



The Raymond and Beverly Sackler Faculty of Exact Sciences

School of Mathematical Sciences

Department of Applied Mathematics

Computer Vision and Machine Learning Methods for Analyzing First Temple Period Inscriptions

Thesis submitted for the degree of Doctor of Philosophy

by

Arie Shaus

This work was carried out under the supervision of **Professor Eli Turkel**

Secondary supervisor: **Professor Israel Finkelstein**

Submitted to the Senate of Tel Aviv University

December 2017

*Dedicated to the blessed memory of
my late nephew Ido Gofer and my
late uncle Ilya Shaus.*

Acknowledgements

I would like to express the sincerest gratitude to my supervisor, Prof. Eli Turkel, for his thoroughly knowledgeable advices, suggestions, corrections and, no less importantly, his kind demeanor and infinite patience, contributing to successful culmination of this thesis. I would also like to thank my secondary supervisor, Prof. Israel Finkelstein, for his ongoing support, thought-provoking riddles, and an inspiration to aim at unearthing the answers to the “big questions”. Although not a formal supervisor, Prof. Eli Piasezky was instrumental in inaugurating and backing my research, while pushing many of the investigations beyond their finishing line, and I’m grateful to him.

I would like to thank the Lautman family and the Adi Lautman Interdisciplinary Program for Outstanding Students for the wealth and scope of knowledge acquired during my first academic steps, including the generous funding and scholarships. I’m also thoroughly grateful to the Azrieli family and to the Azrieli Foundation for the award of an Azrieli Fellowship, to the Foundation’s ever-positive program manager, Ms. Rochelle Avitan, the program coordinator, Ms. Yula Panai, as well as the Foundation’s coordinator at TAU, Ms. Shoshi Noy, for their sympathetic help. This study was also supported by a generous donation of Mr. Jacques Chahine, made through the French Friends of Tel Aviv University. The considerable help of Amalia Biron-Cegla Scholarship, Early Israel Grant (New Horizons project), and Electronic Imaging Student Grant is similarly acknowledged. In addition, the research received funding from the Israel Science Foundation – F.I.R.S.T. (Bikura) Individual Grant no. 644/08, the Israel Science Foundation Grant no. 1457/13, and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 229418.

I would also like to express gratitude to my dear *longue durée* friends, confidants, colleagues, coauthors, reviewers, critics and tee-drinking associates, Shira Faigenbaum-Golovin and Barak Sober. The cooperation with Anat Mendel-Geberovich, in particular on character recognition, documentation and smoothing, was ever so engaging. I'm also grateful to Yariv Aizenbud and Eythan Levy for stimulating discussions and encouragement.

I would also like to acknowledge the valuable assistance, advice and cooperation from Dr. Leah Bar, Ms. Yael Barschak, Dr. Shirly Ben-Dor Evian, Mr. Michael Cordonsky, Ms. Judith Dekel, Ms. Sivan Einhorn, Ms. Noa Evron, Ms. Liora Freud, Mr. Assaf Kleiman, Ms. Tamar Lavee, Prof. David Levin, Prof. Murray Moinester, Ms. Ma'ayan Mor, Prof. Nadav Na'aman, Ms. Myrna Pollak, Prof. Christopher A. Rollston, Prof. Benjamin Sass, Mr. Pavel Shargo, Prof. David M. Steinberg, and Prof. Bruce E. Zuckerman. Your help is greatly appreciated. This research is also in great debt to the late Prof. Itzhaq Beit-Arieh. The remarks of the anonymous reviewers were very beneficial for the improvement of the text.

Finally, I would like to thank my family: my parents, for cultivating my curiosity since early age; my dear sister Maria, her husband Itay, and their cute kids Ido, Noa and Omri for little moments of happiness; my fiancée Anna for her unconditional love and support - you undoubtedly deserve another thesis for your perseverance in developing countless ways to encourage my work on the thesis (شكرا محبوبة!); and our beloved dog Hatti, who helped me finalize my ideas during our mutual walks.

Ostraca images: courtesy of the Institute of Archaeology, Tel Aviv University (photographer: Michael Cordonsky) and the Israel Antiques Authority. Facsimiles:

courtesy of Ms. Judith Dekel; of the Israel Exploration Society (Aharoni 1981); and of Prof. C. A. Rollston (Rollston 2006). Full spectrum color images of the Dead Sea Scrolls courtesy of the Israel Antiques Authority (photographer: Shai Halevi; Dead Sea Scrolls 2016). Paleographic tables are courtesy of the Israel Exploration Society (Aharoni 1981).

Abstract

The thesis concentrates on computational methods pertaining to ancient ostraca - ink on clay inscriptions, written in Hebrew. These texts originate from the biblical kingdoms of Israel and Judah, and dated to the late First Temple period (8th – early 6th centuries BCE). The ostraca are almost the sole remaining epigraphic evidence from the First Temple period and are therefore important for archaeological, historical, linguistic, and religious studies of this era. This “noisy” material offers a fertile ground for the development of various “robust” image analysis, image processing, computer vision and machine learning methods, dealing with the challenging domain of ancient documents’ analysis. The common procedures of modern epigraphers involve manual and labor-intensive steps, facing the risk of unintentionally mixing documentation with interpretation. Therefore, the main goal of this study is establishing a computerized paleographic framework for handling First Temple period epigraphic material. The major research questions, addressed in this thesis are: quality evaluation of manual facsimiles; quality evaluation of ostraca images; automatic binarization of the documents and its subsequent refinement; quality evaluation of binarizations on global and local levels; identification of different writers between inscriptions (two distinct methods are proposed); image segmentation (with improvements over the classical Chan-Vese algorithm); and letters’ shape prior estimation. The developed methods were tested on real-world archaeological and modern data and their results are found to be favorable.

Contents

Acknowledgements.....	5
Abstract.....	9
Contents	11
List of Figures	17
List of Tables	27
1. Introduction	29
2. Quality Evaluation of Manually Created Facsimiles.....	35
2.1 Background and Prior Art.....	35
2.2 Facsimile Evaluation.....	36
2.3 Experimental Results	39
Methodology Verification I.....	39
Methodology Verification II.....	41
2.4 Possible Drawbacks	43
2.5 Summary	44
3. Potential Contrast Quality Measure with Application to Multispectral Imaging .	45
3.1 The Problem.....	45
3.2 Prior Art	46
3.3 Requirements and Measure Definition	50
Requirements.....	50
Assumptions	51
Proposition I (Optimality)	51

3.4 Measure Properties.....	54
Population Separability	54
Complexity	56
Equivalence to Error Estimation.....	56
Symmetry between Foreground and Background	56
Proposition II (Invariance with Respect to Invertible g)	56
3.5 Automated Foreground/Background Selection.....	58
3.6 Experimental Results	59
Experiment Results for Manual Foreground and Background Selection	60
Experiment Results for Automated Foreground and Background Estimation	61
3.7 Application of the Methodology	62
3.8 Summary	65
4. Binarization via Registration-based Scheme.....	67
4.1 Introduction.....	67
4.2 Prior Art	67
Examined Algorithms.....	67
4.3 Proposed Algorithm's Description	74
1) Preliminary Registration	74
2) Unconstrained Elastic Registration	75
3) Constrained Elastic Registration	76
4) Proportional Binarization	77
4.4 Proposed Algorithm's Results	79

4.5 Summary	82
5. Binarization Improvement via Sparse Dictionary Model.....	83
5.1 Problem Statement	83
5.2 Proposed Solution	84
5.3 Experimental Results	87
5.4 Summary	91
6. Quality Evaluation of Binarizations	93
6.1 Introduction.....	93
6.2 Methodological Pitfalls.....	94
6.3 Existing Solutions	95
6.4 Preliminary Definitions and Assumptions	98
6.5 Proposed Measures	99
Proposition I (Equivalence of PSNR and $-L_2$):.....	101
Proposition II (Equivalence of L_1 and L_2):.....	102
6.6 Experimental Setting and Results	103
Experimental Setting	103
Experimental Results.....	105
6.7 Summary	108
7. Quality Evaluation of Individual Characters' Binarizations	111
7.1 Introduction.....	111
7.2 Suggested Character Measures	111
Stroke Width Consistency Measure	112

Edge Noise Proportion Measure.....	116
Measures' Combinations	118
7.3 Experimental Design.....	118
Motivation	118
Dataset	119
Goal	120
Input Data	120
Models' Specifications	121
Models' Score, Selection and Success Criteria	122
Selected Model	122
Selected Model Verification.....	123
7.4 Summary	124
8. Writers' Identification via a Combination of Features, with Historical Implications	125
8.1 Introduction.....	125
8.2 Prior Art	127
8.3 Materials and Methods.....	129
8.4 Algorithmic Apparatus.....	131
Feature Extraction and Distance Calculation	133
Hypothesis Testing	137
8.5 Results.....	141
Modern Hebrew experiment.....	141
Arad Ancient Hebrew experiment.....	144

8.6 Discussion	148
9. Writers' Identification via Binary Pixel Patterns and Kolmogorov-Smirnov Test	153
9.1 Introduction.....	153
9.2 Algorithm's Description	154
Preliminary Remarks	154
Same Writer Statistics Derivation	158
9.3 Modern Hebrew Experiment.....	160
The Basic Settings	160
Parameter Tuning and Robustness Verification	161
Experimental Results.....	162
9.4 Ancient Hebrew Experiment.....	163
The Basic Settings	163
Experimental Results.....	164
9.5 Summary	165
10. Segmentation via Morphologically-based Chan-Vese Framework	167
10.1 Introduction and Prior Art.....	167
10.2 The Chan-Vese algorithm	169
10.3 From Chan-Vese to Alternative Solution	171
10.4 Experimental Results	175
10.5 Summary	176
11. Letter Shape Prior Estimation	179

11.1 Introduction.....	179
11.2 Prior Art	181
11.3 The Proposed Algorithm.....	183
11.4 Results.....	186
11.5 Summary	194
12. Conclusions and Future Research Directions	195
13. References.....	199

