table 1. As can be seen, the major script styles are the square script, semi-cursive script, and cursive script. The geographic location is also influential on script style, so we extracted Oriental, Yemenite, and Spanish script groups from the same source.

Script style	Number of samples	Page list
Square and Semi-Square Oriental	51 images	I-1 – I-51
Semi-cursive Oriental	61 images	I-52 – I-112
Yemenite	42 images	I-113 – I-154
Square Spanish	62 images	II-1 – II-62
Semi-cursive Spanish	99 images	II-63 – II-161
Cursive Spanish	48 images	II-162 – II-209

Table 1: The page numbers in the Hebrew script sample volumes (Beit-Arie, Engel, and Yardeni 1987; Beit-Arie, Engel, and Yardeni 2002) of the document samples used to obtain the dictionary of each script type. “I” marks images extracted from (Beit-Arie, Engel, and Yardeni 1987); “II” marks those extracted from (Beit-Arie, Engel, and Yardeni 2002).

5 Finding Joins

To determine whether two fragments originate from the same manuscript, we compare their vector representation. The comparison can be performed in several ways, and it is often beneficial to combine multiple methods together.

5.1 Similarity Inference

Focusing on just one representation, each leaf is represented by one vector, for example, by the L2 normalized histogram of keypoint types. For every pair of leaves, we need to determine whether they are from the same join or not. Ideally, we would have a similarity function that would return a high value when two leaves are from the same join, and a low value otherwise. In this ideal case, a threshold over the similarity function provides a decision value.

The basic similarity score is obtained by considering, for every two vectors $\mathbf{p}$ and $\mathbf{q}$, the similarity derived from their Euclidean distance $e^{-||\mathbf{p}-\mathbf{q}||}$.

In our work, we also employ learned similarities. Tailoring similarity measures to available training data by applying learning techniques is gaining popularity; see, for example (Bilenko, Basu, and Mooney 2004; Cristianini, Kandola, Elisseeff, and Shawe-Taylor 2002; Hertz, Bar-Hillel, and Weinshall 2004; Shental, Hertz, Weinshall, and Pavel 2002; Weinberger, Blitzer, and Saul 2006; Xing, Ng, Jordan, and Russell 2003). Here, the similarity is to be learned from pairs of samples that are known to belong to the same join or not, and we choose to use a similarity that has been shown to be extremely successful in face-recognition work.

The OSS (Wolf, Hassner, and Taigman 2008; Wolf, Hassner, and Taigman 2009) is a similarity learning technique designed for the same/not-same problem. Given two vectors $\mathbf{p}$ and $\mathbf{q}$, their OSS score is computed by considering a training set of background sample vectors $\mathbf{A}$. This set of vectors contains examples of items different from either $\mathbf{p}$ and $\mathbf{q}$ (that is, they do not belong in the same class as neither $\mathbf{p}$ nor $\mathbf{q}$). Note, however, that these training samples are otherwise unlabeled. In our experiments, we take the set $\mathbf{A}$ to be one split out of the nine splits used for training at each iteration (see Section 5.3).

A measure of the similarity of $\mathbf{p}$ and $\mathbf{q}$ is then obtained as follows. First, a discriminative model is learned with $\mathbf{p}$ as a single positive example and $\mathbf{A}$ as a set of negative examples. This model is then used to classify the second vector, $\mathbf{q}$, and obtain a classification score. The nature of this score depends on the particular classifier used. We, following (Wolf, Hassner, and Taigman 2008), employ an LDA classifier, and the score is the signed distance of $\mathbf{q}$ from the decision boundary learned using $\mathbf{p}$ (positive example) and $\mathbf{A}$ (negative examples). A second such score is then obtained by repeating the same process with the roles of $\mathbf{p}$ and $\mathbf{q}$ switched: this time, a model learned with $\mathbf{q}$ as the positive example is used to classify $\mathbf{p}$, thus obtaining a second classification score. The final OSS is the sum of these two scores. It was shown in (Wolf, Hassner, and Taigman 2009) that LDA based OSS can be computed very efficiently.

5.2 Classification and Combination of Features

For the recognition of joins, we need to convert the similarity values of Section 5.1 to a decision value. Moreover, it is beneficial to combine several similarities together. For both these tasks we employ linear support vector machines (SVM), with fixed parameter value $C = 1$, as was done in (Wolf, Hassner, and Taigman 2008; Wolf, Bileschi, and Meyers 2006).

