# Automatic Extraction of Catalog Data from Digital Images of Historical Manuscripts

Roni Shweka, Yaacov Choueka[*]
The Friedberg Genizah Project
Jerusalem, Israel
Lior Wolf, Nachum Dershowitz
The Blavatnik School of Computer Science
Tel Aviv University
Ramat Aviv, Israel

**Abstract**

The Cairo Genizah is a collection of handwritten historical documents containing approximately 350,000 fragments of mainly Jewish texts discovered in the late 19th century. The fragments are today spread out in more than 70 libraries and private collections worldwide, and there is an ongoing effort to document and catalog all extant fragments.

We explore three levels of extraction of catalog data from digital images of the fragments. First, images should be captured in a way that permits standardized automatic processing. Second, the images can be processed to detect elements such as image foreground, regions of written text and lines of the text, thereby allowing for the automatic assignment of conventional catalog measurements. Third, modern computer-vision tools and statistical inference techniques may be employed to identify fragments that might originate from the same original codex. Such matched fragments, commonly referred to as "joins", were heretofore identified manually by experts, and presumably only a small fraction of existing joins have been discovered to date.

Overall, we present what might be the first effort to address all three levels successfully within a large-scale project, detailing the various design choices and describing the techniques and algorithms used for the Cairo Genizah digitization project.

[*] Professor Emeritus, Department of Computer Science, Bar-Ilan University, Ramat Gan.

# 1 Introduction

The Cairo Genizah (see e.g. (Reif and Reif, 2002) and the very recent (Glickman, 2010, Hoffman and Cole, 2010)) is a collection of handwritten historical documents containing approximately 350,000 manuscript-fragments of Jewish texts discovered in the late 19th century. Most of the fragments were written between the 10th and the 14th centuries, and almost all of them in Hebrew characters, mostly in the Hebrew, Judeo-Arabic and Aramaic languages. Today, the fragments are spread out in more than 70 libraries and private collections worldwide. The Friedberg Genizah Project (www.genizah.org), whose mission is to computerize the whole research world of the Genizah collection, is in the midst of a multi-year process of digitally photographing all of the extant fragments. As of July 2012, the project's virtual library holds over 400,000 digital images. Unfortunately, this huge and critically important collection for Jewish Studies and for the study of the cultural heritage of the Mediterranean societies in the Middle Ages, in general, is far from being entirely cataloged, despite the ongoing hundred-year-old effort to document and catalog all extant fragments. Moreover, existing catalogs vary greatly in the amount and type of data they incorporate. Many of them merely briefly record the content of the fragment, without any information regarding its physical attributes.

We present a system for automatically computing much of the traditional catalog data, as well as some additional interesting attributes not usually included, by extracting them from the fragment's digital image. These are, mainly: the exact dimensions of the fragment; number of columns; number of lines; size of the margins; and the fragment's physical status (torn vertically, horizontally or diagonally, or missing corners). Until now, finding these properties, most of which are expected to be found in any modern catalog, required tedious, time consuming and tiresome manual labor, which had to be done with the original manuscript in hand. Moreover, our system can extract some finer data that may be relevant to paleographic studies, such as density of lines (line height, inter-line space) and density of characters (number of characters in a fixed unit of width). This system is also able to differentiate between bifolios and folios (single pages), and in the former case collects the physical attributes for each page separately.

In addition to the detailed physical description of a single fragment, the huge database generated by the system serves for supporting identification of "join" candidates in the Cairo Genizah. A *join* is a set of manuscript-fragments that originate from the same original codex, but are separated today, and stored under different shelfmarks, possibly scattered in different libraries. In previous work (Wolf et al., 2011b), we described a system for the semi-automatic identification of joins by ascertaining the degree of handwriting similarity between pairs of fragments. By querying the database and applying some basic rules for a good match, taking into account the extracted physical attributes as well as the completeness or incompleteness of the fragments, we can significantly improve on the quality of the results obtained by only analyzing similarity of handwriting.

One of the major aims of this paper is to propose appropriate conditions for taking digital images of manuscripts, which are necessary for achieving this kind of results. We argue that, today, the function of digital imaging is not only conservation and accessibility, but also to enable these images to serve potentially as inputs to artificial-intelligence algorithms and processes, and the computer should be taken into account as one of the "clients" of the images. Hence, proper conditions should be considered in advance when digitizing manuscripts; when such conditions are neglected, the application of computerized methods and harvesting their results become unnecessarily difficult, and the quality of obtained results is adversely affected. These conditions are detailed in the next section. Section 3 describes—in very general terms—the succession of various processes that have to be applied to digital images of manuscripts in order to compute the above-mentioned attributes. In Section 4, the system for automatically suggesting potential joins is described. This is followed by a brief conclusion.

## 2 Capturing Manuscript Images with an AI Eye in Mind

One of the main purposes of this paper is to detail the proper conditions for taking digital images of manuscripts, necessary for achieving the kind of results described in the introduction, viz. automatic extraction of cataloging and other useful data from the fragment's image, supporting a system for the automatic suggestion of possible joins in the collection, and allowing for automated exploration of writer-identification and digital-paleography issues. It is not claimed here that such procedures are necessary and pertinent to all projects of document digitization; if the case at hand is that of a number of well-preserved and well-organized handwritten codices, then most of the automatic measurements detailed here may be unnecessary (getting them from just one typical page of the codex would be sufficient) and join issues are not relevant. Still, it is claimed that the needs of the computer as a user of the images should be taken into account, if not for current needs, then for future and innovative ideas (such as OCR), and, on the other hand, that the procedures suggested here can be of great value for cases similar to that of the Cairo Genizah, namely, large sets of individual, mutilated unattached fragments, such as the Dunhuang collection (idp.bl.uk), or the Papyri collections. (For a comprehensive list of papyri collections, see www.trismegistos.org/coll/list_all.php; for an example of such a digitized collection, see http://www.uni-heidelberg.de/fakultaeten/philosophie/zaw/papy.)