List of Figures

Figure 1.1 Examples of ostraca (ink inscriptions on clay) from the Iron Age fortress of Arad, located in arid southern Judah. These documents are dated to the latest phase of the First Temple Period in Judah, ca. 600 BCE. The texts represent correspondence of local military personnel.....	30
Figure 1.2 Ostracon No. 1 from Tel Arad: (a) an ostracon image; (b) hand drawn facsimile; (c) zoom-in on image and facsimile, the utmost left word of the last line. The leftmost “ <i>nun</i> ” character is documented, yet it is absent upon close inspection; (d) a fragment of a paleographic table, containing “representative” letters from different ostraca.	32
Figure 1.3 A summarizing schematic flowchart of the overall framework. Continuous lines represent direct input, while dotted lines represent auxiliary information.....	34
Figure 2.1 Example of (a) an ostracon image, (b) a facsimile image	38
Figure 2.2 Example of (a) initial facsimile-ostracon fit, (b) CMI-based registration..	38
Figure 2.3 Arad ostracon No. 34.....	39
Figure 2.4 Overlaid facsimile A, CMI = 71.1	40
Figure 2.5 Overlaid facsimile B, CMI = 82.6	40
Figure 2.6 Overlaid facsimile C, CMI = 84.0	41
Figure 2.7 Another image of Arad ostracon No. 34.	42
Figure 3.1 Example of images undergoing grayscale transformations. (a) original image with sampled foreground (in red) and background (in blue). (b) the image after brightness change (+70). (c) the image after histogram rescaling ($\times 1.3$). (d) the image after histogram equalization.....	49
Figure 3.2 An example of misleading naked eye: Two images stemming from the same source, with the same sampled populations (Fig. 3.1a). (a) added Gaussian noise of $\mu=0$,	

$\sigma=32$, $PC=206.28$ (b) narrowing the dynamic range and brightening ($I/4+200$), $PC=255.00$53

Figure 3.3 Example of ambiguous foreground and background. While it is possible that the kettle is the foreground and the chair is the background, writing as a foreground and whiteboard as a background is another viable option.53

Figure 3.4 Example of foreground and background not separable by thresholding, while easily classifiable by the PC framework. (a) original grayscale image (circle=0, writing within the circle=195, writings outside the circle=127, other areas outside the circle=255); (b) example of an image thresholded by 150; (c) circle and its content as foreground (in red) with the rest as background (in blue); (d) PC-binarization based on (c); (e) writing as foreground (in red) with the rest as background (in blue); (f) PC-binarization based on (e).54

Figure 3.5 A natural scene handled by our method. A good contrast enhancement is achieved despite the similarity in foreground and background shade. (a) RGB image of the scene with manual selection of foreground in red and background in blue; R (b), G (c) and B (d) channels, with respective PC values of 244.8, 67.6 and 61.2; the PC-binarizations for R (e), G (f) and B (g).55

Figure 3.6 An example of automatically created saliency-based foreground (a) and background (b) maps.....59

Figure 3.7 Section of Dead Sea scroll No. 124, fragment 001 (Dead Sea Scrolls 2016). (a) Image of a scroll; (b) PC-binarization of (a). 63

Figure 3.8 Images of Horvat Radum ostracon No. 1 (Beit-Arieh 2007, Sober et al. 2014). (a) optimal image at $\lambda=620$ nm, selected by our method; (b) sub-optimal image at $\lambda=950$ nm.64

Figure 3.9 Images of Horvat Uza ostracon No. 3 (Beit-Arieh 2007, Sober et al. 2014).
(a) RGB image; (b) multispectral image taken at $\lambda=660$ nm, selected by our method.
.....64

Figure 3.10 Images of ostracon No. 13.056-01-S01 from Qubur el-Walaydah
(Faigenbaum et al. 2014). (a) RGB image; (b) multispectral image taken at $\lambda=690$ nm,
selected by our method.64

Figure 3.11 *Verso* of Arad Ostracon 16. (a) current color image; (b) 890 nm image taken
via our multi-spectral imaging system.65

Figure 4.1 Lachish No. 3 experiment: (a) ostracon image; (b) manual facsimile. Results
of: (c) Otsu; (d) Bernsen; (e) Niblack; (f) White; (g) Sauvola; (h) Gatos.71

Figure 4.2 Arad No. 1 experiment: (a) ostracon image; (b) manual facsimile. Results
of: (c) Otsu; (d) Bernsen; (e) Niblack; (f) White; (g) Sauvola; (h) Gatos.72

Figure 4.3 Arad No. 34 experiment: (a) ostracon image; (b) manual facsimile. Results
of: (c) Otsu; (d) Bernsen; (e) Niblack; (f) White; (g) Sauvola; (h) Gatos.73

Figure 4.4 Example of ostracon-facsimile correspondence before (a) and after (b) the
registration.75

Figure 4.5 An example of ostracon-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 in (b), some CC's were “swallowed” by the others. 76

Figure 4.6 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 in (a).....77

Figure 4.7 An example of improvement between the second (a) and third (b) registration
stages. Note the reappearance of the missing CC's.....77

Figure 4.8 Lachish No. 3: (a) ostrakon image; (b) bounding octagons; (c) binarization result; (d) binarization result with stain removal.	80
Figure 4.9 Arad No. 1: (a) ostrakon image; (b) bounding octagons; (c) binarization result; (d) binarization result with stain removal.	81
Figure 4.10 Arad No. 34: (a) ostrakon image; (b) bounding octagons; (c) binarization result; (d) binarization result with stain removal.	82
Figure 5.1 A collection of patches, illustrating stages 1 and 2 of the algorithm.	86
Figure 5.2 Fragment of Arad #1: (a) binarization from Section 4 – in blue good patches reflecting the writing practice, in red non-representative “noisy” patches; (b) binarization improvement, with representative patches maintained with minimal changes, while non-representative patches replaced.	86
Figure 5.3 Fragment of Arad No. 34: (a) binarization from Section 4, in red non-representative “noisy” patches; (b) binarization improvement, with non-representative patches replaced.	86
Figure 5.4 Arad No. 1: (a) ostrakon image; (b) binarization from Section 4; (c) k-medians result; (d) k-medoids result; (e) extensive dictionary result. Zoom on right-center: (f) binarization from Section 4; (g) k-medians result; (h) k-medoids result; (i) extensive dictionary result.	88
Figure 5.5 Arad No. 34: (a) ostrakon image; (b) binarization from Section 4; (c) k-medians result; (d) k-medoids result; (e) extensive dictionary result. Zoom on top-left: (f) binarization from Section 4; (g) k-medians result; (h) k-medoids result; (i) extensive dictionary result.	89
Figure 5.6 Arad No. 1: experiment testing the robustness of k-medians, initial DB size reduced by a factor of: (a) 3 (b) 21 (c) 75; experiment testing the robustness of k-medoids, initial DB size reduced by a factor of: (d) 3 (e) 21 (f) 75.	90

Figure 6.1 Standard binarization quality evaluation process. The document image is gray-scale, while the binarization and the ground truth are black and white images. The quality metric measures the adherence of the binarization to the ground truth.	94
Figure 6.2 Proposed binarization quality evaluation process. The quality of binarization or ground truth is assessed by measuring their adherence to the document image.	94
Figure 7.1 Example of local-scale stroke width discontinuity due to stains and letter erosion (discontinuities marked in red).....	112
Figure 7.2 A demonstration of shortest stroke width = segment selection for a particular foreground pixel (in green – the shortest segment, in red – other segments considered).	113
Figure 7.3 An illustration of Step 2 in average edge curvature measure computation.	115
Figure 7.4 An example of edge pixel p , possessing three neighbors p_1 , p_2 and p_3 . This requires an adjustment in M_{AEC} calculations.	116
Figure 7.5 (a) Clean character, (b) Corrupted character.	117
Figure 7.6 Expert’s ranking of one character, in decreasing quality order. (a) Original image, (b) Sauvola $w = 200$, (c) Shaus et al. inc. unspeckle stage, (d) Shaus et al., (e) Otsu, (f) Niblack $w = 200$, (g) Niblack $w = 50$, (h) Sauvola $w = 50$, (i) Bernsen $w = 50$, (j) Bernsen $w = 200$	120
Figure 7.7 The selected regression model, a “forced” tree with 9 leaves. The leaves indicate the mean predicted rank (prior to re-ranking; after applying the ranking function 1.591 will become 1, 3.57 will become 2, etc.). Note that all four proposed measures are utilized by the selected model.	123
Figure 8.1 Main towns in Judah and sites in the Beer Sheba Valley mentioned in the current section.	126

Figure 8.2 Ostraca from Arad (Aharoni 1981): No. 5 (A), No. 24 (B) and No. 40 (C). The poor state of preservation, including stains, erased characters and blurred text, is evident..... 132

Figure 8.3 Artificial illustration of H_0 rejection experiment (containing only *alep* letters): (A) two compared documents; (B) unifying their sets of characters; (C) automatic clustering; (D) the clustering results vs. the original documents. 139

Figure 8.4 An example of a Modern Hebrew alphabet table, produced by a single writer (with 10 samples of each letter)..... 142

Figure 8.5 Comparison between several specimens of the letter *lamed*, stemming from: Arad 1 (A, B); Arad 7 (C, D) and Arad 18 (E, F). Note that our algorithm cannot distinguish between the author of Arad 1 and the author of Arad 7, or the authors of Arad 1 and Arad 18. On the other hand, Arad 7 and Arad 18 were probably written by different authors ($P=0.015$ for the letter *lamed* and $P=0.004$ for the whole inscription, combining information from different letters)..... 145

Figure 8.6 Reconstruction of the hierarchical relations between authors and recipients in the examined Arad inscriptions; also indicated is the differentiation between combatant and logistics officials..... 149