In the case of one-similarity, the similarity is fed to the SVM as a one-dimensional vector and training is performed on all training examples. In this case, the SVM just scales the similarities and determines a threshold for classification.

To combine several similarities together we use the SVM output (signed distance from hyperplane) obtained from each similarity separately and construct a vector. This vector is then fed to another SVM. The value output by the last classifier is our final classification score. This method of combining classifier output is called “stacking” (Wolpert 1992).

5.3 The Genizah Benchmark

To evaluate the quality of our join-finding efforts, we have constructed a comprehensive benchmark. Our benchmark, which is modeled after the LFW face recognition benchmark (Huang, Ramesh, Berg, and Learned-Miller 2007), consists of 31,315 leaves, all from the New York (ENA), Paris (AIU), and Jerusalem (JNUL) collections.

The benchmark consists of ten equally sized sets. Each contains 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 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, nine sets are taken as training, and the results are evaluated on the tenth set. Results are reported by constructing an ROC curve for all splits together (the outcome value for each pair is computed when this pair is a testing pair), by computing statistics of the ROC curve (area under curve, equal error rate, and true positive rate at a certain low false positive rate) and by recording average recognition rates for the 10 splits.

The ROC curve is an accepted form of measuring classification success. It is a graph (see Figure 4) in which the trade-off between false positive results and the recall rate is displayed. One would like to obtain perfect recall (identifying all joins) while make no false-positive errors, that is, without identifying non-joins as joins. However, in reality the task is challenging and therefore a certain number of false detections is expected for reaching high levels of recall.

The most interesting statistic from the practical point of view is the recall (true-positive) rate for a low-false positive rate. Since there are many theoretical join candidates in the Genizah, and since human verification effort is limited, any practical join-finding system should mark non-joins as joins only for a small percentage of these candidates.

5.4 Benchmark Results

We compare the performance of several methods, each based on a separate source of information. Not surprisingly, combining these methods together yields the best results.

Domain classification Over 95% of the Genizah documents are classified by their content's domain. The classification contains classes such as "Biblical", "Letters", "Liturgy", "Arabic Tafsir", "Aramic translation", and more. A similarity of -1 is assigned to pairs of documents of incompatible classification. A score of +1 is given if the classifications are compatible, and a score of 0 is given if the compatibility cannot be determined.

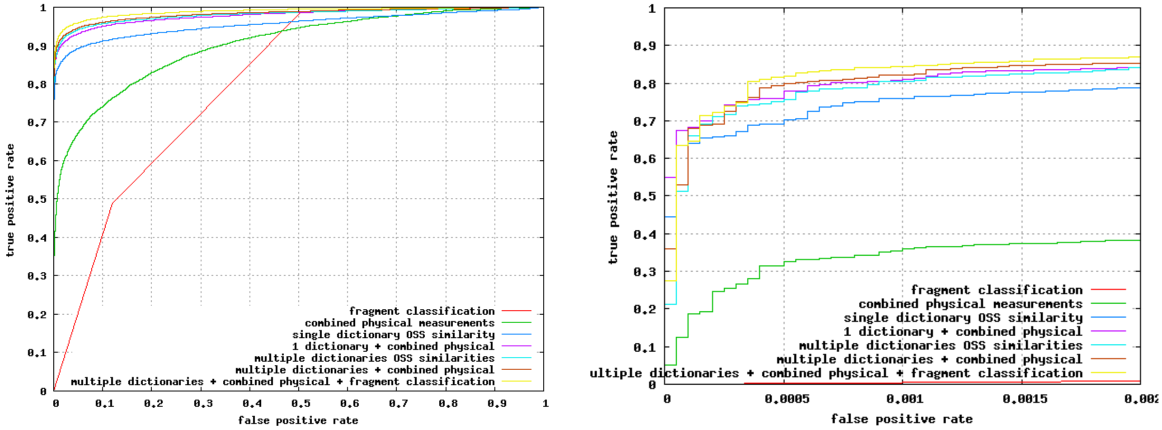
Physical measurements Eight measurements are considered: number of lines, average line height, standard deviation of line height, average space between lines, standard deviation of interline space, minimal bounding box width, minimal bounding box height, and area of the minimal bounding box. Each one of these measurements is hardly discriminative; however, combined together, they are able to discriminate pretty reliably between joins and random pairs, although not as well as the handwriting approach below.