Digitizing collections of historical manuscripts, a flurry of activity happily flourishing in the last few years, is usually justified by assessing the usefulness of the three functions of digitization: conservation, accessibility and manipulability. This activity is accompanied by the compilation of practical guides and technical papers that present best practices for digitizing textual documents, manuscripts and print (FADGI, 2009), such as the Federal Agencies Digitization Initiative Technical guidelines (FADGI, 2010). Indeed, most of this literature is directed solely toward the objectives mentioned above. We consider these functions in turn.

## 2.1 Conservation

More than once have depots of manuscripts in major libraries around the world been threatened by natural disasters such as floods, fires and earthquakes, and more than once have such libraries actually lost some of their precious holdings this way. High-quality digital images are considered today to be an adequate replacement of the originals, at least for practical research purposes.

By “high-quality”, we mean images taken by a professional photographer, in a full-color, 24-bit depth mode, at high resolution (see below), in a lossless (TIFF) format, taking care of all lighting (cold light), reflection and shadow issues. This would be considered today as the minimal standard for acceptable replacement. For very special collections, images are expected to be obtained in many different spectra, including, for example, infrared and ultraviolet, as is in fact is being done currently for the Dead Sea Scroll collection at the Israel Antiquities Authority.

Digital images can be easily replicated and deposited in various locations, thus assuring a reasonable—if somewhat imperfect—sustainability of items of importance. For the record, we wish to say that this new conservation approach is not commensurate with the older ones of making microfiches or microfilms available. We do not go here into detailed reasons why, but these should become clear to the reader from the details that follow.

Regarding resolution, the rule of thumb in vogue in the beginning of this century, and still valid today, has been that a resolution of 400 or 600 dpi is adequate for conservation of textual documents, and that, in most applications, not much is to be gained from higher resolutions. Most of the Genizah collections where indeed digitized according to our requirements with 600 dpi. In a few collections the resolution—for technical reasons—was somewhat lower, but in any case it was (virtually always) no less than 450 dpi.

### 2.1.1 DPI

A word of clarification about the exact meaning in our context of this popular, constantly used, but very often misused and improperly understood unit, *dpi*. DPI (= dots per inch) is a unit appropriate in principle for printouts (as produced by a printer), assessing the printout attribute of having 600 dots of ink per linear inch of paper. The unit is also adequate for scanners, assessing that the scan of a document was done with 600 lighting-samples per inch. Borrowing it for digital imaging, however, is problematic, because of the extra parameter of the distance of the camera from the object. So, the proper unit in this context should be SPI (= samples per inch), denoting the number of “dots” *per real inch of the original document* sampled by the digital camera when capturing the document. The numbers usually assessed for a given image by image processors such as Photoshop, even those supposedly assessed by the camera manufacturers, are inadequate for our purposes. They usually represent a recommended conversion factor for printouts, and do not represent the real resolution of the captured document as defined above.

Realizing a fragment's image with 600 SPI (or DPI) resolution can be achieved as follows (similar computations are valid for any other resolution). Every digital camera comes with a digital "back" or board, consisting of a rectangular array of $X \times Y$ light-cells, as specified by the manufacturer. Let $X/600 = A$ and $Y/600 = B$. Take a rectangular background of $A \times B$ inches. Any fragment that fits completely on it and is shot by the camera at a 1:1 ratio (meaning, the entire background fits into the camera viewer exactly) has been indeed captured with a 600 SPI resolution. (Larger fragments have to be photographed in parts, but we'll not discuss this issue here.)

Given a digital image, how can we assess its real resolution? Assuming a ruler is part of the image, we can delimit an exact inch on this ruler, when the image is viewed in its original size (sans magnification or compression). The number of pixels in that inch (which can be determined by a pixel ruler) is the resolution of the image.

## 2.2 Accessibility

Digitizing a collection of manuscripts and having the images freely available over the Internet for every interested user saves him or her the trouble of traveling to the library where the manuscript resides, or of using an inert microfilm version that can barely be manipulated. It gives one immediate access to a true replica of the manuscript from anywhere and at any time.

## 2.3 Manipulability

Looking at a digital image of a manuscript rather than directly at the original fragment gives the user access to a rich set of important processing capabilities that can greatly help in "reading" the manuscript (an especially tricky task for historical manuscripts usually damaged by age), such as magnifying, rotating, reversing (white on black instead of black on white), mirroring and manipulating contrast, brightness and luminosity of an image in order to obtain the best possible readability conditions.

## 2.4 Image Processing

We argue that, today, an additional objective presents itself, that of digitizing a manuscript so as to have the resulting image serve as a potential input to artificial-intelligence algorithms and image-processing processes, which can greatly benefit the analysis and research of that manuscript. Thus, the computer must be taken into account from now on as one of the "clients" of the imaging process, and proper conditions should therefore be considered in advance when digitizing manuscripts to make the computer's tasks efficient and effective, and, in fact, even possible.