Figure 9.1 A comparison of handwriting analysis schemes. Left: common frameworks, producing an abstract distance between the documents as a final output. Center: the method of Section 8, performing the analysis on per-letter basis, yielding (number of letters) experimental p-values to be combined via Fisher’s method. Right: the current technique, performing Kolmogorov-Smirnov tests for each feature and each letter, yielding (num. of features) x (num. of letters) experimental p-values to be combined via Fisher’s method..... 154

Figure 9.2 A toy example of the same writer statistics derivation for two hypothetical inscriptions. Inscription I consists of two instances of the letter *alep*, and four instances of the letter *bet*, while Inscription II consists of three instances of the letter *alep*, and two instances of the letter *bet*. The only patches with enough statistics are patches numbers 1 and 2. Four comparisons of appropriate samples (for each letter and each patch) are performed via Kolmogorov-Smirnov test, yielding four different p-values. These p-values are then combined via Fisher’s method. 159

Figure 9.3 Testing the combined probability of FP+FN errors as a function of character area (in pixels) as well as different p-value thresholds: 0.2 in blue, 0.1 in red and 0.05 in green. Taking into account the performance of the algorithm in Section 8 (FP+FN≈0.043), all the tested thresholds and all the areas between 1000 and 40,000 pixels would yield reasonable and comparable performance. Slightly better results are achieved in the range of 8,000-20,000 pixels, with 0.1 threshold. 162

Figure 10.1 Five options of neighborhood of an A_1 borderline pixel. Only (a,b) require a re-assignment of the central pixel, due to a positive curvature. 173

Figure 10.2 Segmentation of an object of smooth contour: original image (*left*), vs. result with default setting (*center*), vs. result with radius=11 (*right*). 176

Figure 10.3 Segmentation of a satellite image of Europe night-lights: original image (*left*), vs. Otsu binarization (*center*), vs. result with the default setting (*right*). Image courtesy NASA/Goddard Space Flight Center Scientific Visualization Studio, public domain. 177

Figure 10.4 Segmentation of a spiral art-work: original image (*upper left*), vs. Otsu binarization (*upper right*), vs. result with the default setting (*lower left*), vs. result with radius=2 (*lower right*). Image courtesy José-Manuel Benito Álvarez, public domain. 177

Figure 10.5 Segmentation of a fragment of Arad ostracon No. 1: original image (<i>upper left</i>), vs. Otsu binarization (<i>upper right</i>), vs. result with the default setting (<i>lower left</i>), vs. result with radius=2 (<i>lower right</i>).	177
Figure 10.6 Segmentation of Lachish ostracon No. 3: original image (<i>upper left</i>), vs. Otsu binarization (<i>upper right</i>), vs. result with the default setting (<i>lower left</i>), vs. result with radius=3 (<i>lower right</i>).	178
Figure 11.1 Manually created paleographic table, recording “typical” representatives for each letter in the alphabet.	180
Figure 11.2 Arad 1 - an ostracon image (left) and its corresponding facsimile (right). The various colors of the facsimile indicate different letter types.	187
Figure 11.3 Arad 2 - an ostracon image (left) and its corresponding facsimile (right). The various colors of the facsimile indicate different letter types.	187
Figure 11.4 Arad 24b - an ostracon image (left) and its corresponding facsimile (right). The various colors of the facsimile indicate different letter types.	188
Figure 11.5 An example of the algorithm’s flow for “yod” letter, Arad 24b. Top: a median-based initialization of a prior (utilizing information from 14 characters), and an estimation of two consequent priors, with no attempt at regularization (smoothing). Bottom: three consecutive priors are regularized by an algorithm introduced in Section 10, with median radius set to 5.	189
Figure 11.6 Steps for a regularized prior computation of “mem” from Arad 2 (based on 10 characters).	189
Figure 11.7 The letter “ayin” from Arad 1 (based on 3 characters). Top: computation of letter prior for full resolution imagery, regularization with radius=5. Bottom: computation of letter prior for partial resolution (halved in each axis), with no regularization, radius=5 and radius=10.	189

Figure 11.8 Results of experiment #1 for different ostraca. 192

Figure 11.9 Results of experiment #2 for different ostraca. 193

List of Tables

Table 2.1 Results for two verifications of facsimile quality methodology	42
Table 3.1 Contrast measures comparison based on Fig. 3.1.....	50
Table 3.2 PC measure based on Fig. 3.1.....	57
Table 3.3 Manual foreground and background selection: Ratios between the measures of transformed images with respect to “initial” image (predicted invariance marked in red).	61
Table 3.4 Automatic foreground and background estimation: Ratios between the measures of transformed images with respect to “initial” image (predicted invariance marked in red).	62
Table 6.1 Results for Salt and Pepper Deterioration.....	106
Table 6.2 Results for Dilation of the Foreground	107
Table 6.3 Results for Erosion of the Foreground.....	107
Table 7.1 Comparison of quality measures, activated on clean (Fig. 7.5a) and corrupted (Fig. 7.5b) images.	117
Table 7.2 Agreement with success criteria.	123
Table 8.1 Features and distances used in our algorithm.	135
Table 8.2 Results of the Modern Hebrew experiment.	143
Table 8.3 Letter statistics for each text under comparison.	145
Table 8.4 Comparison between different Arad ostraca.	147
Table 9.1 Example of character histograms.....	157
Table 9.2 Results of Modern Hebrew Experiment.	163
Table 9.3 Results of Ancient Hebrew Experiment, indicating separation of writers between texts (in red background).....	166
Table 10.1 Description of the algorithm, including various options.	175

Table 11.1 Experiments' settings.....	190
Table 11.2 Results of experiment #1 for Arad 1 ostrakon	191
Table 11.3 Results of experiment #1 for Arad 2 ostrakon	191
Table 11.4 Results of experiment #1 for Arad 24b ostrakon	191
Table 11.5 Results of experiment #2 for Arad 1 ostrakon	192
Table 11.6 Results of experiment #2 for Arad 2 ostrakon	193
Table 11.7 Results of experiment #2 for Arad 24b ostrakon	193

1. Introduction

What's between applied mathematics and biblical archaeology? This combination would have been considered peculiar a few decades ago. Yet, these days, the amalgamation of these disciplines is not only reasonable, but even sought-after. Indeed, from the archaeological side, an ever-deepening cooperation with “hard” scientific disciplines (including, yet not limited to physics, chemistry, material sciences, geophysics, geology, genetics, botany and zoology) provides answers to long-standing issues and raises new questions (Shaus et al. 2017b). For several examples of such fruitful multi-disciplinary studies see (Finkelstein et al. 2012; Finkelstein et al. 2015, describing a major research project under the auspices of the European Research Council, with the current study as one of its tracks). On the other hand, “noisy” material stemming from the excavations offers a fertile ground for the development of various “robust” analytical methods, pushing the boundaries of science.

Inter-related research domains such as image processing, computer vision, pattern recognition, data mining, machine learning, text processing and other computational tools are not exceptional, and also become increasingly applicable in archaeological and historical setting (e.g., Gilboa et al. 2004; Brown et al. 2008; Lipschits et al. 2008). One of the fields resulting from this collaboration, is the emerging challenging domain of ancient documents' analysis (e.g., Dinstein and Shapira 1982; Schomaker et al. 2007; Bar-Yosef et al. 2007; Ben Messaoud et al. 2011). Beside the already mentioned mathematical subjects, this fascinating topic also pertains to linguistics, philology, epigraphy, paleography, theology, history and of course archaeology.

This thesis concentrates on particular type of ancient inscriptions, originating from the biblical kingdoms of Israel and Judah. These texts, bearing the Greek name of

ostraca (singular: *ostrakon*), are inscribed on clay sherds, and in our case are written in ink (and not incised). The majority of these documents were created towards the end of the First Temple period (8th – early 6th centuries BCE), also known as Iron Age II. The largest groups of ostraca were discovered in the excavations of Samaria (Reisner et al. 1924), Lachish (Torczyner et al. 1938), Arad (Aharoni 1981) and Horvat 'Uza (Beit-Arieh 2007), with several dozen relatively “lengthy” (encompassing 3-12 lines of text) ostraca in each corpus. Some examples of ostraca from the desert fortress of Arad (Aharoni 1981), dated to ca. 600 BCE, can be seen in Fig. 1.1.



Figure 1.1 Examples of ostraca (ink inscriptions on clay) from the Iron Age fortress of Arad, located in arid southern Judah. These documents are dated to the latest phase of the First Temple Period in Judah, ca. 600 BCE. The texts represent correspondence of local military personnel.

The ostraca are written in a language close to Biblical Hebrew, in ancient Paleo-Hebrew alphabet. Typically, these texts are of “mundane” nature, containing lists of

names, inventories of food and other items, taxation records, as well as day-to-day administrative and military correspondence. However, with the primarily religious, literary and executive documents written on papyri, and therefore not surviving the journey down the millennia in the local humid climate, the ostraca are almost the sole remaining epigraphic evidence from the First Temple period. Hence, their utmost importance for historical, anthropological, linguistic, philological and religious studies of Israel and Judah during this era.

The practice of modern epigraphers (experts on ancient texts) specializing on Iron Age, comprises the following stages. An ostrakon (or its photograph) is manually drawn, resulting in a facsimile (black and white depiction of the document). Facsimiles of various ostraca are utilized for the purpose of creating a “paleographical table”, containing “representative” letter instances for each inscription. Subsequently, the paleographical table serves as a basis for various typological studies, which compare the handwritings’ similarities and discrepancies between different documents, corpora and localities, and attempt to trace the evolution of the letters across the ages. Naturally, such procedures are extremely labor-intensive, and moreover, face the almost certain risk of unintentionally mixing documentation and interpretation. An example of an epigraphic procedure for ostrakon No. 1 from Tel Arad can be seen at Fig. 1.2. It begins with an ostrakon (Fig. 1.2a), depicted in a manually created facsimile (Fig. 1.2b). Unfortunately, a close inspection of the facsimile shows a mixture of documentation and interpretation (Fig. 1.2c). Then, the most representative characters, chosen by the epigrapher, populate the paleographic table (Fig. 1.2d), utilized for further tasks of typological analysis.