Handwriting The handwriting is represented using the bag of visual keywords approach. Using a global dictionary, the best performing method uses the One-Shot-Similarity (OSS) of Section 5.1.

Multiple script type encoding The OSS scores obtained from the various dictionaries described in Section 4.4 are combined using the stacking technique of described in Section 5.2. This method provides a noticeable improvement over the single dictionary method.

Combination	Area under ROC	Equal error rate	Mean success ± Standard Error	TP rate at FP rate of 0.001
Domain classification	0.7932	0.3081	0.4935 ± 0.0042	0
Physical measurements	0.9838	0.0531	0.9735 ± 0.0070	0.8286
Single dictionary	0.9703	0.0725	0.9583 ± 0.0048	0.7276
Single dictionary + physical	0.9785	0.0627	0.9566 ± 0.0028	0.8116
Multiple script style dictionaries	0.9805	0.0564	0.9596 ± 0.0029	0.8053
Multiple dictionaries + physical	0.9830	0.0524	0.9625 ± 0.0028	0.8229
Multiple + physical + domain	0.9888	0.0430	0.9680 ± 0.0024	0.8451

Table 2: Results obtained for various similarity measures and combinations thereof. See text for the description of each method.



(a) (b)

Figure 4: ROC curves averaged over 10 folds. The plots compare the results obtained for the physical measurements (SVM similarity on the 8D vectors); baseline L2 normalized histograms with L2, SVM, and SVM+ similarity; L1 normalized histogram with OSS similarity; and the latter combined with the SVM similarity of the physical measurements. (a) Full ROC curves. (b) A zoom-in onto the low false positive region.

In addition, we combine the handwriting based scores (single or multiple dictionaries) with the physical score and with the domain-classification score.

The results are depicted in Table 2. It can be seen that the best method, the one that combines the multiple script-style dictionaries with the physical measurements and the domain classification, obtains a recall rate of up to 84.5% at a false positive rate of 0.1%. The obtained ROC curves are depicted in Figure 4(a). While some of the improvements seem incremental, they actually make a significant difference in the low-false positive region (Figure 4(b)).

6 Newly-Found Joins

We have conducted three sets of examinations to evaluate the value of our system in finding joins beyond the settings of the benchmark.

First, we applied the system in order to locate joins in the newly discovered Geneva collection. The search for joins using our tools was pretty efficient, with about 30% of the top 100 matches returned turning out to be actual joins. Figure 5 shows a variety of previously-unknown joins proposed by our method. Example (a) consists of two leaves from the same copy of the Mishnah, written on vellum in Hebrew in a square script. The texts are from different tractates of *Order Zeraim*. The left page is from the recently recovered Geneva collection and the right one from the small collection of the Jewish National and University Library. Other leaves from the same manuscript are in Oxford and Cambridge.[1] Example (b) shows fragments from a codex of the Bible, both from the book of Exodus (Hebrew, square script, on vellum), one from Geneva and the other from the Jewish Theological Seminary (JTS) in New York, part of a batch of 69 fragments from various biblical manuscripts (partially vocalized and with cantillation signs). Such codices are written using a very rigid set of typographic rules, and the identification of such joins based on handwriting is considered extremely challenging. Example (c) is in alternating Hebrew and Aramaic (*Targum*, square script), one page from Geneva and the other from the New York JTS collection. Example (d) shows a join of two leaves of Hebrew liturgical supplications from Geneva and from Pennsylvania, in rabbinic script. Example (e) is from a book of precepts by Saadiah ben Joseph al-Fayyūmi, a lost halakhic work by the 10th century gaon. The left page is from the Geneva collection and right one is from JTS. The language is Judeo-Arabic, and the text is written in a square oriental script on vellum. This join is a good example of how joins can help identify new fragments from lost works. Once one of the pair is identified correctly, the identification of the second one is self-determined. Example (f) is from a responsum in Hebrew (rabbinic script). Both leaves are from the Alliance Israélite Universelle Library in Paris, but they are catalogued under different shelfmarks.