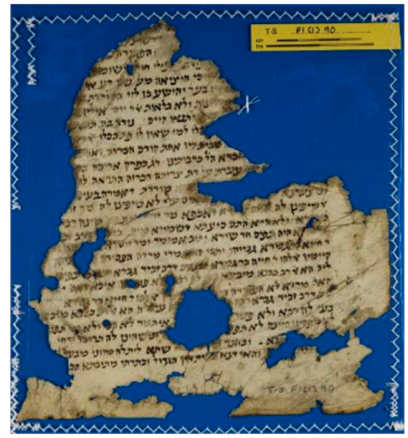
The main conditions are now described.

### 2.4.1 Background

A fragment to be photographed is usually placed on a fixed background of a certain color, this background serving as the common one for all fragments in the collection. Taking into account, however, that the first process of computerized analysis of an image in fact includes separating the fragment from its background, it becomes evident that the background color to be chosen should be the most contrasting one with the fragment color, both in terms of its material as well as its ink, so as to make the foreground-background separation task both efficient and precise. In Fig. 1(a), the background color used is very close to that of the foreground, making it almost impossible for the computer to distinguish between the two. Having a black background is also not recommended, because in this case the computer would not be able to distinguish between characters written in black ink and holes in the fragment through which the black background shows. Thus, the common practice in some libraries of digitizing on a white, cream, brown or black background should be considered not ideal, because these colors do not contrast well with that of the manuscript material and its ink.

Sampling a large number of fragments in different points, we found the average color of these fragments in terms of the RGB scheme to be (255, 136, 0). We thus concluded that the most contrasting color would be (0, 120, 255), which is a shade of blue, and this is indeed our recommendation. A librarian might argue that, while such a color would indeed make life easy for the computer, it would alienate ordinary users who would find it bizarre and unattractive. Our rejoinder is however quite simple: since, with this background-color choice, the computer can easily and exactly isolate the background, the system can change it, pixel by pixel, to any color desired, from off-white to black, through brown or rose, displaying the image to the user with any desired background color and texture.

(a)



(b)

Figure 1: (a) Poor contrast (Geneva) vs. (b) ideal contrast (Cambridge).

When the project of digitizing the huge Genizah collection at Cambridge University Library was started, we decided to use this shade of blue as the standard background for all images, with excellent results. The same practice was followed in the digitization of the Genizah collection at the British Library in London. See Fig. 1(b). Still, the images are presented on the Genizah website, following these libraries' requests, with a background in a shade gray that was carefully chosen and approved by them. The outcome is recognized as suitable by all parties.

### 2.4.2 Ruler

Including a ruler in the image is necessary, as explained above, to assess the real resolution of the image, and, in fact, for calibrating it. An alternative might be to use a background with a grid in inches (or centimeters); such material, however, usually comes in a very light color (not to interfere with the fragment's image), and so should be avoided because of the background color issue noted above.

Having a ruler in the image is crucial, especially in cases when different images are taken with different lenses or with the camera not fixed in the same position throughout the entire shooting process.

The ruler should be clearly distinctive from the fragment. Hence, a brown wooden ruler or see-through plastic one should be avoided. Rather, a metallic ruler is recommended. Also, it is recommended that the ruler be short, so that it can fit in its entirety in the image. If this is not possible, care should be taken to always align the starting end of the ruler with the left edge of the background.

### 2.4.3 Artifacts

The use of clips, weights and notes, as in Fig. 2(a), should also be avoided. For a proper analysis of the image by the computer, every significant element in the image should be identified and be easily recognizable, and the best segmentation is achieved by color separation. If such extra elements are unavoidable, we recommend that they be of the same distinctive color as the background (i.e. the special blue defined above). Our recommendation was followed in the British Library digitization project, as shown in Fig. 2(b).

Textual notes (such as shelfmarks) should be of a fixed size and shape, preferably with an easily recognizable icon, so as to enable software to recognize them easily.

In summary, taking care of the computer needs when digitizing will pay for itself handsomely by enabling the computer to supply us automatically with much useful data and with intelligent suggestions.


(a) (b)

Figure 2: (a) Fragment from the Strasbourg collection with label, clip and weight bags vs. (b) one from the British Library using a contrasting color.

# 3 Extracting a Fragment's Physical Attributes from its Digital Image

We now consider the successive steps of image analysis to be applied to a fragment's digital image so as to discover the fragment's physical attributes, as described above.

Incidentally, it should be mentioned that, although it is recommended to capture the images of fragments with a 600 dpi resolution (in the sense specified above), this was mainly for the purposes of conservation and manipulability. For measurement purposes, we found, after several experiments, that a resolution of 150 dpi is sufficient. Compressing the original image to 150 dpi greatly reduces the memory requirements and CPU resources, and reduces by several orders of magnitude the time needed for processing images.

The image-processing pipeline begins with a series of pre-processing steps by which the original image is transformed into a derived image suitable for the analysis stage in which the fragment's measurements are extracted from that derived image. The technical aspects of the process were described in previous work (Wolf et al., 2011b). Here we content ourselves with a high-level description of the various steps and their output.