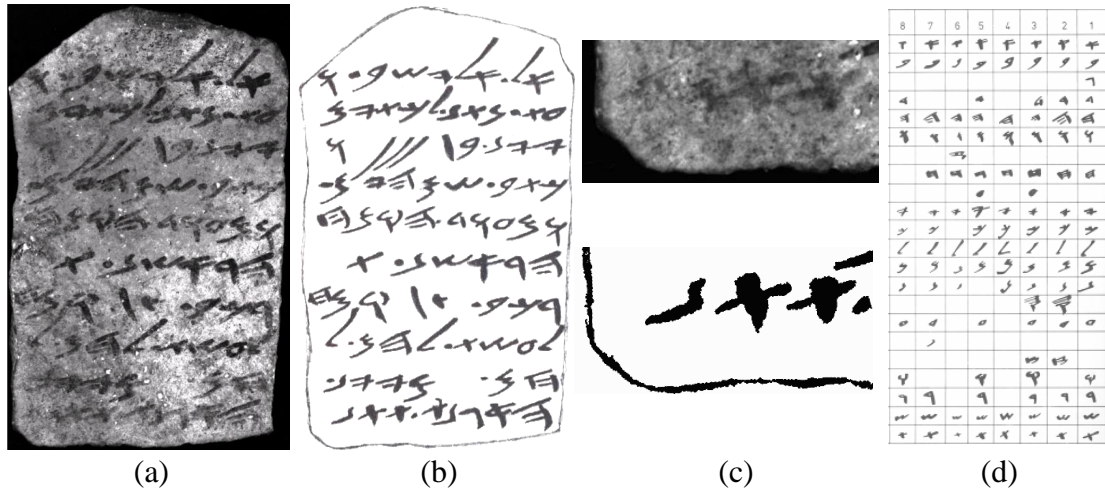


Figure 1.2 Ostracon No. 1 from Tel Arad: (a) an ostracon image; (b) hand drawn facsimile; (c) zoom-in on image and facsimile, the utmost left word of the last line. The leftmost “nun” character is documented, yet it is absent upon close inspection; (d) a fragment of a paleographic table, containing “representative” letters from different ostraca.

The main goal of this study is establishing a computerized paleographic framework for dealing with First Temple period epigraphic material. This toolbox can be compared with other similar projects and toolkits dealing with historical documents of other languages, eras and writing systems. Examples include the Gamera project (Droettboom et al. 2012), the Hadara framework for historical Arabic documents (Pantke et al. 2013), the Monk handwritten documents engine (Van der Zant et al. 2009; Van Oosten and Schomaker 2014; Schomaker 2016); as well as several ventures dealing with Hebrew writing from other ages and media, in particular the Dead Sea Scrolls (Grossman 2010; Lavee 2013; Dead Sea Scrolls 2016) and the Cairo Genizah (Wolf et al. 2010; Potikha 2011). Undoubtedly, inspiration can, and will be drawn below from these and many other references. However, the distinctive challenges (e.g., small amount of very short, fragmentary and highly degraded texts; unskilled authors with significant intra- and inter-writer characters’ variability; stained, cracked, uneven, nonuniform, fluorescent and difficult to image medium; many hotly debated issues among epigraphers, leading to the absence of any agreed-upon “ground-truths”), as well

as the unique research questions related to Hebrew Iron Age epigraphy, necessitate the development of an original computational apparatus.

Below is a concise description of the major research questions, handled by the corresponding sections of the thesis.

- **Section 2:** How can the quality of manually created facsimiles be evaluated?
- **Section 3:** How can the quality of various images of the ostraca be evaluated?
- **Section 4:** How can automatic binarizations, possibly encompassing the beneficial information of manual inexact facsimiles, be created?
- **Section 5:** How can binarizations be improved via sparse methods?
- **Section 6:** How can the quality of binarizations be evaluated?
- **Section 7:** How can the quality of individual characters within the binarizations be evaluated?
- **Sections 8 and 9:** How can different writers be detected within a given corpus?
(Two distinct methods are proposed.)
- **Section 10:** How can a fast image segmentation be achieved?
- **Section 11:** How can a letter's prior be estimated?

A schematic flowchart of the overall framework is provided in Fig. 1.3.

We aimed at making each section of the thesis as self-contained as possible, with links to other sections supplied whenever necessary. Some of the following results were previously presented in papers quoted below, as well as within some brief overview articles (Faigenbaum-Golovin, Shaus, Sober et al. 2015; Shaus et al. 2016a; Faigenbaum-Golovin, Shaus, Sober et al. 2017). This thesis refines, improves, finalizes and connects these developments.

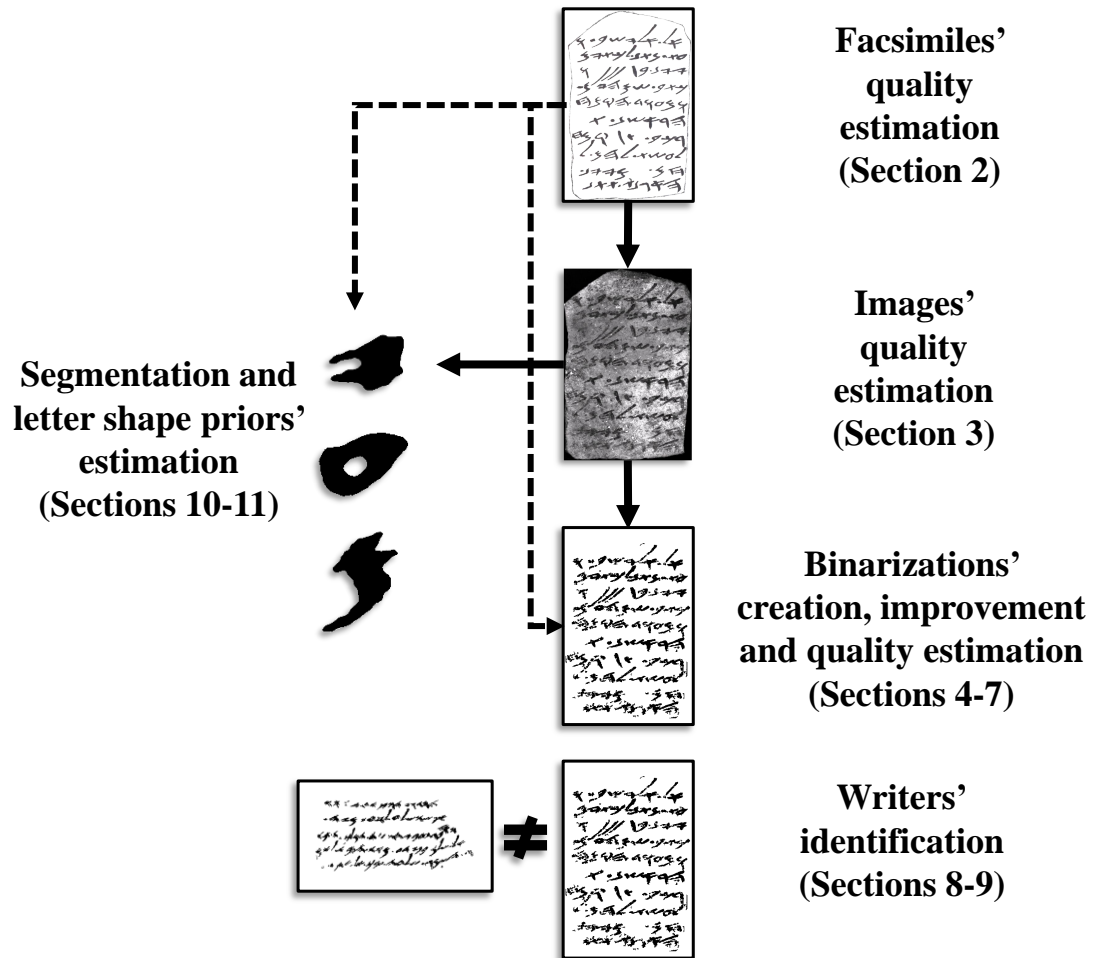


Figure 1.3 A summarizing schematic flowchart of the overall framework. Continuous lines represent direct input, while dotted lines represent auxiliary information.

Unless stated otherwise, all the methods were implemented by the author of the thesis via the Python programming language (Python 2010), utilizing libraries such as NumPy (Van der Walt et al. 2011), SciPy (Jones et al. 2001), scikit-learn (Pedregosa et al. 2011), scikit-image (Van der Walt et al. 2014), Matplotlib (Hunter 2007), PIL (PIL 2009) and Pillow (Pillow 2010).

2. Quality Evaluation of Manually Created Facsimiles

2.1 Background and Prior Art

The discipline of Iron Age epigraphy relies heavily on manually-drawn facsimiles (binary documents) of 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 facsimile quality evaluation. Some epigraphical publications (e.g. Hunt et al. 2001 and Barkay et al. 2003; although they do not deal with ostraca) superimpose the facsimile over the inscription image, but this is performed manually with no attempt at measuring the quality of the fit. On the other hand, various methods from the domain of document analysis, dealing with quality estimation of binarizations (e.g. Ntirogiannis et al. 2008, Pratikakis et al. 2010, Gatos et al. 2011), require the creation of a manual or semi-automatic ground truth (which can be potentially influenced by the human factor, see Barney Smith 2010). Candidate binarized images (facsimiles) are then graded in one way or another, according to the quality of their fit to the ground truth, with no reference to the inscription image.

As an alternative, we shall establish an effective facsimile quality measure, simple enough to be explained to epigraphers. The measure will be based upon registering the facsimile directly to an inscription image (kept constant). The performance of the measure will be tested in order to assess its reliability. The overall approach was first presented in (Shaus et al. 2010, 2012).