A second set of outside-benchmark experiments was done on an initial benchmark we created for a preliminary work (Wolf, Littman, Mayer, Dershowitz, Shweka, and Choueka 2009). This benchmark was much smaller and contained ten splits each containing 196 positive pairs and 784 negative ones. All images were taken from the ENA and AIU collections. As mentioned above, the negative pairs we work with are not necessarily negative. This does not affect the numerical results much,

[1] It turns out that this specific automatically-proposed join has already been discovered and is documented in the recent Geneva catalog (Rosenthal 2010), and in the forthcoming Sussmann Catalog.

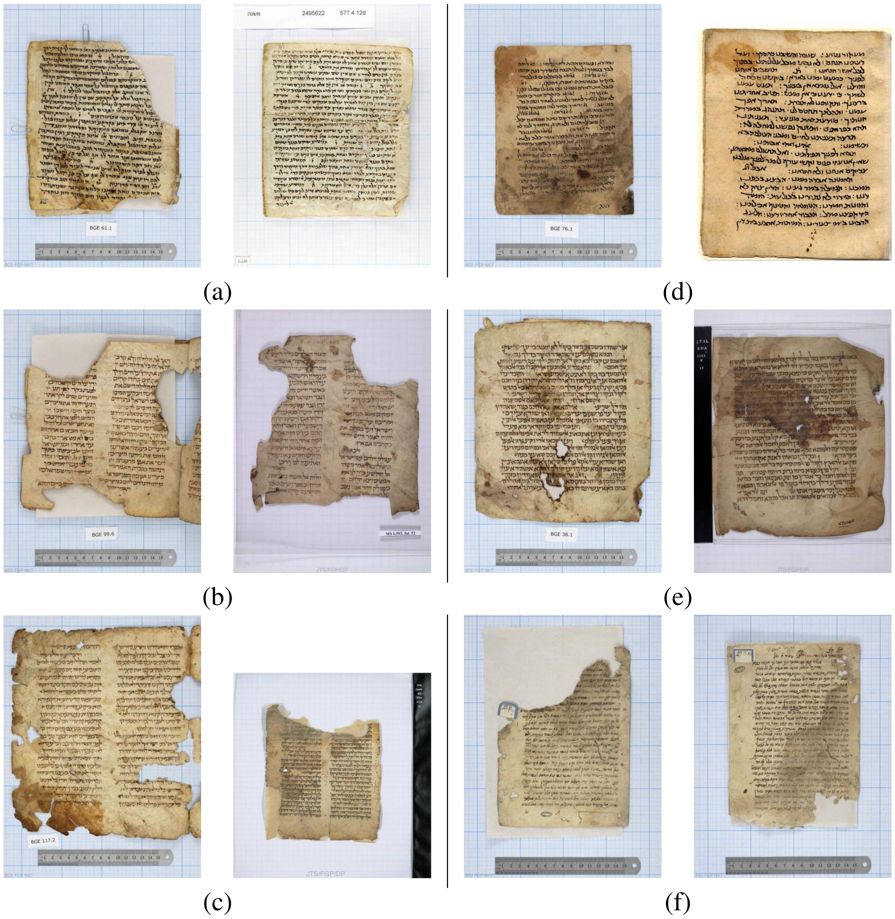

(a) (d)
(b) (e)
(c) (f)

Figure 5: Examples of unknown joins discovered by our system. See text for details.

since the fraction of joins is overall low; however, it implies that there may exist unknown joins in the set of leaves that are currently available to us.

We applied our classification technique to all possible pairs of leaves and then looked at the 30 leaf pairs that were not known to be joins, but which received the highest matching scores. The resulting pairs were submitted to a human expert for validation. The manual labor involved was about $2\frac{1}{2}$ hours. Eighty percent of the newly detected join candidates were actual joins. Seventeen percent are not joins, and one pair could not be determined.

A third set of join-seeking efforts was conducted on all between-collection pairs of fragments unknown to be joins in the ENA, AIU and JNUL collections, as well as in smaller European collections of mixed quality. Note that inter-collection joins are harder for humans to find, and are more challenging and rare. The top scoring 9,000 pairs were detected. After some further analysis of the catalogical information, some additional known pairs were removed resulting in 8,790 pairs. The first

Range	Strong join	Weak join	Total Join	Excluding emplty
1–2000	17.05%	6.95%	24.00%	44.8%
5791–8790	7.16%	6.20%	13.37%	18.0%

Table 3: The percentile of verified new joins out of the candidate joins suggested by our system.

