

Binarization of First Temple Period Inscriptions – Performance of Existing Algorithms and a New Registration Based Scheme

Arie Shaus, Eli Turkel

The Department of Applied Mathematics
Tel Aviv University
Tel Aviv 69978, Israel

E-mails: ashaus@post.tau.ac.il, turkel@post.tau.ac.il

Eli Piassetzky

The Sackler School of Physics and Astronomy
Tel Aviv University
Tel Aviv 69978, Israel

E-mail: eip@tauphy.tau.ac.il

Abstract—The discipline of First Temple Period epigraphy (the study of writing) relies heavily on manually-drawn facsimiles (black and white images) of ancient inscriptions. This practice may unintentionally mix up documentation and interpretation. As an alternative, this article surveys the performance of several existing binarization techniques. The quality of their results is found to be inadequate for our purpose. A new method for automatically creating a facsimile is then suggested. The technique is based on a connected-component oriented elastic registration of an already existing imperfect facsimile to the inscription image. Some empirical results, supporting the methodology, are presented. The procedure is also relevant to the creation of facsimiles for other types of inscriptions.

Facsimile; binarization; elastic registration; CMI; epigraphy; First Temple Period; Iron Age; ostrakon.

I. INTRODUCTION

Most of the Hebrew texts of the Iron Age (First Temple period) in Israel and Judah were written in ink on papyrus. However, these documents did not survive the journey down the millennia. The most abundant among the meaningful surviving texts were written in ink on ostraca (pieces of pottery). Important ostraca corpora were unearthed in Samaria [1], Lachish [2] and Arad [3].

The discipline of Iron Age epigraphy relies heavily on manually-drawn facsimiles (binary documents) of these ostraca inscriptions. However, facsimiles crafted by hand may unintentionally mix up documentation with interpretation. Surprisingly, despite their importance for the field of epigraphy, to the best of our knowledge no attention has thus far been devoted to automatic facsimile creation.

We first survey the performance of several known computerized binarization techniques, either general-purpose (Otsu [4], Bernsen [5] and Niblack [6]), or specifically adapted for document analysis (White [7], Sauvola [8] and Gatos [9]). The resulting binarizations are found to be of insufficient quality. We then propose a new method for automatically creating a facsimile. It is based on a connected-component oriented elastic registration of an already existing imperfect facsimile to the inscription image. The registration will utilize a simple target function (explained briefly in [10] and elaborated upon in [11]), on

both large and small scales. The performance of the new binarization will also be tested.

II. PREVIOUS WORK

A. Examined Algorithms

For the purpose of comparing the quality of the results stemming from available binarization methods to a facsimile manually drawn by an epigrapher, six binarization algorithms are considered. These include three general-purpose binarization algorithms with wide acceptance: Otsu [4], Bernsen [5] and Niblack [6], as well as the White [7], Sauvola [8] and Gatos [9] methods, which focus on the domain of document analysis, in particular in a low quality (e.g. historical) setting. In addition to being the most popular, some of these techniques also serve as a basis for other binarization algorithms. This is apparent from the survey, performance comparison and methodological articles [12-16].

Otsu [4] maximizes the between-class variance criteria:

$$\omega_0 \omega_1 (\mu_1 - \mu_0)^2 \quad (1)$$

where μ_0 and μ_1 are averages of the two pixel “populations” (determined by a threshold), and ω_0 , $\omega_1 = 1 - \omega_0$ are their appropriate proportions.

Bernsen's method [5] is based on a “contrast measure” $C(x, y) = z_{high} - z_{low}$, i.e. the difference between the brightest and the darkest pixels. If $C(x, y) < l$ (l is a parameter, [17] recommends a value of $l = 15$), the local population is assumed to be homogeneous, and is marked as background. Otherwise, the threshold is:

$$T(x, y) = (z_{high} + z_{low}) / 2 \quad (2)$$

Bernsen's criterion suffers from a non-robust behavior, especially in the presence of salt-and-pepper type of noise.

The Niblack [6] binarization uses the threshold:

$$T(x, y) = m(x, y) + k \cdot s(x, y) \quad (3)$$

where $m(x, y)$ is a local mean, $s(x, y)$ is the local standard deviation and k is a parameter (with a recommended value of $k = -0.2$). Since $s(x, y) > 0$, and $k < 0$, $T(x, y) < m(x, y)$. Therefore, given a reasonable distribution of pixels, their majority is expected to be assigned to the (white) background.