## 3.1 Preprocessing

**A. Computing image dpi.** Finding the exact dpi of the image in the above sense is achieved by using the ruler included in every image. A reference image of the ruler was photographed independently, and used for locating a similar ruler in each image. The identification is done by employing a randomized algorithm, RANSAC (Fischler and Bolles, 1981), in combination with scale-invariant feature transform (SIFT) keypoint matching (Lowe, 2004). This template-based approach identifies and localizes the specific object (the ruler) and does not just remove it. For the latter task, a classical method such as the binarization method of (Sauvola and Pietikäinen, 2000) that also contains a method for identifying textual regions could be used.

Once identified, we can measure the ruler's size in pixels (assuming it is totally contained in the image) and divide it by its known size. When only part of the ruler appears in the image, we can still get the required result, either by using the distance between the keypoints matched in relation to the reference image, or by detecting two consecutive ticks and measuring the distance between them.

In some of the collections the fragment images were captured without a ruler, but on graph paper. Detecting the grid and counting the number of pixels between the lines (and knowing of course beforehand the distance between the grid's lines) proved to provide an accurate value for the image dpi.

For evaluating the results obtained by these two methods, we analyzed the AIU (Paris) Genizah collection in which both methods were available, that is, images in this collection were photographed with a background of blue graph paper and, in addition, a steel ruler was present in every image. The collection contains 13,230 images, most of them captured at 450 dpi. Table 1a lists the dpi values that were extracted in this collection by identifying the ruler, while Table 1b lists the dpi values extracted by identifying the graph paper. As can be seen, there is only a tiny deviation of 1% in the results obtained by the two methods (6 pixels at 450 dpi). While the identification of the graph paper was fully successful, the identification of the ruler failed (dpi=0) for 255 images (about 2%), due to a missing ruler or to a ruler partly covered by the fragment.

Processing images in other collections, which were taken at 600 dpi, produces values in the range of 595-603 dpi, thereby providing an accuracy of 0.2 mm in one inch (5 pixels in a 600 dpi image).

| DPI | 0 | 130-172 | 178 | 208-274 | 441-453 | 592-596 | Total |
|---|---|---|---|---|---|---|---|
| Count | 255 | 80 | 389 | 13 | 12202 | 291 | 13230 |

Table 1a DPI as extracted from ruler

| | 0 | 144-174 | 182 | 208-276 | 450-453 | 604 | Total |
|---|---|---|---|---|---|---|---|
| Count | 1 | 79 | 389 | 13 | 12457 | 291 | 13230 |

Table 1b DPI as extracted from graph paper

**B. Separating foreground from background.** The goal of this step is to detect the image of the fragment itself, isolating it from the accompanying background.

The process of separating fragment(s)—there may be many small separate fragments in the image—from the background in the image depends solely on color separation. It is therefore crucial to have a good distinction between the color of the background and that of the fragment. As described above, the collections in Cambridge and at the British Library were indeed taken on the blue background recommended above, which gives excellent separation. Other collections were taken on graph paper in which the grid lines were colored in a light hue of blue. In either case, an automatic classifier was first applied to identify foreground pixels (in contrast to background ones) based on RGB color values (or HSV values). To create a region-based segmentation of the fragment(s), the connected components of the detected foreground pixels were marked, and the convex hull of each component calculated (connected component = a contiguous region of foreground pixels; convex hull = the smallest possible encompassing polygon with angles opening inward). These procedures retain almost all of the relevant parts of the images, while excluding most of the background.

**C. Detecting and removing irrelevant components.** In some collections, each fragment is held in a slip attached to a binder; see Fig. 3. The system detected these binders by the combination of their color and shape, and removed them from the image. In other collections, images included a label with the fragment's shelfmark. These labels were also detected by the system and ignored. This was the case with the large Cambridge Genizah collection in which we have processed 309,480 images with labels. Out of this large number our algorithm failed to find the label in only 11,084 images (0.3%). By contrast, the detection of clips and weight bags was much more challenging and quite problematic. Luckily, such practices were rare, so we solved the problem with semi-automated tools, augmented by some manual labor.


Figure 3: Fragment from the JTS collection with a black binder.

**D. Separating multi-fragment images into components.** In many cases, more than one fragment was captured in a single image; see Fig. 4. In the AIU collection, for instance, out of 13230 images, 2224 (16.8%) were detected as having more than one fragment and 98 (0.7%) were detected as having more than five. Each such fragment (a "component" of the image) was identified and given a unique identifier (serial number) and, subsequently, handled independently of other components in the image. In such cases, however, there was a need to relate the components in the recto image of a fragment to the ones in its verso image (so as to have the same identifiers for both images). This was done automatically by mirroring one image and matching the components in both images by size and shape.

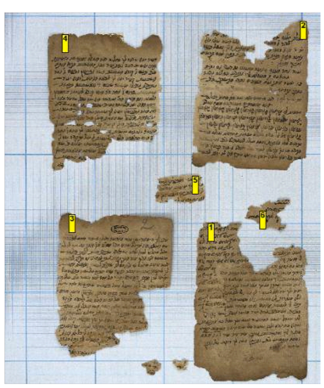


Figure 4: Multiple fragments in one image.

**E. Binarization.** The regions detected in the foreground segmentation process are then binarized, that is, every ink pixel is assigned a value of 1 (representing black), and all other pixels are assigned a value of 0 (for white). This is done using the auto-binarization tool of the ImageXpress 9.0 package by Accusoft Pegasus. To cope with failures of the Pegasus binarization, we binarized the images a second time using a 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 those two binarization steps. Pixels near the fragment boundary are set to 0. More sophisticated binarization methods, such as (Bar-Yosef et al., 2007), are being experimented with.