2.2 Facsimile Evaluation

Given a gray-scale $O(p)$ ostracon image, and the facsimile image $F(p)$, (in both cases $p \in [1, m] \times [1, n]$) several image-fit functions can be defined (as will be explained later, given images of different sizes, a registration of the facsimile image to the ostracon image is required). Natural candidates for comparing different versions are the commonly used L_1 and L_2 norms. While the latter may entail nice analytic properties, it also has the tendency to heavily penalize large deviations, which might lead to non-robust behavior. Thus, we prefer the L_1 norm. Since the facsimile documents are binary we denote $I = \{p \mid F(p) = 0\}$ (“ink pixels”) and $C = \{p \mid F(p) = 255\}$ (“clay pixels”), which will function as a partition of $O(p)$ induced by $F(p)$. We begin with the following measure which we wish to *minimize*:

$$E_1(F, O) := \sum_{p \in I \cup C} |F(p) - O(p)|. \quad (2.1)$$

As the facsimile image is restricted to 0=ink and 255=clay values, it is easy to show that *minimizing* E_1 is equivalent to *maximizing*

$$E_2(F, O) := \sum_{p \in C} O(p) - \sum_{p \in I} O(p) - \sum_{p \in I \cup C} F(p). \quad (2.1)$$

It is expected that among the various facsimiles depicting a given inscription, the relative proportions of ink and clay pixels (as opposed to their location) would be almost constant. Thus, the rightmost sum can be neglected. A possible problem with this measure is the dominance of the left component over the middle one, as the “ink” pixels (within the facsimile image) are relatively rare. A more “egalitarian” approach is to use averages (i.e. $\mu_D = \text{Average}_{p \in D} \{O(p)\}$, where D is a domain within an image) instead of sums, thus biasing the measure towards the ink pixels. Define:

$$CMI(F, O) := \mu_c - \mu_l, \quad (2.3)$$

where μ_c is the average “clayness” while μ_l is the average “inkness”. The overall measure is abbreviated as CMI (“clayness minus inkness”). The CMI index exhibits a connection to the Otsu binarization measure (Otsu 1979), which is equivalent to:

$$\omega_0 \omega_1 (\mu_1 - \mu_0)^2. \quad (2.4)$$

Here, μ_0 and μ_1 are averages of the two pixel “populations” (in our case these are μ_c and μ_l) and ω_0 , $\omega_1 = 1 - \omega_0$ are their appropriate proportions. Since, $\omega_0 \omega_1$ reaches a maximum when $\omega_0 = \omega_1 = 0.5$, Otsu's criteria may be viewed as the square of the CMI measure, biased towards the histogram median (which splits the pixels into two populations of equal proportions). On the other hand, it should be noted that the underlying problems are quite different: while Otsu deals with an unknown pixel population separated by histogram thresholding, our mission is to evaluate existing pixel populations induced by another (facsimile) image.

The main difficulty of comparing two documents is that the manually-crafted facsimile may depict the ostrakon from a slightly different angle, or to be somewhat rotated with respect to the ostrakon image. For that reason, there arises the need for a registration between the facsimile image and the ostrakon image, resulting in the facsimile registered image fitting the dimensions of the ostrakon. For registration purposes, we use the same CMI target function. We design a registration that only allows for rotations $R_\theta(F)$ of the facsimile image, with subsequent height/width adjustments in order to impose the dimensions of the ostrakon image on the facsimile (the simplicity of the registration minimizes our intervention in the work of the epigrapher; see more sophisticated registration in Section 4). Therefore, the

optimization is only performed with respect to one parameter, the angle θ , (sampled herein with 0.1 degrees' resolution):

$$\theta_{\max} = \arg \max_{\theta} CMI(R_{\theta}(F), O). \quad (2.5)$$

An example of an ostracon image, a facsimile image, their initial fit and their CMI-based registration can be seen at Figs. 2.1 and 2.2 (depicting Arad ostracon No. 1, see Aharoni 1981).

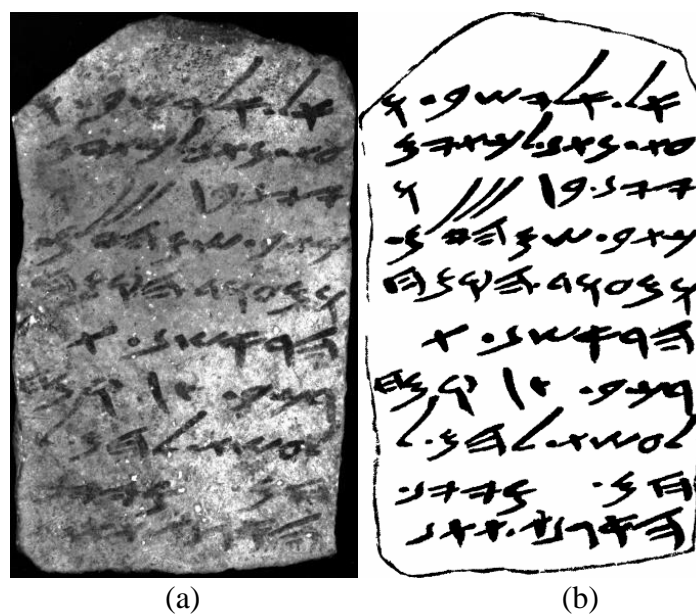


Figure 2.1 Example of (a) an ostracon image, (b) a facsimile image



Figure 2.2 Example of (a) initial facsimile-ostracon fit, (b) CMI-based registration

2.3 Experimental Results

Methodology Verification I

We compare several facsimiles of the same Arad No. 34 ostrakon (containing hieratic, i.e. Egyptian, numerals; see Aharoni 1981, Rollston 2006 and Fig. 2.3), created by different individuals. Two of the facsimiles were drawn by epigraphers and one by an artist. In order to avoid identifying these scholars, they are denoted below as A, B and C. The results of the CMI-based registration and evaluation can be seen in Figs. 2.4-2.6. They confirm the soundness of the approach.

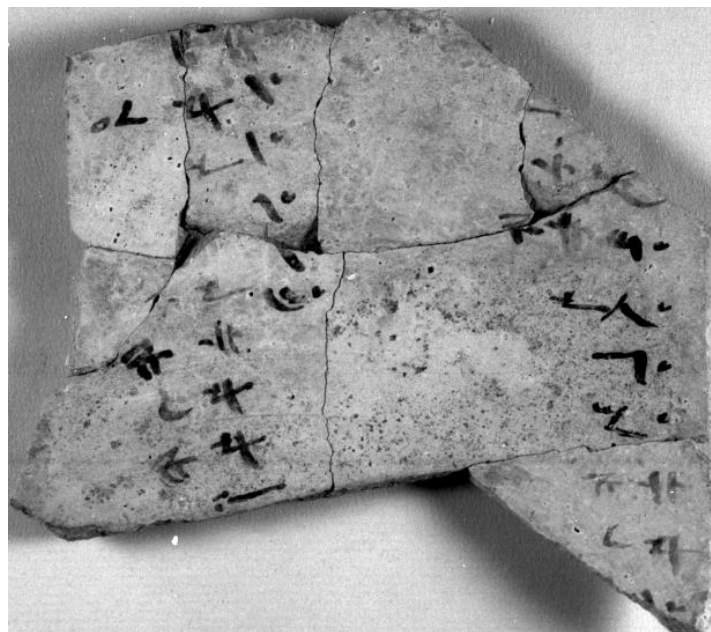


Figure 2.3 Arad ostrakon No. 34

Fig. 2.4 shows the ostrakon image compared with facsimile A. The registration of the facsimile is excellent (attesting to the effectiveness of the CMI measure). The overall fit of the overlaid facsimile A is good. Nevertheless, the facsimile characters are not always correlated with the ostrakon image characters (see for example Fig. 2.4 in the lower left), which results in typical “shadows” (un-obstructed ink). The strokes are not always long enough (e.g., Fig. 2.4, lower right). The strokes themselves are somewhat wide. The resulting CMI score is 71.1.

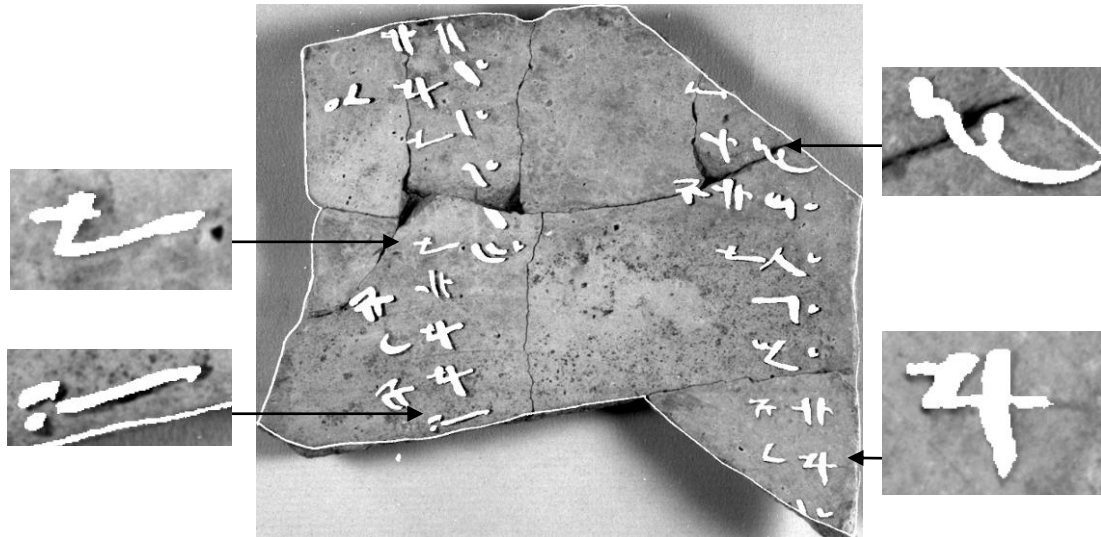


Figure 2.4 Overlaid facsimile A, CMI = 71.1

The overlaid facsimile B, seen on Fig. 2.5, again has good registration. This time, the fit is also good and the facsimile characters seem to be in better correlation with the ostracon image. On the other hand, the character strokes are sometimes a bit too wide (e.g., Fig. 2.5, upper left) and the overlap is not always perfect. In addition, notice cases where the strokes are not long enough (e.g., Fig. 2.5, upper left and lower right). Overall, the facsimile is of better quality than A. The CMI measure, 82.6, is understandably higher.