The White algorithm uses a running average scheme, constantly updated by the current pixel values in a non-linear fashion. Look-ahead considerations in both image directions are also present. For additional details, see [7], Dynamic Threshold Algorithm section.

The Sauvola method [8] is composed of two stages. The first, a region analysis (extricating textual and non-textual regions) does not perform well for our purpose. We therefore concentrate on the second stage, adaptive thresholding. The local threshold is defined as:

$$T(x, y) = m(x, y) \cdot \left[1 + k \cdot \left(\frac{s(x, y)}{R} - 1 \right) \right], \quad (4)$$

where $m(x, y)$ is the local mean, $s(x, y)$ is the local standard deviation, k and R are parameters (with recommended values of $k = 0.5$ and $R = 128$). Since $s(x, y) < R$, and $k > 0$, $T(x, y) < m(x, y)$. Therefore, a majority of the pixels are again expected to be assigned to the (white) background.

The Gatos binarization technique is intended to handle low quality historical documents. In its original form [9], it consists of a pre-processing utilizing a Wiener filter, an estimation of foreground regions using Niblack's approach (see above), a background surface interpolation, a thresholding by comparing the estimated background surface to the original image, and a post-processing procedure. In the following, the last stage was ignored in order to compare the different binarization algorithms on an equal basis.

B. Binarization Results for Existing Algorithms

The experiments presented below (also see Section IV, remark b) were performed on three images of different ostraca, Lachish ostracon No. 3 (reversed side), Arad ostracon No. 1 (both Lachish No. 3 and Arad No. 1 contain ancient Hebrew writing), and Arad ostracon No. 34 (containing Hieratic, i.e. Egyptian, numerals). In all cases, the recommended parameters were used and the width of moving window was chosen as $W = 101$ (that way, the window encompasses even the largest characters). No pre- or post-processing was performed. The experimental results for the ostraca of Lachish No. 3, Arad No. 1 and Arad No. 34 can be seen respectively in Figs. 1, 2 and 3.

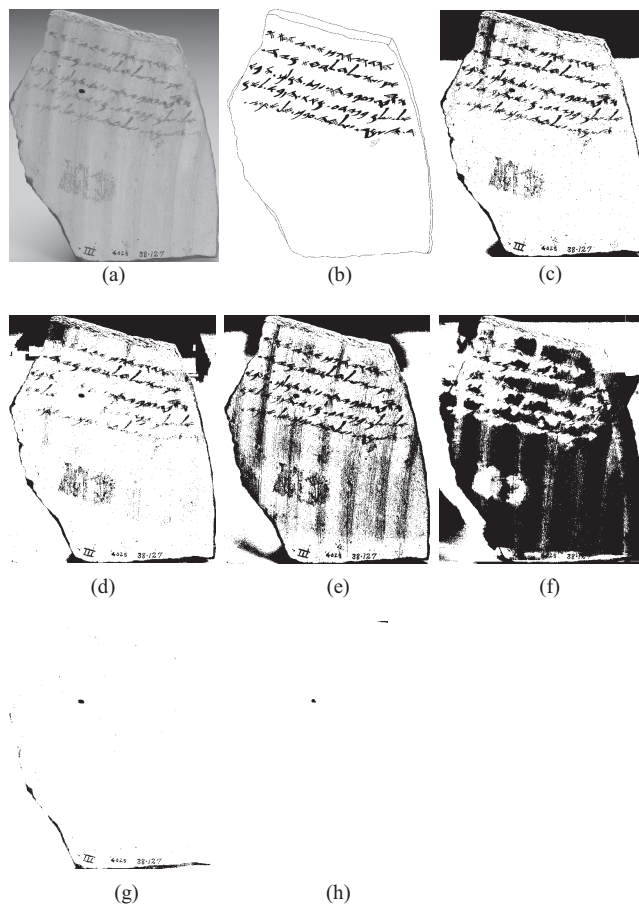


Figure 1. Lachish No. 3 experiment. (a) ostracon image (b) manual facsimile (c) Otsu (d) Bensen (e) Niblack (f) White (g) Sauvola (h) Gatos.