**F. Auto-alignment.** Although in most cases fragments were imaged placed upright, in many other cases the fragment was tilted. This depended on how the fragments were conserved: in Mylar envelopes, bound in volumes or in loose storage. The need for alignment is two-fold: first, to enable the correct measurement of the fragment's various attributes (such as width and length), and, second, to enable proper application of the handwriting-matching algorithm described below.

Alignment is achieved by rotating the image until the lines of text are horizontal, using a simple method akin to those in (Baird, 1992, Srihari and Govindaraju, 1989). For each possible rotation angle, we consider the ratio of black to white pixels in each horizontal line. We then calculate the variance of the projection for each angle, and select the angle for which the variance is maximal. This method may fail in two ways. If the fragment was originally written diagonally or in uneven lines, the system will fail to straighten the fragment, and might even undo what was originally a correct setup. Though this will result in erroneous measurements, it will still be beneficial for the handwriting matching algorithm. Furthermore, our algorithm might not detect an upside-down image. Although this will not affect the proper measurement of the fragment, it will hinder the handwriting matching algorithm. In principle, were we able to identify the script style, the algorithm could determine the correct orientation. Unfortunately, the diversity of handwritings in the Genizah material is so rich and variable that we are as yet unable to achieve this goal, but do hope to in the near future.

To evaluate the quality of the auto-alignment algorithm, we inspected the distribution of projection values in the chosen angle. In well-aligned documents, we expect the standard deviation of this distribution to be high.

On the other hand, when the distribution of projection values tend to be flat, we assume that the alignment is of poor quality. Using one standard deviation as the threshold, we mark each fragment as either having a good line-alignment or not.

We applied this method to the totality of the 469,940 processed fragments. Of these, 323,396 (68%) fragments were dsignated as being well-aligned, while 146,544 (31%) were marked as having poor line-alignment. This high failure rate has several causes: fragments with a blank side, tiny fragments with only a few legible characters, fragments in poor physical condition (mutilated or badly damaged) and fragments with diagonally-written lines. All of these contributing factors (which occur abundantly in the Genizah collection) yield poor binarization and, consequently, poor line-alignment.

## 3.2 Physical measurements

Having accurately determined the exact dpi of an image, we can measure different attributes of the fragment with very high accuracy. The most obvious ones are the dimensions of the fragment, which are recorded (manually) in most Genizah catalogs (but these catalogs cover, as noted, only a very small fraction of the Genizah collections). Few catalogs also record the inner dimensions of the fragment, that is, the dimensions of the written part. Besides these, our system also records the exact width of all four margins.

Needless to say, these measurements should be considered valid only for complete and intact folios. In the more common case for the Genizah of damaged pages, these figures should be considered as lower bounds only. There is a need, therefore, to record the state of the fragment: whether complete or damaged, and in the later case, in what directions(s) the damage occured (horizontal, vertical or both). The system evaluates the state of the fragment by its shape and by the appearance of the margins. The number of lines in the fragment, commonly recorded in many catalogs, is also computed. Note, however, that, while catalogs usually give the number of lines for one sample page from each shelfmark (which might contain several pages), our system records this value for every page, recto and verso.

Obviously, the quality of all these automatically-generated values largely depends on the state of preservation of the fragment. Stained, very dark or faded fragments and otherwise poorly preserved ones yield noisy or poor binarization, which will result in inaccurate measures and values. Therefore, a module for the evaluation of data quality is applied to all available images. This is performed by inspecting several factors, including the noise level in the binary image, the homogeneity of the text portion and the deviation of the obtained results from typical values.

Another method for evaluating the measurements' quality is comparing the measurements' values for the recto and the verso sides of the same document. A certain difference between the values of the measurements for the two sides of a fragment is unavoidable, as these measurements depend on the fragment's line-alignment quality, which is calculated separately for each side. Slightly different angles of the lines in the opposite sides of the original document will yield different alignment of the two images and therefore different measurements of the same document. Moreover, poor quality line-alignment on one side may result in incorrect alignment of the fragment and erroneous dimensions measures.

Analyzing 3654 documents from the AIU collection with only one fragment per image, we found that 89% of them had less than a 10% difference in the width dimensions. We consider these documents to have good auto-alignment and sufficiently accurate measurements.

Our system also offers a graphic tool for the website user that draws the derived rectangles bounding the fragment and the textual portion on top of the original image, providing a visual indication of the validity of the inferred measures. Another function marks every detected textual line in the fragment, allowing for manual indication of lines that the system may have skipped or marked by mistake. See Fig. 5. These drawings can be toggled on and off, so they need not interfere with the readability of the fragment.

The system also records some finer attributes, which are not to be found in catalogues, but are of great importance for matching joins, including average line height, average inter-line space, and average density of characters. These attributes may characterize a manuscript and can contribute to the join module, described next.