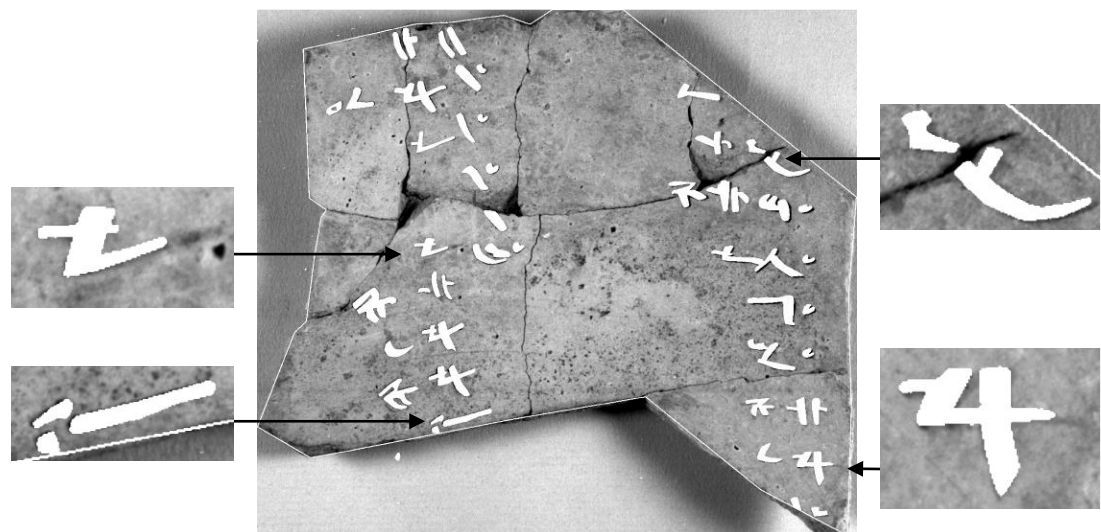


Figure 2.5 Overlaid facsimile B, CMI = 82.6

In the case of facsimile C, Fig. 2.6, the registration is outstanding. The characters are narrow and “crisp”; they seem to be in excellent agreement with the

ostracon image. The CMI score, 84.0, is justifiably the highest among the three facsimiles. This is despite one possibly missing character, taken for a scratch or stain (Fig. 2.6, upper right); owing to the fact that empirically, the CMI measure prefers mistaking ink for clay than vice versa (i.e. it is “conservative” with respect to “character-invention”, but will not heavily penalize for dropping dubious character).

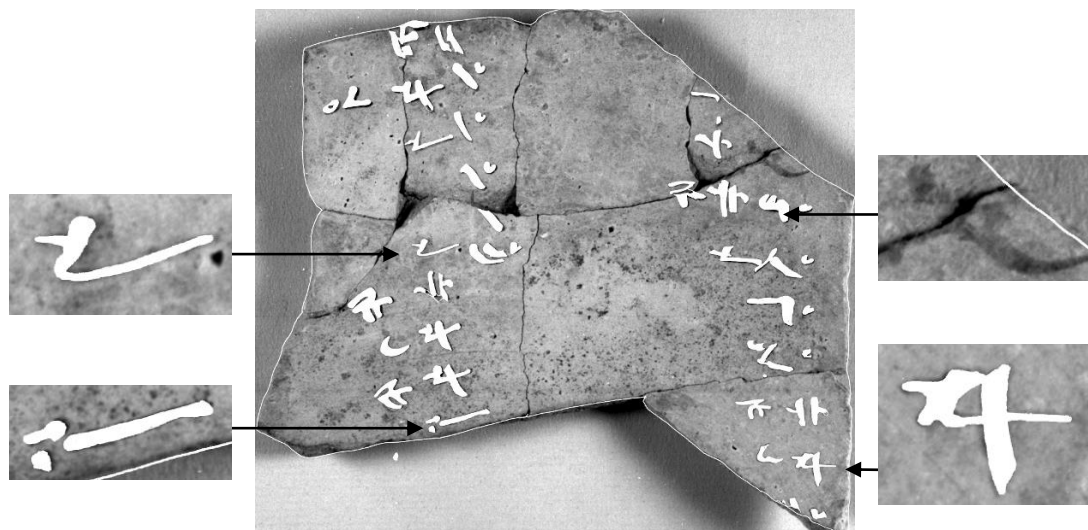


Figure 2.6 Overlaid facsimile C, CMI = 84.0

In conclusion, the procedure correctly indicates that facsimile C is the best of the three. The superb registration, also based on a CMI index, is also a good indicator of the soundness of this measure.

Methodology Verification II

It follows from the definition (Eq. 2.3), that the CMI index depends on the ostracon image. Hence, it is reasonable to assume that camera position and angle (vis-à-vis the ostraca), as well as illumination characteristics, are significant factors in the image and so may change not only the CMI scores, but also their relative rankings. In order to empirically test the degree of invariance of the CMI measurements and their rankings to ostracon image change, we used yet another image of the same ostracon, which can be seen in Fig. 2.7.

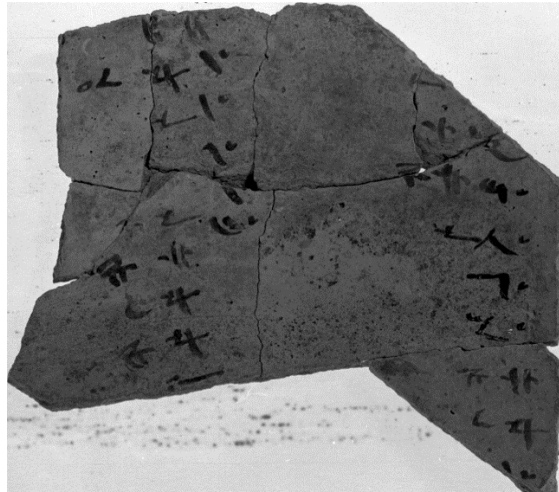


Figure 2.7 Another image of Arad ostracon No. 34.

Comparing Figs. 2.3 and 2.7, it is obvious that the latter image is markedly different from the former. It is viewed from a different angle, it is slightly rotated, the background is brighter and lacks shadows, and the ostracon itself is darker. We repeated the previous methodology verification stage, the protocol included the usage of the unchanged A, B and C facsimiles and an application of the same CMI registration and quality estimation apparatus. Table 2.1 summarizes the results of the first and the second methodology verifications.

Table 2.1 Results for two verifications of facsimile quality methodology

Facsimile	CMI score using Image #1	CMI score using Image #2
A	71.1	64.5
B	82.6	71.6
C	84.0	75.1

The change in the magnitude of the results is hardly surprising, as the image has a different grayscale level distribution. What is important is that the order of the CMI scores is maintained despite the completely different ostracon images. The A score is lower than B, while the C score is higher than both A and B. Therefore, despite using substantially different ostracon images, the relative results of the facsimile evaluation

remain effectively the same. This current empirical validation shows, that the facsimile rankings are fairly invariant under certain ostrakon image alterations.

2.4 Possible Drawbacks

Several shortcomings in the method and its verification ought to be mentioned:

- In any given quality assessment metric, some cases can lead to misleading results. The CMI index is no exception. As an illustration, assume an extremely faint character, with gray levels comparable to typical clay gray levels. In such a case, omitting the character from the facsimile might be preferable from the CMI index point of view. A compromise could be to draw only a silhouette of such a faint character (in fact, this is an accepted epigraphical practice). Another example is that of a dark stain. From the CMI index perspective it may be better to record it on the facsimile as if it were a character. As already stated, the CMI score is “conservative” with respect to “character-invention”, and is not expected to benefit substantially from the addition of a letter.
- The CMI-based evaluation depends on registration of the facsimile to the ostrakon image. We use a registration of a very simple type, which empirically works for our purpose. More sophisticated registrations can be considered (see Zitová and Flusser 2003 for a survey on the subject). Registering on a per-character basis, for instance (cf. Section 4), may lead to another quality measure and allow for low scale correction of the drawing. Such a method of registration may also compensate for nonlinear camera distortions.
- The results presented here were obtained from a limited number of test cases. In addition to these, we successfully experimented with several other ostraca (e.g. Fig.

2.8, Lachish ostracon No. 3) and tested the technique on different scales (1/4 and 1/8). Subsequent usages of the CMI score (see below) strengthened the confidence in this methodology.

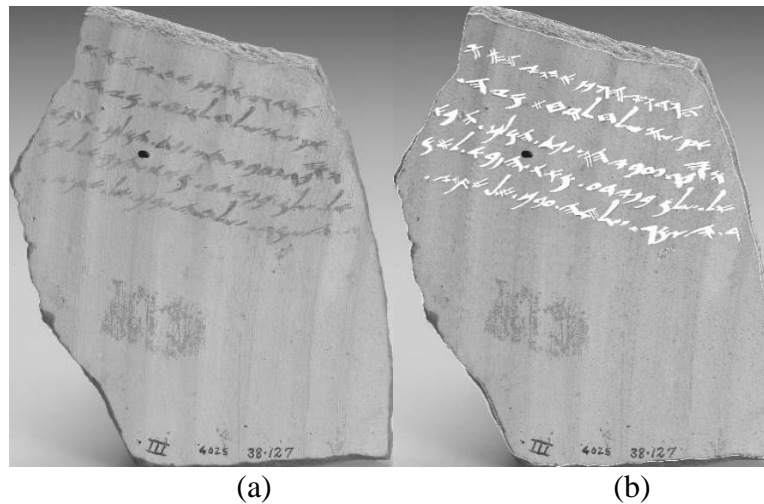


Figure 2.8. Another example of (a) ostracon image, Lachish ostracon No. 3; (b) a fit to a high-quality facsimile

2.5 Summary

We presented a facsimile (or binarization) quality measure (CMI), based upon registering the facsimile directly to an inscription image. The technique was tested on different ostraca, scales, facsimiles and ostraca images. The CMI grades received for the facsimiles reflect their relative merits. Based on the CMI scores, the rankings of the facsimiles are empirically invariant to the ostracon image. It can therefore be concluded that the proposed technique is sound and can be used to evaluate the accuracy of a facsimile in relation to the original ostracon.