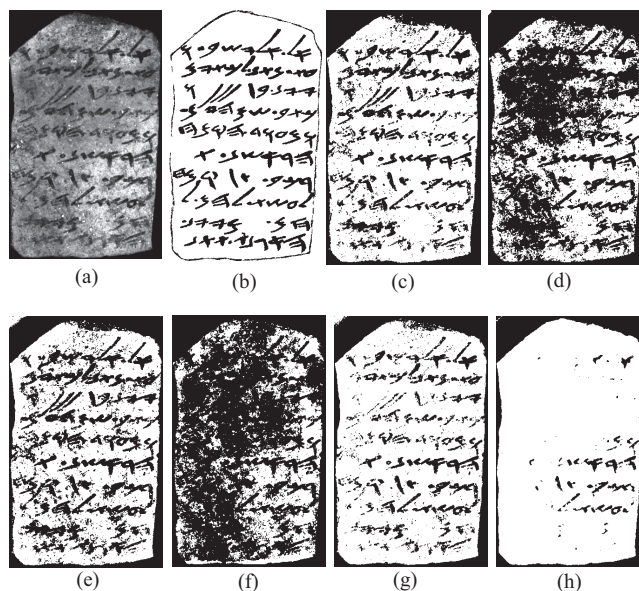


Figure 2. Arad No. 1 experiment. (a) ostracon image (b) manual facsimile (c) Otsu (d) Bensen (e) Niblack (f) White (g) Sauvola (h) Gatos.

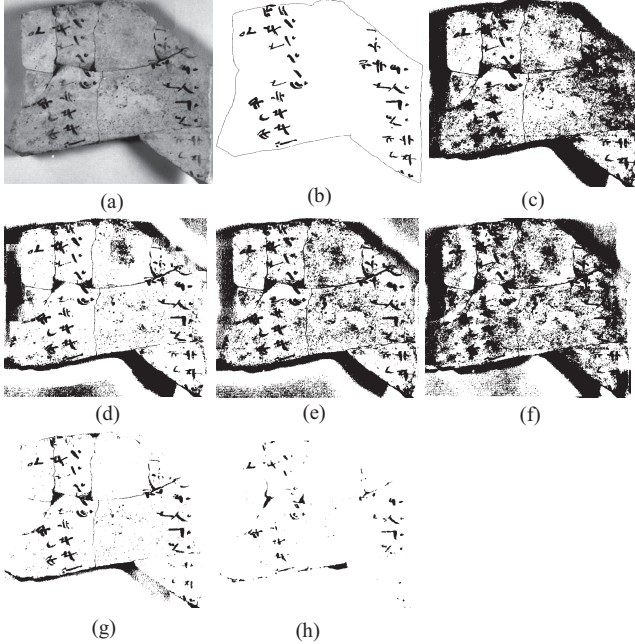


Figure 3. Arad No. 34 experiment. (a) ostracon image (b) manual facsimile (c) Otsu (d) Bernsen (e) Niblack (f) White (g) Sauvola (h) Gatos.

The experiments show that no algorithm was able to achieve binarization results that compare favorably to a manually drawn facsimile. The reason for that is the degraded, exceedingly non-uniform medium (i.e. input image), the presence of non-Gaussian and cross-pixel-dependent noise, broken strokes, cracks and stains mistaken for characters etc. Subsequently, in the next section, an alternative binarization scheme, taking into account information from the facsimile itself, will be presented.

III. PROPOSED REGISTRATION BASED BINARIZATION

A. Algorithm Description

We now present a new binarization algorithm. It is based on registering a pre-existing (not completely accurate) binary facsimile to the ostracon image. The ostracon image is always held constant, while the binary image undergoes various transformations. The registration procedure reduces the distortions imposed on the characters within the registered facsimile (for a survey of less restrictive registration algorithms see [18]). Finally, the registered facsimile information is utilized in order to produce an ostracon image binarization.

The algorithm steps, presented below, will be demonstrated on the Arad No. 1 ostracon and the facsimile images ([3]).