2,000 pairs and the last 3,000 fragments of this list were studied. The results are given in Table 3. It separates between strong joins, meaning same scribe and same manuscript, and weak joins – a join between different manuscripts that seems to be written by the same scribe. In contrast to strong joins, the certainty of a weak join is doubtful, and in many cases should be examined carefully again by an expert. In any event, a join that was marked as weak always represents a good candidate for fragments written by the same scribe, and as such it is considered a success.

As can be seen, 24% of the top discoveries are true joins, mostly strong. More than 13% of the 6th, 7th, and 8th thousands of matches are validated joins. At least half of those are strong joins. Going over the examples, it became apparent that many of the proposed joins are artifacts caused by normalized vectors arising from empty documents. This was to be expected, since the benchmark that was used to develop the join-discovery tool was not designed to handle blank documents. After the removal of 49 empty fragments and all their discovered joins, the recognition rates grew considerably.

Overall, throug these efforts, approximately 1000 new joins were found. Given that the overall number of joins found in over a century of Genizah research and by hundreds of researchers is only a few thousands, our system proved its scalability and value. The main limiting factor in finding more joins is the availability of human experts. We hope to alleviate this constraint by making our join candidates available over the Internet to the Genizah research community.

7 Unsupervised Grouping by Script Style

As we have found, the most distinguishing visual information between the fragments arises from the handwriting. The search for joins focuses on minute differences that exist between various scribes. We now turn our attention into grouping the documents by a much coarser distinction: the one between script styles.

We sample 300 leaves from the Genizah collection that have been classified into one of 12 script styles: “Square Ashkenazi”, “Square Italian”, “Semi-cursive Oriental”, “Square Oriental”, “Cursive Oriental”, “Semi-cursive Spanish”, “Square Spanish”, “Cursive Spanish”, “Semi-cursive Yemenite”, “Square Yemenite”, “Square North-African”, “Cursive North-African”. We then attempt to group the leaves automatically, a process called “clustering”.

We found that conventional clustering algorithms such as k-means work poorly for separating the documents into script-styles. Indeed, k-means focuses on clusters of similar sizes, and might produce unintuitive results for data that is not distributed homogeneously in the parameter space.

We therefore employed the following method that was developed in order to deal with an unknown number of clusters, variability in cluster size, and inhomogeneous data.

First, each leaf is represented as a vector using the bag of visual keyword approach and a single global dictionary. Multiple dictionaries would not be appropriate here, since we would like to obtain the script styles from the data, and not impose it on the representation.

Recall that the vector representing each leaf contains visual “keyword” frequencies.To eliminate noise and remove spurious correlations between documents, we focus on the most prominent keywords for each document. This is done by replacing each keyword frequency that is less than half of the maximal frequency by zero.

In the next step, we build a graph in which every leaf is a node, and an edge exists between two nodes if the correlation between their modified vectors is above 0.5. The connected components of this graph are taken as the initial clusters. Connected components that contain single points are referred to below as “singletons” and are considered unclustered.

We then refine these clusters by iterating, until convergence, two alternating steps. In the first step, pairs of clusters for which the distances between each cluster’s points and the cluster mean point are similar to the distances between the two clusters are merged. In the second, singletons are assigned to clusters if their distance to the closest cluster is not larger than three times the standard deviation of distances within that cluster.

After convergence, this procedure yields 18 clusters and 34 singletons. The clusters are pretty homogenous with regard to script style: 93% of the documents are clustered within clusters in which their script-style is the most frequent script-style; 7% are clustered in clusters in which they belong to the minority.

The distribution of documents of various script styles among the 18 clusters is shown in the confusion matrix presented in Figure 6. Each row of this matrix corresponds to one script style, and each column to one cluster.

Figure 7 shows samples from representative clusters. As can be seen, confusion is often a result of script styles that are superficially similar. Naturally, a more detailed analysis of individual letters would lead to more accurate results; however, this requires accurate optical character recognition, which is beyond current state of the art for the vast majority of Genizah images.