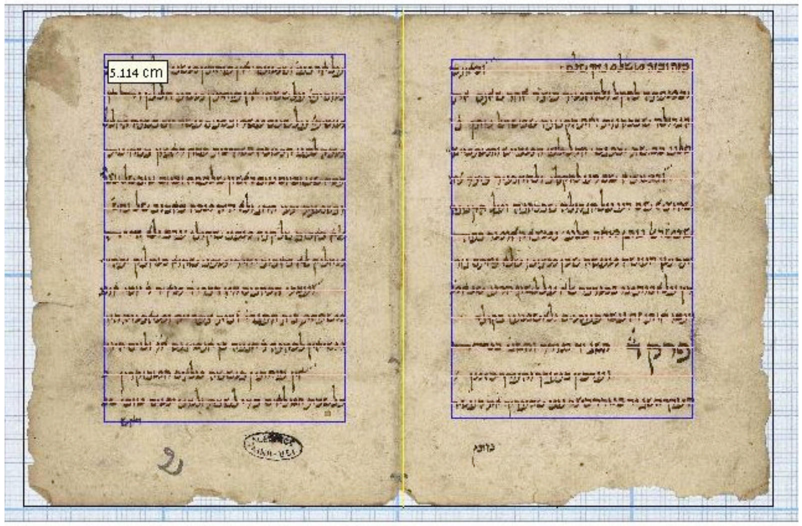



Figure 5: Bounding rectangles and medial lines, representing derived measurements for a fragment.

## 4 An Automatic System for Discovering Joins

One of the most critical issues in Genizah research is that of discovering "joins", that is, different fragments originating from the same codex that have been relocated in different locations (through the unavoidable deterioration of the originals over so many centuries and the random acquisition and trade of manuscripts), one fragment being found, for instance, in Cambridge and another in Vienna. Over the past hundred years, a few thousand joins have been discovered manually, by the sheer erudition, memory and intelligence of scholars. Can a computerized system help solve this problem today?

It is clear that if the images of two fragments are to be declared a join, then their handwriting should be similar, since both fragments were in all probability written by the same scribe. Hence the need to develop a computerized system that can assess—within a given probability—that two images of handwritten material bear similar handwriting, and were presumably written by the same hand, and thus constitute a potential join. In addition to comparing handwriting, the join-discovery system should take into consideration the appropriate cataloging data, as extracted from the images in the system described above.

Having successfully developed such a join-suggestion system proves the usefulness of making digital images of manuscripts available, as well as the usefulness of a computerized system for extracting catalog data from these images.

## 4.1 Similar Handwriting

For determining similarity of handwriting, we employed a general framework for image representation that has been shown to excel in domains far removed from document processing, namely, that of face recognition, adopting a method based on a "bag of visual keywords" (Dance et al., 2004, Lazebnik et al., 2006). The technical details of our system are provided elsewhere (Wolf et al., 2011b). Here, we provide only a non-technical high-level description.

In this approach, we do not analyze the individual handwritten letters and their shapes, but rather use a global comparison scheme, vaguely similar to the way with which two portraits are compared by the computer and found to be portraits of the same person. First we compute for each image a characteristic "signature", which is then converted into a numerical vector. To determine whether the handwriting of the two images is similar, their signatures are compared. This comparison employs a learned metric, that is, we use a training set of known joins to estimate the parameters of a similarity function that describes how likely any two images are to be a join, given their signatures.

We have conducted three experiments to evaluate the usefulness of our system for finding joins. These evaluations have produced a long list of brand new joins, never before recorded in Genizah research, which have just been published for a scholarly audience (Shweka et al., 2011).

## 4.2 Benchmarks

### 4.2.1 A First Small Benchmark

A set of experiments was performed on an initial benchmark we created (Wolf et al., 2009), with all images taken from the JTS (Jewish Theological Seminary in New York) and AIU (Alliance Israélite Universelle in Paris) collections. We compared all possible pairs of images from these two collections and submitted the 30 pairs that received the highest scores, but were not already known to be joins, to a human expert for validation, which took a couple of hours. Eighty percent of the newly detected candidates were found to be actual joins, 17% were found to be non-joins, and the status of one pair could not be readily determined.

#### 4.2.2 Benchmark with the Geneva Collection

We then asked our system to find joins with the recently recovered Geneva collection, which is characterized by mostly large, neat, clear and quite well-conserved folios. The search using our tools was pretty efficient, with about 30% of the top 100 matches turning out to be joins. Fig. 6 shows a variety of previously-unknown joins proposed by our algorithm (the leaf from Geneva is in each case is on the left). Example (a) consists of two leaves from the same copy of the Mishnah (Hebrew, square script, on vellum), with the right one being from the small collection of the National Library (NL) in Jerusalem (additional leaves from the same manuscript are in Oxford and Cambridge). Example (b) shows fragments from a codex of the Bible (Hebrew, square script, on vellum), with the right fragment from JTS. Such codices are written using a very rigid set of calligraphic rules, and the identification of such joins based on handwriting is considered extremely challenging. Example (c) is from a codex of alternating Hebrew and Aramaic text (square script), the right-hand one from JTS. Example (d) shows a join of two leaves of Hebrew liturgical supplications (rabbinic script), the second one from Pennsylvania. Example (e) is from a book of precepts by Saadiah Gaon, a lost halakhic work by the 10th century head of the Academy in Sura (Judeo-Arabic, square oriental script, on vellum), the right one from JTS. This is a good example of how joins can help identify new fragments from lost works. Once one is identified correctly, the identification of the other is automatically determined. Example (f) is from a Hebrew responsum (rabbinic script), where both leaves are from AIU, but given different shelfmarks.