1) Preliminary Registration

This stage attempts at establishing an initial high-level registration. The only permitted degree of freedom for the registration is the rotation angle of the facsimile with respect to the ostracon image. Following the rotation, the facsimile image is automatically adjusted in order to fit the ostracon image dimensions. The target function for this, and all the subsequent stages, is:

$$CMI(F, O) = \mu_c - \mu_i, \quad (5)$$

where $O(p)$ is the ostracon image, $F(p)$ is the facsimile image ($p \in [1, M] \times [1, N]$). μ_c and μ_i are respectively the averages of ostracon image pixels corresponding to the clay (255) and ink (0) pixels of the registered facsimile image, denoted as “clayness” and “inkness”. The combined CMI (“clayness minus inkness”) measure strives to maximize the clayness (averaging bright ostracon pixels), while simultaneously minimizing the inkness (averaging dark ostracon pixels). A concise explanation of the CMI index is given in [10], with a more in-depth analysis supplied in [11]. Fig. 4 illustrates the results of the registration on superimposed facsimile and ostracon images. It can be seen that the target function performs well for registration purposes. On the other hand, the remaining “shadows” near certain characters indicate that on a low level, a better registration is needed, leading to the next registration stages.

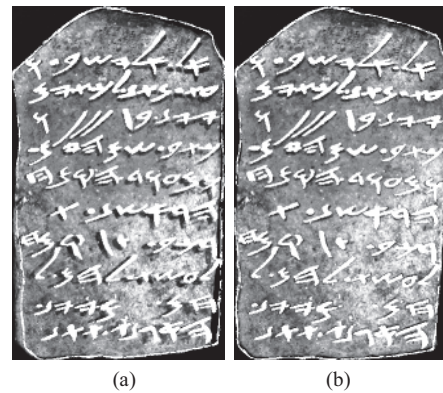


Figure 4. Example of ostracon-facsimile correspondence before (a) and after (b) the registration.

2) Unconstrained Elastic Registration

This stage attempts to achieve a more accurate low-level registration. The preliminarily registered *facsimile* is decomposed into *connected components* (CC). Each CC is given an $O(p)$ window, within which it is allowed to “float” freely. In other words, the CMI index within the window is optimized with respect to the CC's position. A brute-force implementation of such a local registration within a $W \times W$ window would require $O(W^2)$ computations. However, due to the typically observed convexity of the local CMI function, a simple “hill-climbing” technique works almost just as well (the exceptional cases handled, among other phenomena, on the next step), considerably reducing the complexity to $O(W)$. An example of an overall unconstrained elastic registration is shown in Fig. 5. The improvement is apparent, though due to the unrestricted nature of the registration, some CC's settled on a local CMI maxima, “merging” with the others.

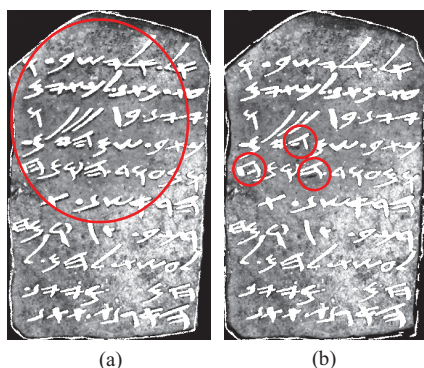


Figure 5. An example of ostracum-facsimile correspondence before (a) and after (b) the unconstrained elastic CC registration. The old and the new misalignments are marked by red color. Notice that within the new registration, some CC's were “swallowed” by the others.

3) Constrained Elastic Registration

The goal of this stage is to regularize and synchronize the movement of the facsimile image CC's. For every CC, the x and y movements of the previous stage are documented. Each displacement, in each coordinate, is then replaced by the median of the movements of the surrounding CC's (akin to the median filter). Hence, the displacements of CC's not correlated with the movement of the surrounding CC's (going “against the flow”) are easily detected and handled. Afterwards, beginning at the new (“median”) starting position, each CC is again allowed to find a CMI-optimized location. Fig. 6 illustrates the CC's movements before and after the application of median filter and re-registration. Fig. 7 shows the improvement in the ostracum-facsimile correspondence.