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10	cluster 11	cluster 12	cluster 13	cluster 14	cluster 15	cluster 16	cluster 17	cluster 18	unclustered
Square Ashkenazi	0.00	0.00	0.00	0.33	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
Square Italian	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Semi-cursive Oriental	0.00	1.00	1.00	0.67	0.00	0.00	0.20	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
Square Oriental	0.00	0.00	0.00	0.00	0.64	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18
Cursive Oriental	0.00	0.00	0.00	0.00	0.04	0.00	0.80	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.03
Semi-cursive Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
Square Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
Cursive Spanish	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.15
Semi-cursive Yemenite	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Square Yemenite	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.06
Square North-African	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.09
Cursive North-African	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	1.00	0.00

Figure 6: A confusion matrix that shows that frequency within each obtained cluster of each script style. For example, the first cluster is composed entirely out of leaves of square Italian script style, while the forth cluster is one-third square Ashkenazi and two-thirds Semi-cursive Oriental.

8 Conclusion

We have presented a framework for identifying joins in Genizah fragments, which already provides results of value to Genizah researchers by identifying unknown joins. We have shown how handwriting data, especially when combined with prior knowledge of script styles, physical measurements, and domain classification, can produce a reliable system.

We also explored the grouping of Genizah documents in a top-down manner, and have shown that, when the heterogenous nature of the data set is accounted for, the paleographic information emerges as the most visually prominent characteristic.

References

Aiolli, F. and A. Ciula (2009). A case study on the system for paleographic inspections (spi): challenges and new developments. In *Proceeding of the 2009 conference on Computational Intelligence and Bioengineering*, Amsterdam, The Netherlands, The Netherlands, pp. 53–66. IOS Press.

Aussems, J. F. A. (2006). Christine de pizan and the scribal fingerprinta quantitative approach to manuscript studies. Utrecht, Masters Thesis.

Aussems, M. and A. Brink (2009). Digital palaeography. In *Malte Rehbein, Patrick Sahle & Torsten Schassan (eds), Palaeography and Codicology in the Digital Age*, pp. 293–308.

Baird, K. (1992). Anatomy of a versatile page reader. *Proceedings of the IEEE 80*(7).

Bar-Yosef, I., I. Beckman, K. Kedem, and I. Dinstein (2007). Binarization, character extraction, and writer identification of historical hebrew calligraphy documents. *Int. J. Doc. Anal. Recognit. 9*(2), 89–99.

Bar Yosef, I., K. Kedem, I. Dinstein, M. Beit-Arie, and E. Engel (2004). Classification of hebrew calligraphic handwriting styles: Preliminary results. In *DIAL '04: Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*, Washington, DC, USA, pp. 299. IEEE Computer Society.

Beit-Arie, M., E. Engel, and A. Yardeni (1987). *Specimens of Mediaeval Hebrew Scripts, Volume 1: Oriental and Yemenite Scripts (in Hebrew)*. Jerusalem: The Israel Academy of Sciences and Humanities.

Beit-Arie, M., E. Engel, and A. Yardeni (2002). *Specimens of Mediaeval Hebrew Scripts, Volume 2: Sefardic Script (in Hebrew)*. Jerusalem: The Israel Academy of Sciences and Humanities.

Bensefia, A., T. Paquet, and L. Heutte (2003). Information retrieval based writer identification. In *Int. Conf. on Document Analysis and Recognition*.

Bilenko, M., S. Basu, and R. Mooney (2004). Integrating constraints and metric learning in semi-supervised clustering. In *ICML*.

Bres, S., V. Eglin, and C. V. Auger (2006). Evaluation of Handwriting Similarities Using Hermite Transform. In *Frontiers in Handwriting Recognition*.

Brink, A. A., J. Smit, M. L. Bulacu, and L. R. B. Schomaker. Quill dynamics feature for writer identification in historical documents. Forthcoming.

Bulacu, M. and L. Schomaker (2007a). Automatic handwriting identification on medieval documents. In *Int. Conf. on Image Analysis and Processing*.

Bulacu, M. and L. Schomaker (2007b). Text-independent writer identification and verification using textural and allographic features. *IEEE Trans. Pattern Anal. Mach. Intell. 29*(4), 701–717.

Bulacu, M., L. Schomaker, and L. Vuurpijl (2003). Writer identification using edge-based directional features. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Washington, DC, USA, pp. 937–941. IEEE Computer Society.

Casey, R. G. and E. Lecolinet (1996). A survey of methods and strategies in character segmentation. *PAMI 18*.