Despite its apparent simplicity, the measure was extended and used as a basis for the purposes of quality evaluation of multispectral images (Section 3), automatic derivation of ostraca binarizations (Section 4), as well as quality evaluation of such binarizations (Section 6). Moreover, the facsimiles and their content were utilized as a rough “preliminary draft” for further analysis in Sections 8-11.

3. Potential Contrast Quality Measure with Application to Multispectral Imaging

3.1 The Problem

During the course of the research, many efforts were invested in obtaining the best possible images of the ostraca, utilized as inputs for subsequent algorithms described in this thesis. The most promising technique was the multispectral imaging method (see additional details in Faigenbaum et al. 2012). The investigation demonstrated that typically, the most favorable signal was obtained by imaging in particular wavelength, unique for each inscription. Dimension reduction techniques such as PCA, commonly applied in multispectral imagery context, turned out to be less favorable than selecting the most signal-saturated imaging band, with the remaining bands containing traces of the same signal with increasing noise levels. That basically reduces the problem to an allegedly easier one, that of choosing the most contrasted grayscale image from a given group of registered images - be that an RGB channel or one of 10 or 50 multispectral bands.

Establishing the contrast of an image is a well-studied problem in the fields of Optics and Image Processing. Several measures have been proposed, for that purpose, in the past. Among these are the contrast measures of Weber (Peli 1990), Michelson (Michelson 1927, Peli 1990), root-mean-square contrast and its enhancements (Pavel et al. 1987, Shio 1989), CMI (see previous section), as well as measures based on frequency domain analysis (Peli 1990, Li et al. 2009), wavelet transforms (Lai and Kuo 2000, Li et al. 2009) and edge detection (Leu 1992, Négrate 1992); see additional details regarding some of these methods below.

However, the problem is complicated by the immense set of transformations which can be applied to the image, potentially improving its contrast. Given a proliferation of the available Image Processing software solutions, applying such enhancements is almost indispensable. Therefore, the real challenge, which was not dealt with in the previous literature, is *measuring the contrast of an image taking into account all its possible transformations*. Herein, we will limit ourselves to finding an *analytical* solution to the wide range of *grayscale transformations*; the relevant results were published in (Shaus et al. 2017a).

3.2 Prior Art

Various algorithms were designed to give an objective contrast measure that correlates with human assessment. In what follows, we consider grayscale images of the form $I : [1, L] \times [1, M] \rightarrow [0, 255]$ (the intervals are assumed to be subsets of integers). We review several popular contrast measures, stating their relative shortcomings.

A simple way of measuring a bi-population image contrast is calculating the ratio between foreground and background:

$$\text{SimpleContrast} := \mu_B / \mu_F, \quad (3.1)$$

where μ_B and μ_F are the averages of the sampled background and foreground luminance values, respectively.

A more commonly used measure (closely related to *SimpleContrast*) is Weber's contrast ratio (Peli 1990) defined as:

$$Weber := \frac{\mu_B - \mu_F}{\mu_B} = 1 - \frac{1}{SimpleContrast}. \quad (3.2)$$

Another prominent contrast ratio measure is given by Michelson (Michelson 1927, Peli 1990):

$$Michelson_{\min \max} := \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \quad (3.3)$$

where I_{\max} and I_{\min} are the maximal and minimal luminance values of the entire image, respectively. This definition can be adapted to the case of bi-population images as follows:

$$Michelson := \frac{\mu_B - \mu_F}{\mu_B + \mu_F}, \quad (3.4)$$

The ratios in Eqs. 3.1, 3.2 and 3.4 result in different values for a single image. Nevertheless, given a set of images, the *ordering* based upon them will be identical. This can be verified via algebraic manipulations.

A different statistical approach is the root-mean-square contrast (Pavel et al. 1987):

$$RMS := \left[\frac{1}{LM} \sum_{l=1 \dots L, m=1 \dots M} (I(l, m) - \bar{I})^2 \right]^{1/2}, \quad (3.5)$$

where $I(l, m) \in [0, 1]$ is a normalized gray level and $\bar{I} = \frac{1}{LM} \sum_{l=1 \dots L, m=1 \dots M} I(l, m)$. Another, closely related statistical-based measure is suggested by (Shio 1989).

A very simple, yet valuable contrast measure, defined in Section 2 (and utilized in Sections 4-6), is the CMI index, restated here as:

$$CMI := \mu_B - \mu_F, \quad (3.6)$$

This measure will play an important role in the current section.

Some additional approaches are based on frequency domain analysis (e.g. Peli 1990, Li et al. 2009), wavelet transforms (e.g. Lai and Kuo 2000, Li et al. 2009) and edge detection (e.g. Leu 1992, Négrate 1992, which also deal with contrast improvements).

Popular image enhancements bear the potential of improving the image quality. These include brightening and darkening, histogram stretching and equalization - all performed by grayscale transformations. Unfortunately, all of the above-mentioned measures are affected, to some extent, by such transformations. For instance, the Weber and Michelson ratios are not invariant to grayscale shifts, the CMI is not invariant to grayscale rescalings, while all the measures are not invariant to histogram equalizations. This aspect is demonstrated in Fig. 3.1 and Table 3.1. The RMS seems relatively stable with respect to most of the grayscale transformations, except for histogram equalizations. Unfortunately, although its definition represents the standard deviation of the image, which is an important statistic, the RMS does not quantify the quality of separation between foreground and background. Indeed, random permutation of pixels within the image would yield the same RMS value.



(a)



(b)



(c)



(d)

Figure 3.1 Example of images undergoing grayscale transformations.

(a) original image with sampled foreground (in red) and background (in blue).

(b) the image after brightness change (+70).

(c) the image after histogram rescaling ($\times 1.3$).

(d) the image after histogram equalization.

Table 3.1 Contrast measures comparison based on Fig. 3.1.

Image	Weber	Michelson	RMS	CMI
(a) Original I	0.535	0.365	1.42×10^{-4}	90.6
(b) Brightened $I+70$	0.378	0.233	1.42×10^{-4}	90.6
(c) Rescaled $I \cdot 1.3$	0.536	0.366	1.43×10^{-4}	117.7
(d) Equalized $Eq(I)$	0.33	0.197	1.27×10^{-4}	72.1

3.3 Requirements and Measure Definition

Requirements

Given a contrast measure m , and an image I , the task is finding a grayscale transformation $g \in G := \{[0, 255] \rightarrow [0, 255]\}$ maximizing $m(g \circ I)$. At first glance, this may seem as a computational-intensive undertaking, since the set of transformations of a given image is immense ($2^{2^{B+\log_2 B}}$ for images of bit-depth B). The main contribution of this section is a constructive procedure for finding the optimal transformation g *analytically*, for a particular measure m . This would lead to a definition of a new, “*potential*” contrast measure, possessing the following properties:

1. Quantifying the difference between foreground and background pixels (i.e. the measure is a meaningful one).
2. Images will be judged according to their **potential** for improvements via **all possible** grayscale transformations (i.e. the measure is “aware” of the possibility to perform image enhancements such as brightening, rescaling and equalizing its grayscale levels).
3. In particular, the measure ought to be invariant to **invertible** grayscale transformations (as the inherent information of the image is preserved and the potential for image improvement after such transformation is maintained).

Assumptions

In order to deal with this problem analytically, we restrict ourselves to the CMI measure defined in Eq. 3.6, $m = CMI$ (the analysis presented below will not hold for others measures we're aware of). Furthermore, we deal with a case of sampled histograms (“populations”) of foreground and background pixels, as is observed in Fig.

3.1a. These are respectively denoted as $\{p_F(t)\}_{t=0}^{255}$ and $\{p_B(t)\}_{t=0}^{255}$ (satisfying

$$0 \leq p_F(t) \leq 1, 0 \leq p_B(t) \leq 1 \text{ and } \sum_{t=0}^{255} p_F(t) = \sum_{t=0}^{255} p_B(t) = 1).$$

We begin with finding the maximal $CMI(g \circ I)$ for an image I , with the wealth of optional grayscale transformations g , proceeding with the definition of a new measure.

Proposition I (Optimality)

For a given image I , with sampled populations $\{p_F(t)\}_{t=0}^{255}$ and $\{p_B(t)\}_{t=0}^{255}$ (as denoted above), the optimal grayscale transformation with respect to the CMI measure is:

$$g_I^{opt}(t) := \arg \max_{g \in G} CMI(g \circ I) = \begin{cases} 0 & p_F(t) > p_B(t) \\ 255 & p_F(t) \leq p_B(t) \end{cases}. \quad (3.7)$$

Proof:

$$\begin{aligned} CMI(g \circ I) &= \sum_{t=0}^{255} g(t)p_B(t) - \sum_{t=0}^{255} g(t)p_F(t) = \sum_{t=0}^{255} g(t)[p_B(t) - p_F(t)] \leq \\ &\leq \sum_{\substack{t=0, \\ p_B(t) \geq p_F(t)}}^{255} 255 \cdot [p_B(t) - p_F(t)] + \sum_{\substack{t=0, \\ p_B(t) < p_F(t)}}^{255} 0 \cdot [p_B(t) - p_F(t)] = \\ &= \sum_{t=0}^{255} g_I^{opt}(t)p_B(t) - \sum_{t=0}^{255} g_I^{opt}(t)p_F(t) = CMI(g_I^{opt} \circ I) \end{aligned}$$

■

Definition of Potential Contrast

The *Potential Contrast* (PC) of an image is:

$$PC(I) := CMI(g_I^{opt} \circ I). \quad (3.8)$$

Remarks:

1. Due to its nature, the PC measure reflects the innate image quality, not necessarily compatible with immediate human impression. Consider a pair of images created from the same source (Fig. 3.1a), one with added Gaussian noise (Fig. 3.2a), while the other brightened to some extent (Fig. 3.2b). Although the former may be viewed as more contrasted, in fact the latter has considerably higher Potential Contrast (PC=206.28 vs. PC=255.0). This is due to the fact that it possesses the same information as the original image, unlike the image with Gaussian noise.
2. Foreground and background selection can be performed in numerous ways. These choices represent diverse, often incompatible, needs of human operators. For example, in Fig. 3.3, what are the expected foreground and background? Are they respectively the kettle and the chair? Or maybe the writing and the whiteboard? Therefore, in our view, no “ultimate” background and foreground selections encompassing all feasible tasks can be defined. This explains our preference for sampled foreground and background populations – the foreground and the background are in the eyes of the beholder. Nonetheless, a “naïve” suggestion for automatic foreground and background estimation is proposed below.