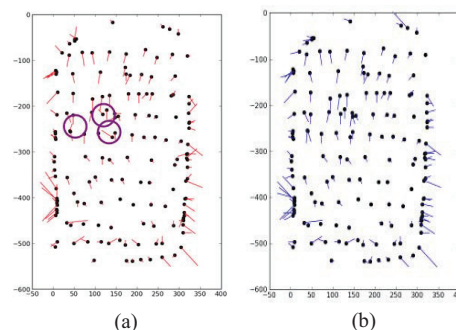


Figure 6. An example of per-CC movement (in pixels) before (a) and after (b) median filter and re-registration. Note the disappearance of the old misalignments, marked by violet color.

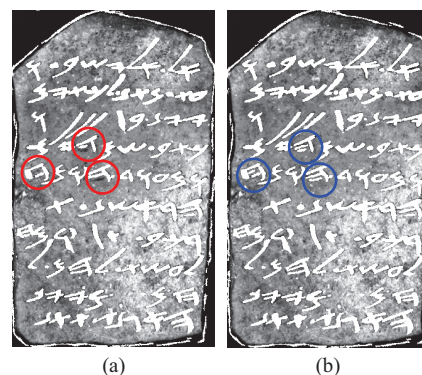


Figure 7. An example of improvement between the second (a) and third (b) registration stages. Note the reappearance of the missing CC's.

4) Proportional Binarization

The last stage utilizes the current registration in order to achieve a satisfying binarization of the ostracum image. For each CC, a bounding structure is defined. A convex hull (which is more accurate than a bounding rectangle) is a reasonable option. However, in our case, a *bounding octagon* (BO) was preferred for simplicity reasons. The BO's can be thought of as image areas within either ostracum or the registered facsimile image. The BO's are somewhat dilated in order to account for certain inaccuracies in the manual facsimile.

Binarization is then performed within each BO of the ostracum image. The classical algorithms mentioned in Section II, performed at the BO level, result in a binarization of disappointing quality. This can be explained by the fact that within the BO, the ink pixels proportion tends to be unusually high. This may contradict the assumptions regarding the background prominence (see Section II.A). Therefore, some of the methods are either stuck in sub-optimal maxima, or have to be adapted by ad-hoc tuning of their parameters.

A different, simple option is therefore preferred. Though the manual facsimile contains inaccuracies stemming from the epigrapher's cognitive world, within each BO, the proportion of ink pixels to be expected is roughly the same as in the manual facsimile. Therefore, we first calculate the

ink proportion IP_j for each BO_j within the registered manual facsimile ($RF(p)$):

$$IP_j = \frac{\#\{p \mid p \in BO_j \wedge RF(p) = 0\}}{\#\{p \mid p \in BO_j\}} \quad (6)$$

Second, for each BO_j of the ostracon image, we find the appropriate threshold T_j such that:

$$\frac{\#\{p \mid p \in BO_j \wedge O(p) < T_j\}}{\#\{p \mid p \in BO_j\}} \cong IP_j \quad (7)$$

Finally, every BO_j within $O(p)$ is thresholded according to the T_j in (7). In addition, small denoising procedures (e.g. morphological operations) can be performed, either within each BO, or on a global scale. In what follows, we present results without denoising, as well as results with simple stain (CC below certain size) removal.

B. Binarization Results

Following the experimental setting of section II.B, the ostraca and facsimile images of Lachish No. 3, Arad No. 1 and Arad No. 34 were analyzed. The results of the new registration and binarization algorithm can be seen respectively in Figs. 8, 9 and 10. In all cases, the ostraca border pixels were removed.

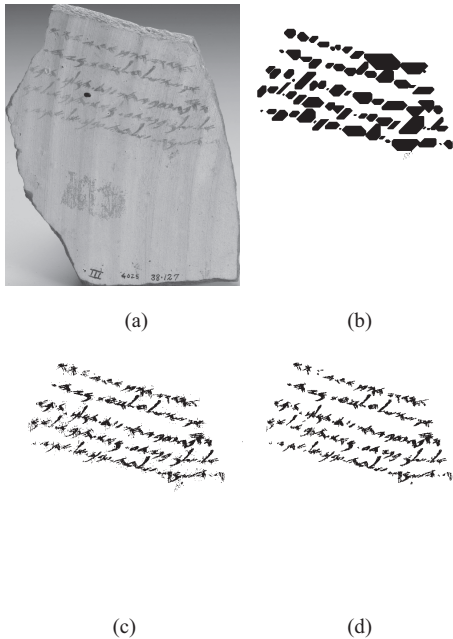


Figure 8. Lachish No. 3. (a) ostracon image (b) bounding octagons (c) binarization result (d) binarization result with stain removal.