Ciula, A. (2005). Digital palaeography: using the digital representation of medieval script to support palaeographic analysis. In *Digital Medievalist*.

Cristianini, N., J. Kandola, A. Elisseeff, and J. Shawe-Taylor (2002). On kernel-target alignment. In *NIPS*.

Dance, C., J. Willamowski, L. Fan, C. Bray, and G. Csurka (2004). Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*.

Dinstein, I. and Y. Shapira (1982a). Ancient hebraic handwriting identification with run-length histograms. *Systems, Man and Cybernetics, IEEE Transactions on*.

Dinstein, I. and Y. Shapira (1982b). Ancient hebraic handwriting identification with run-length histograms. *IEEE Trans. Syst. Man Cybern. 12*, 405–409.

Fischler, M. A. and R. C. Bolles (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM 24*(6).

Fournier, J.-M. and J. Vienot (1971). Fourier transform holograms used as matched filters in hebraic paleography. *Isr. J. Technol.*, 281–287.

Hertz, T., A. Bar-Hillel, and D. Weinshall (2004). Boosting margin based distance functions for clustering. In *ICML*.

Huang, G. B., M. Ramesh, T. Berg, and E. Learned-Miller (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. UMASS, TR 07-49.

Lazebnik, S., C. Schmid, and J. Ponce (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.

Leedham, G., S. Varma, A. Patankar, and V. Govindarayu (2002). Separating text and background in degraded document images. In *Frontiers in Handwriting Recognition*.

Lerner, H. G. and S. Jerchower (2006). The Penn/Cambridge Genizah fragment project: Issues in description, access, and reunification. *Cataloging & Classification Quarterly 42*(1).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*(2), 91–110.

Mallon, J. (1952). *Paléographie romaine*. Madrid: Consejo Superior de Investigaciones Científicas, Instituto Antonio de Nebrija, de Filología.

Panagopoulos, M., C. Papaodysseus, P. Rousopoulos, D. Dafi, and S. Tracy (2009). Automatic writer identification of ancient Greek inscriptions. *PAMI 31*(8).

Plamondon, R. and G. Lorette (1989). Automatic signature verification and writer identification – the state of the art. *Pattern Recognition 22*(2), 107 – 131.

Reif, S. C. (2000). *A Jewish Archive from Old Cairo: The History of Cambridge University's Genizah Collection*. Richmond, England: Curzon Press.

Rosenthal, D. (2010). *The Cairo Genizah Collection in Geneva Catalogue and Studies*. Jerusalem.

Said, H. E. S., T. N. Tan, and K. D. Baker (2000). Personal identification based on handwriting. *Pattern Recognition 33*(1), 149–160.

Serre, T., L. Wolf, and T. Poggio (2005). Object recognition with features inspired by visual cortex. In *CVPR*.

Shental, N., T. Hertz, D. Weinshall, and M. Pavel (2002). Adjustment learning and relevant component analysis. In *ECCV*.

Sirat, C. (1981). L'examen desécritures: l'oeil et la machine. Essai de méthodologie, Thesis, Paris, Editions du Centre National de la Recherche Scientifique.

Srihari, S. N. and V. Govindaraju (1989). Analysis of textual images using the Hough transform.

Weinberger, K., J. Blitzer, and L. Saul (2006). Distance metric learning for large margin nearest neighbor classification. *NIPS*.

Wolf, L., S. Bileschi, and E. Meyers (2006). Perception strategies in hierarchical vision systems. In *CVPR*.