#### 4.2.3 Benchmark of Joins between Collections

A third set of join-seeking efforts was conducted on all between-collection pairs of fragments unknown to be joins in ENA, AIU, NL and smaller European collections of mixed quality. Note that inter-collection joins are harder and more challenging for scholars to find manually. The top-scoring 9,000 pairs were extracted and then reduced, for practical reasons, to 8,790 pairs. The first 2,000 pairs and the last 3,000 fragments of this list were studied. The results are given in Table 1. The columns distinguish between "strong" joins, meaning the same scribe and the same manuscript, and "scribal" joins—a join between different manuscripts that appear to be written by the same scribe. The latter are also of potential interest to scholars and are considered a successful hit. 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, and at least half of those are strong. Going over the examples, it became apparent that many of the proposed joins were artifacts caused by normalized vectors arising from blank pages. 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 such pages and all their supposed joins, the recognition rates increased considerably.

It should be added that any Genizah scholar would be extremely happy to check a list of, say, a hundred pairs, finding maybe only one pair to be a true join, since his chances of finding this join by himself are practically nil.


Figure 6: Examples of heretofore unknown joins discovered by the system. See text for details.

| Non-blank | Total join | Strong join | Scribal join | Range |
|---|---|---|---|---|
| 45% | 24% | 17% | 7% | 1–2000 |
| 18% | 13% | 7% | 6% | 5791–8790 |

Table 2: Percentage of verified new joins out of candidate joins suggested by the system.

#### 4.2.4 Computational Benchmark

To facilitate collaboration between researchers and to easily evaluate the contribution of various components of our system, we also devised an accessible benchmark that can be utilized without the need of expert evaluation. This benchmark is modeled after the widely popular face-recognition benchmark called "Labeled Faced in the Wild" (Huang et al., 2007).

Our benchmark consists of 31,315 leaves, from the New York (ENA), Paris (AIU), and Jerusalem (JNUL) collections. It consists of “positive pairs”, each consisting of two fragments known to be from the same join, and of negative pairs, which are pairs that are not known to be positive. Since the expected number of unknown joins is very limited in comparison to the total number of pairs, the vast majority of the negative pairs are believed to be non-joins.

There are two views of the dataset: View 1, which is meant for parameter tuning, and View 2, meant for reporting results. View 1 contains three splits, each containing 1000 positive pairs of leaves belonging each to the same join, and 2000 negative pairs of leaves that are not known to belong to the same join. When working on View 1, one trains on two splits and tests on the third.

View 2 of the benchmark consists of ten equally sized sets. Each also contains 1000 positive pairs of images taken from the same joins, and 2000 negative 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 on View 2, one repeats the classification process ten times. In each iteration, nine sets are taken as training, and the results are evaluated on the tenth set. Results are reported by constructing a receiver operating characteristic (ROC) curve for all splits together (the outcome value for each pair is computed when this pair is a testing pair) and by computing statistics of the ROC curve (area under curve and true positive rate at a certain low false positive rate).

By using handwriting similarity alone, we are able to show high recognition rates on this benchmark: the obtained area under curve is 0.98, and the recall at the false positive rate of 0.001 is as high as 79%.

## 4.3 Incorporating Catalog Information with Handwriting Similarity to Suggest Joins

We found that the most distinguishing visual information between fragments arises from the handwriting, and the search for joins focuses on minute differences that exist between various scribes, as described above. However, other sources of information are also valuable for finding joins. Applied in tandem with the handwriting similarity, these can help disambiguate difficult cases and improve the overall accuracy.

The physical measurements, the extraction of which was described in Section 3.2, are highly indicative for finding joins. Eight measurements are considered: number of lines, average line height, standard deviation of line height, average space between lines, standard deviation of interline space, and the inner dimensions of the fragment: height, width, and area.

In order to verify that two documents are potential joins, we compare the measurements of the two documents. In good condition, two leaves of the same manuscript would produce similar measurememtns. The Genizah, as described above, however, contains highly degraded documents. Therefore, joins display vastly different measurements in many cases. Conversely, many pairs of fragments have similar measurements by chance.

Therefore, it is not surprising that each one of these measurements is hardly discriminative by itself; however, combined together, they are able to discriminate pretty reliably between joins and random pairs, although not nearly as well as handwriting similarity. The exact details are reported in (Wolf et al., 2011a).

Another source of available information is subject classification. A significant part of the digitized Genizah documents have already been manually classified by subject matter. The classification contains the categories Hebrew Bible, Bible translations, Bible commentaries, Talmud, liturgy, Judeo-Arabic literature, plus several more. Since every manuscript is expected to belong to one classification, this information is relevant for excluding improbable joins. However, the utility of this information is rather limited due to inconsistent classifications and, sometimes, multiple conflicting classifications for even the same fragment. Quite often, fragments of the same unknown join might have unrelated classifications.

Running a battery of tests, as described in (Wolf et al., 2011a), we found that handwriting is significantly more informative than physical measurements, which are more informative than subject classification. Still, the combination of the three, by means of multivariate regression, produces results that are considerably more accurate than using handwriting similarity alone. Specifically, for the benchmark of Section 3.2.4, recall at the false positive rate of 0.001 increases to 85%.

## 4.4 Discussion

A related task to that of join finding is that of scribe identification, where the goal is to identify the writer by morphological characteristics of his handwriting. This 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 (Bensefia et al., 2003, Panagopoulos et al., 2009). The work of Bres, Eglin and Auger (2006) uses text-independent statistical features, while other efforts combine both local and global statistics (Bulacu and Schomaker, 2007, Dinstein and Shapira, 1982). Within this spectrum, our approach for handwriting similarity is local, with the property that the graphemes are automatically extracted without human supervision.