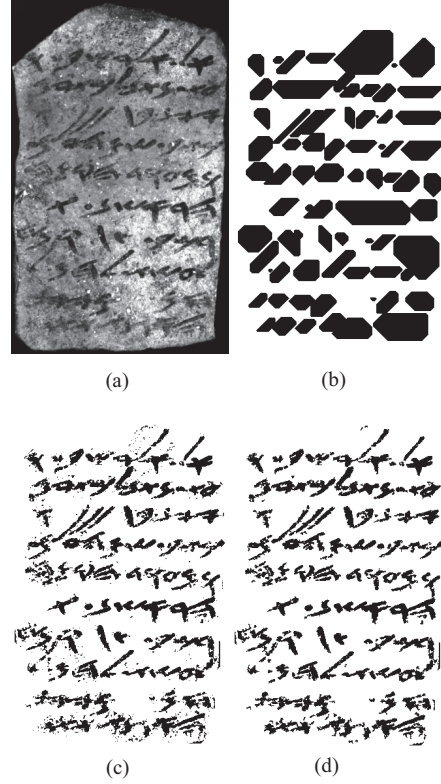


Figure 9. Arad No. 1. (a) ostracon image (b) bounding octagons (c) binarization result (d) binarization result with stain removal.

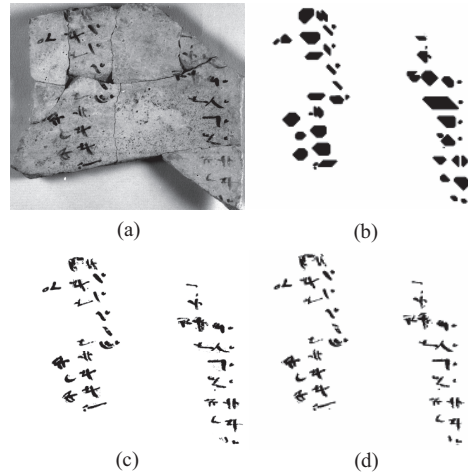


Figure 10. Arad No. 34. (a) ostracon image (b) bounding octagons (c) binarization result (d) binarization result with stain removal.

The quality of the output, albeit not ideal, clearly indicates that the new binarization compares favorably to other surveyed algorithms, and in some cases, to the manual facsimiles. This is not surprising, as harvesting information from the *facsimile*, however imperfect it may be, appears to be beneficial for identifying the interesting *ostraca* image areas and their properties. On the other hand, cracks and

stains, which might be mistaken for characters, are avoided unless they fall in close proximity to real letters.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

Six prominent binarization techniques, several of them specializing on low quality historical documents, were tested on a set of three quite different ostraca. Their overall results are far from satisfying. Our new binarization algorithm, based on registering a pre-existing inexact facsimile to an ostraca image, was also tested, with superior results. It can therefore be concluded that the proposed method is sound and can be used for automatic creation of facsimiles. It can also be used as a facsimile-drawing aid for epigraphers, providing a possibility for subsequent facsimile refinements.

Several further developments of the algorithm are worth mentioning:

- a. A comparison of the newly created facsimiles vis-à-vis manual facsimiles (exceeding the scope of this paper) was performed with tools such as mentioned in [10] and [11], showing the superiority of the new binarization. Additional tests are expected to strengthen the confidence in this methodology.
- b. The results presented here were obtained from a limited number of test cases. In addition to these, we successfully experimented with several other ostraca and tested the technique on different scales (1, 1/4 and 1/8).
- c. The new binarization method presupposes an existing manual facsimile for each ostracon. Though this is true in the domain of Hebrew First Temple Period Inscriptions, this may not be the case in other types of documents. An alternative may be utilization of automatically created facsimile, in a manner reminiscent of [9].
- d. The computerized facsimile is not invariant with respect to the manual one. Thus, it can benefit from cases when more than one manual facsimile exists for a given ostracon.
- e. On a bounding octagon level, more sophisticated binarization techniques, taking into account different character constraints, can be considered.
- f. The presented method may be easily adapted for other types of inscriptions with pre-existing facsimiles.