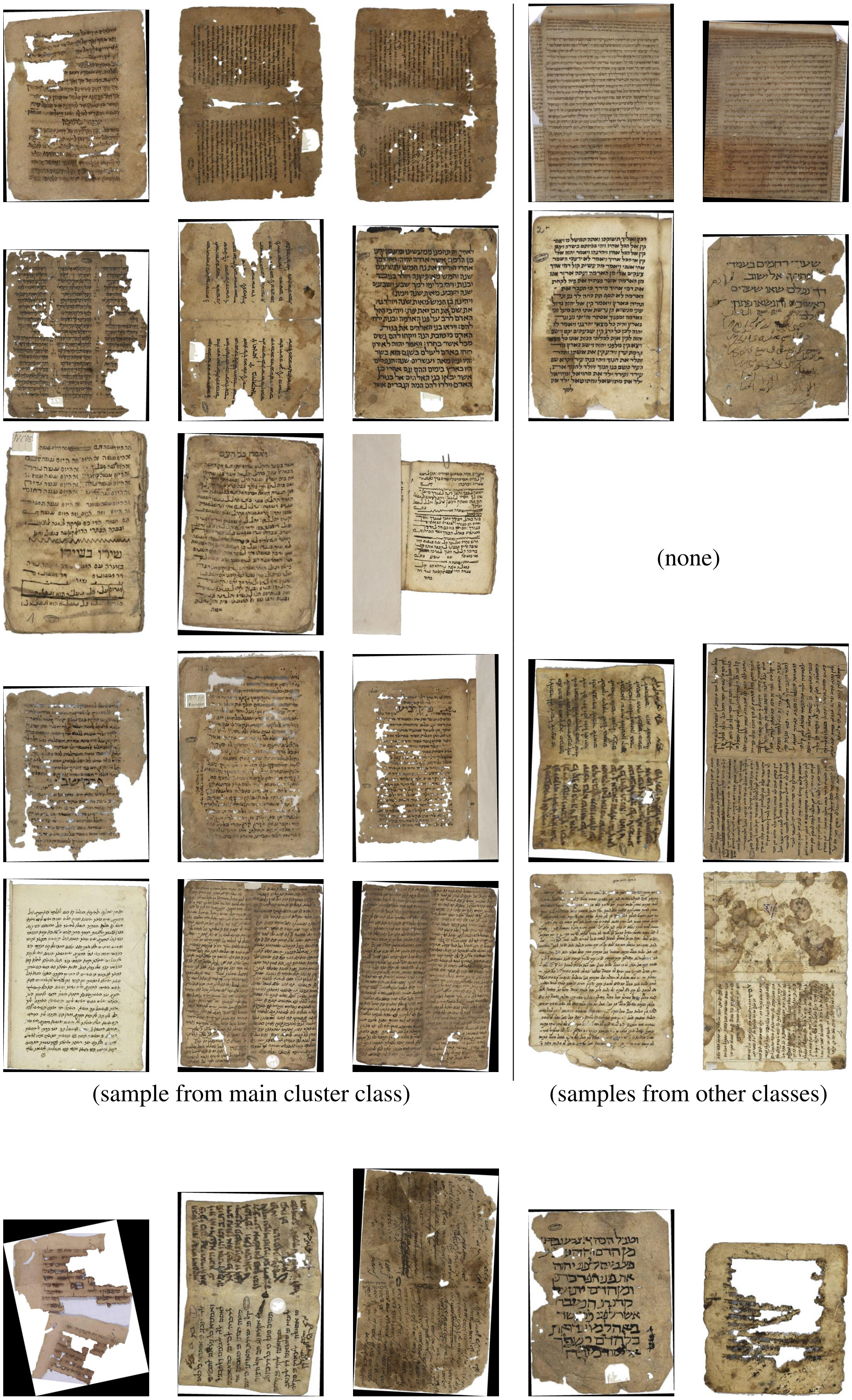
Wolf, L., T. Hassner, and Y. Taigman (2008). Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop in ECCV*.

Wolf, L., T. Hassner, and Y. Taigman (2009, Sept.). The one-shot similarity kernel. In *IEEE International Conference on Computer Vision (ICCV)*.

Wolf, L., R. Littman, N. Mayer, N. Dershowitz, R. Shweka, and Y. Choueka (2009, September). Automatically identifying join candidates in the cairo genizah. In *Post ICCV workshop on eHeritage and Digital Art Preservation*.

Wolpert, D. H. (1992). Stacked generalization. *Neural Netw. 5*(2), 241–259.

Xing, E., A. Y. Ng, M. Jordan, and S. Russell (2003). Distance metric learning, with application to clustering with side-information. In *NIPS*.



(none)

(sample from main cluster class)

(samples from other classes)

(samples of remaining singletons)

Figure 7: Each row shows samples from a single cluster. The left three samples are from the main script style of this cluster, while the two samples to the right are samples of the same cluster that belong to other script styles. Shown, from top to bottom are clusters 4, 5, 6, 8, 17. Also shown (last row) are samples of singletons, i.e., documents that were not assigned to any of the clusters.