Previous contributions to handwriting recognition identify the writer of the document from a list of known authors. Here, we concentrated on finding join candidates, and did not assume a labeled training set. 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 focused on this task. The handwriting techniques we use are not entirely suitable for distinguishing between different manuscripts penned by the same writer. However, the additional data employed, such as genre and topic classifications and physical parameters, help distinguish different manuscripts by the same writer.

Interestingly, there is a specialization to individual languages, employing language-specific letter structure and morphological characteristics (Bulacu and Schomaker, 2007, Dinstein and Shapira, 1982, Panagopoulos et al., 2009). Since the Genizah contains a multitude of script styles and languages, our solution has to be generic by design.

## 5 Conclusion

Digitizing collections of historical manuscripts is rapidly becoming one of the main tasks of librarians and preservers of such collections. While it is understood that digital images are required for conservation, sustainability, accessibility and manipulability needs, we maintain that they are also important as input to advanced artificial-intelligence image-analysis techniques that can contribute great benefits to the study of these manuscripts. The digitization effort should, therefore, take into account the fact that the computer itself will be one of the most important "consumers" of the digital images, and its needs must be taken into account when planning a digitization project. These requireents are: including a ruler in every image, choosing a contrasting background of a specific blue hue, avoiding the use of artifacts and capturing images at 600 dpi (in the specific meaning explained earlier).

Having done so, the computer can process each image to automatically extract a large amount of useful data about the fragment's physical attributes, including its dimensions, the number of rows, margins, character and line density, etc. Moreover, artificial-intelligence techniques can be applied to a pair of images to ascertain the likelihood that they were written by the same scribe, and by incorporating measurements extracted from the images, to determine whether they are indeed a join, originating from the same manuscript.

This approach is being applied to the huge and very important Cairo Genizah collection (comprising 350,000 fragments), whose manuscripts are currently dispersed all over the world. Using the methods described here might well enable us to reconstruct the original Genizah collection, thus assuring a quantum leap in facilitating Genizah research.

## References

Federal Agencies Digitization Initiative. Digital Conversion – Documents and Guidelines, A Bibliographic Reference, 2009. Available at http://www.digitizationguidelines.gov/guidelines/Guidelines_Bibliography-2009rev.pdf

Federal Agencies Digitization Initiative. Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files, 2009 (revised: August 2010). Available at http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf

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

I. Bar-Yosef, I. Beckman, K. Kedem, and I. Dinstein. Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents. *Int. J. Doc. Anal. Recognit.*, 9(2):89–99, 2007. ISSN 1433-2833. doi: http://dx.doi.org/10.1007/s10032-007-0041-5.

A. Bensefia, T. Paquet, and L. Heutte. Information retrieval based writer identification. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 946–950, 2003.

S. Bres, V. Eglin, and C. Volpilhac Auger. Evaluation of handwriting similarities using Hermite transform. In Guy Lorette, editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France) France, 2006. Suvisoft. Available at http://hal.inria.fr/inria-00108345/en/.

M. Bulacu and L. Schomaker. Automatic handwriting identification on medieval documents. In *14th International Conference on Image Analysis and Processing, ICIAP 2007*, pages 279 –284, 10-14 2007. doi: 10.1109/ICIAP. 2007.4362792.

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

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

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

M. Glickman. *Sacred Treasure — The Cairo Genizah*. Jewish Lights Publishing, Woodstock, Vermont, 2010.

A. Hoffman and P. Cole. *Sacred Trash: The Lost and Found World of the Cairo Geniza*. Nextbook-Schocken, New York, 2010.

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

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* volume 2, pages 2169–2178, 2006. doi: 10.1109/CVPR.2006.68.

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

M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, and S. Tracy. Automatic writer identification of ancient greek inscriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1404–1414, Aug. 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.201.

S. Reif and S. Reif. *The Cambridge Genizah Collections: Their Contents and Significance*. Cambridge University Press, 2002.

J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33: 225-236. 2000.

R. Shweka, Y. Choueka, L. Wolf, and N. Dershowitz. 'Veqarev otam ehad el ehad': Zihui ktav yad vezeruf qitei hagnizah be-emzaut mahshev (=Identifying handwriting and joining Genizah fragments by computer). *Ginzei Kedem*, 7:171–207, 2011. (In Hebrew.).

S. N. Srihari and V. Govindaraju. Analysis of textual images using the Hough transform. *Machine Vision and Applications*, 2(3):141–153, 1989.

L. Wolf, N. Dershowitz, L. Potikha, T. German, R. Shweka, and Y. Choueka. Automatic paleographic exploration of Genizah manuscripts. In *Codicology and Palaeography in the Digital Age II*. Norderstedt: Books on Demand, Germany, 2011a.

L. Wolf, R. Littman, N. Mayer, N. Dershowitz, R. Shweka, and Y. Choueka. Automatically identifying join candidates in the Cairo Genizah. In *Post ICCV Workshop on eHeritage and Digital Art Preservation*, Sept. 2009.

L. Wolf, R. Littman, N. Mayer, T. German, N. Dershowitz, R. Shweka, and Y. Choueka. Identifying join candidates in the Cairo Genizah. *International Journal of Computer Vision*, 94(1):118–135, 2011b.