ACKNOWLEDGMENT

This study was supported by the F.I.R.S.T. (Bikura) Individual Grant no. 644/08 and by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement no. 229418.

Arie Shaus is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. We would also like to acknowledge the valuable help of Ms. Yael Barschak of the Israel Antiquities Authority, Prof. Bruce E. Zuckerman of the University of Southern California, and Ms. Judith Dekel, Ms. Shira Faigenbaum, Prof. Israel Finkelstein, Ms. Myrna Pollak, Prof. Benjamin Sass, Mr. Pavel Shargo and Mr. Barak Sober, all of Tel Aviv University.

Ostracon images: courtesy of the Institute of Archaeology, Tel Aviv University; and of the Israel Antiquities Authority. Facsimiles: courtesy of Ms. Judith Dekel; and of the Israel Exploration Society [3]. The algorithms of White and Gatos were implemented via the Gamera toolkit [19].

REFERENCES

- [1] G. A. Reisner, C. S. Fisher, D. G. Lyon, *Harvard Excavations at Samaria*. Cambridge, 1924.
- [2] H. Torczyner et al. *Lachish I: The Lachish Letters*. London, 1938.
- [3] Y. Aharoni, *Arad Inscriptions*. Israel Exploration Society, 1981.
- [4] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Systems Man Cybernet.* Vol. 9 (1), 1979, pp. 62–66.
- [5] J. Bernsen, "Dynamic thresholding of grey-level images," *Proceedings of the Eighth International Conference on Pattern Recognition*, Paris, France, 1986, pp. 1251–1255.
- [6] W. Niblack, *An Introduction to Digital Image Processing*, Prentice-Hall, 1986, pp. 115–116.
- [7] J. M. White, G. D. Rohrer, "Image thresholding for optical character recognition and other applications requiring character image extraction," *IBM J. Res. Dev.* 27(4), 1983, pp. 400–411.
- [8] J. Sauvola, M. Pietikainen, "Adaptive document image binarization," *Pattern Recognition*, Vol. 33, 2000, pp. 225–236.
- [9] B. Gatos, I. Pratikakis, S. J. Perantonis, "An adaptive binarization technique for low quality historical documents," *Lecture Notes in Computer Science* 3163, 2004, pp. 102–113.
- [10] A. Shaus, I. Finkelstein, E. Piasezky, "Bypassing the eye of the beholder: automated ostraca facsimile evaluation," *Maarav*, in press.
- [11] A. Shaus, E. Turkel, E. Piasezky, "Quality evaluation of facsimiles of Hebrew First Temple Period inscriptions," *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems (DAS 2012)*, Gold Coast, Australia, 2012.
- [12] M. Sezgin, B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging* 13(1), 2004, pp. 146–165.
- [13] P. Stathis, E. Kavallieratou, N. Papamarkos, "An evaluation survey of binarization algorithms on historical documents," *Journal of Universal Computer Science*, Vol. 14, No. 18, 2009, pp. 3011–3030.
- [14] J. He, Q. D. M. Do, A. C. Downton, J. H. Kim, "A comparison of binarization methods for historical archive documents," *Proceedings of the Eight International Conference on Document Analysis and Recognition (ICDAR 2005)*, Vol. 1, 2005, pp. 538–542.
- [15] B. Gatos, K. Ntirogiannis, I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*, 2009, pp. 1375–1382.
- [16] E. Kavallieratou, "An objective way to evaluate and compare binarization algorithms," *Proceeding of 23rd Annual ACM Symposium on Applied Computing*, 2008, pp. 397–401.
- [17] D. Trier, T. Taxt, "Evaluation of binarization methods for document images," *IEEE Trans. Pattern Anal. Mach. Intell.* 17, 1995, pp. 312–315.
- [18] B. Zitová, J. Flusser, "Image registration methods: a survey," *Image and Vision Computing* 21, 2003, pp. 977–1000.
- [19] <http://sourceforge.net/projects/gamera>