(a)

(b)

Figure 3.2 An example of misleading naked eye: Two images stemming from the same source, with the same sampled populations (Fig. 3.1a).

(a) added Gaussian noise of $\mu=0, \sigma=32$, **PC=206.28**

(b) narrowing the dynamic range and brightening ($I/4+200$), **PC=255.00**.

3. The CMI was chosen as a basis for the Potential Contrast definition due to the possibility of optimizing analytically the measure for all possible grayscale transformations. We did not succeed to similarly utilize other measures.

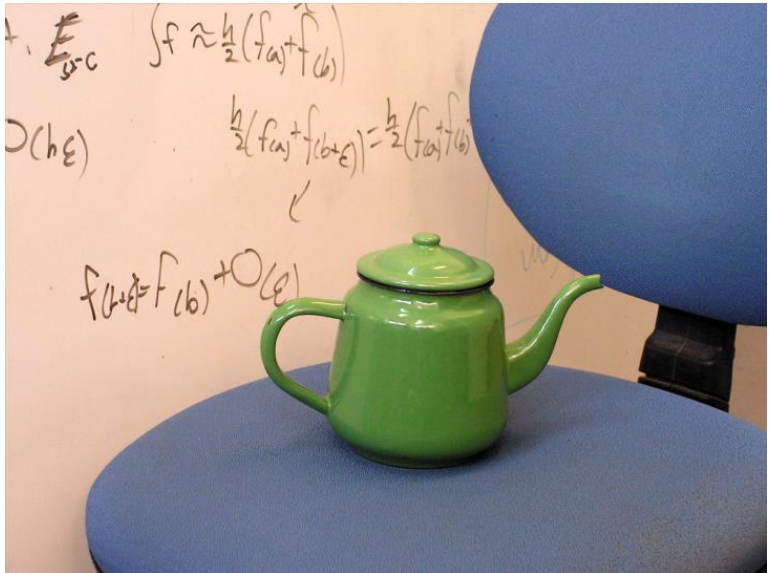


Figure 3.3 Example of ambiguous foreground and background. While it is possible that the kettle is the foreground and the chair is the background, writing as a foreground and whiteboard as a background is another viable option.

3.4 Measure Properties

Population Separability

The optimal grayscale transformation g_I^{opt} may be viewed as a function separating between foreground and background populations. This function serves as a *classifier*, denoted herein as *PC-binarization*. If the populations are separable by a certain threshold (e.g. two non-overlapping modes), the function can be represented as:

$$g_I^{opt}(t) = \begin{cases} 0 & t \leq T \\ 255 & t > T \end{cases} \quad (3.9)$$

However, this is not the general case (which can be seen in Eq. 3.7). Fig. 3.4 provides an example of grayscale histogram not separable by thresholding, while easily classifiable by the PC framework.

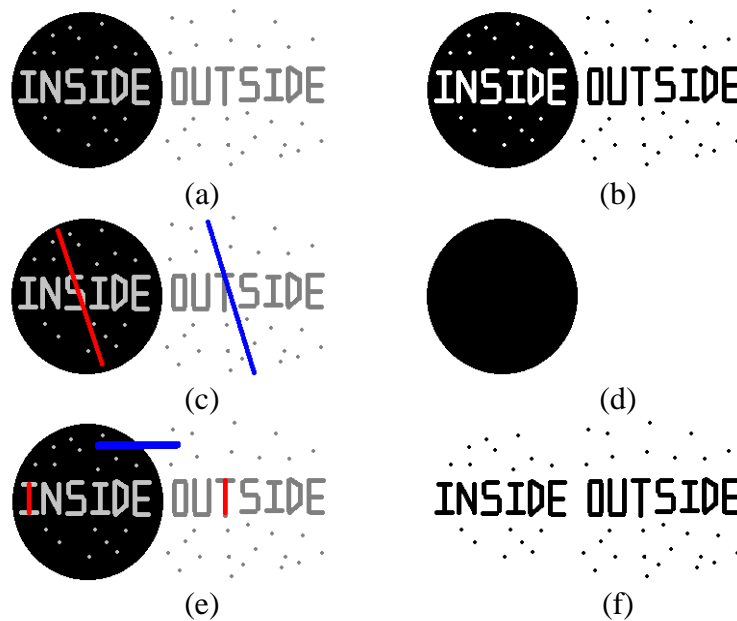


Figure 3.4 Example of foreground and background not separable by thresholding, while easily classifiable by the PC framework. (a) original grayscale image (circle=0, writing within the circle=195, writings outside the circle=127, other areas outside the circle=255); (b) example of an image thresholded by 150; (c) circle and its content as foreground (in red) with the rest as background (in blue); (d) PC-binarization based on (c); (e) writing as foreground (in red) with the rest as background (in blue); (f) PC-binarization based on (e).

In fact, even a slight difference in gray levels between the two populations may suffice to achieve a reasonable separation, i.e. binarization. See an example of “challenging” contrast enhancement in Fig. 3.5, based on the RGB decomposition of the original image, with several resulting PC-binarizations.

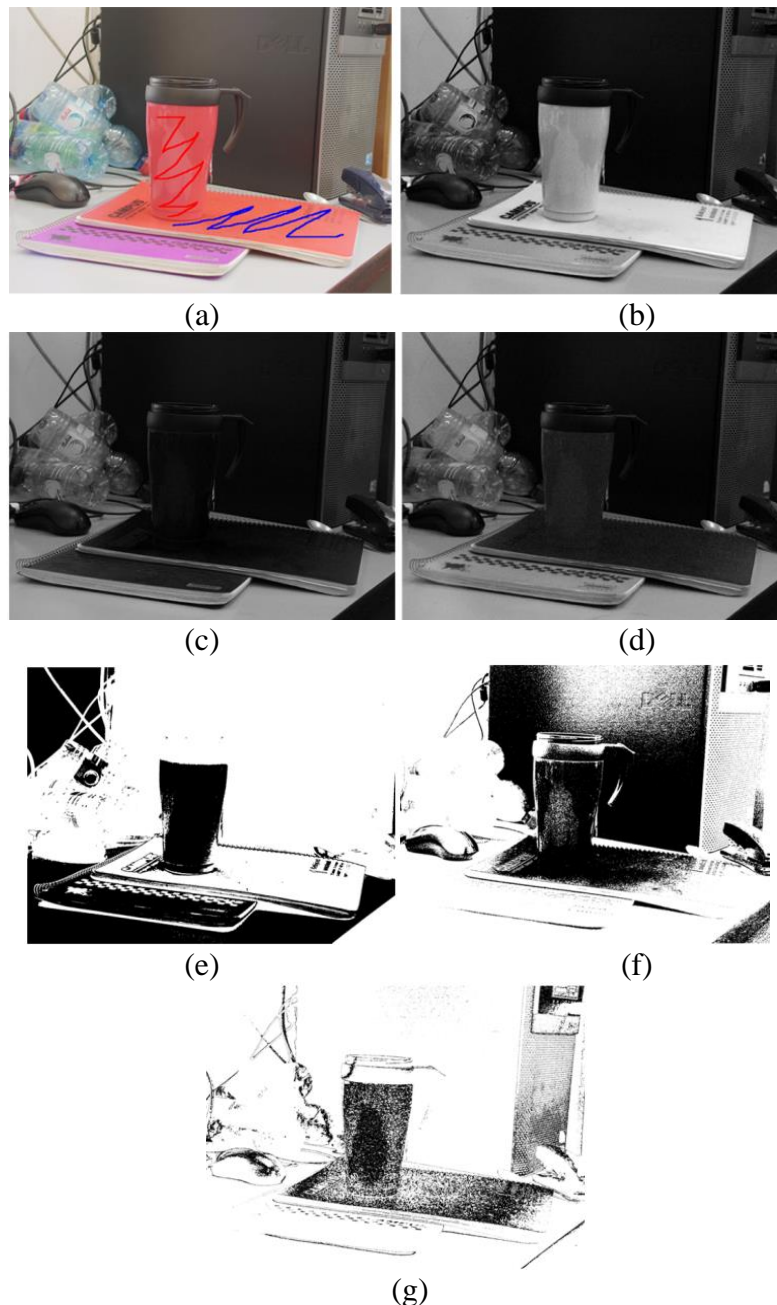


Figure 3.5 A natural scene handled by our method. A good contrast enhancement is achieved despite the similarity in foreground and background shade. (a) RGB image of the scene with manual selection of foreground in red and background in blue; R (b), G (c) and B (d) channels, with respective PC values of 244.8, 67.6 and 61.2; the PC-binarizations for R (e), G (f) and B (g).

Complexity

The calculation of foreground and background histograms is linear in the number of pixels ML , which tends to be small. The construction of g_I^{opt} is only dependent on the number of levels in the histogram. Therefore, for a grayscale image of 256 levels, the overall complexity is $O(ML + 256)$. Hence, the complexity is linear with respect to the number of pixels.

Equivalence to Error Estimation

The PC measure can be viewed as a measure minimizing the rate of false positives (FP) and false negatives (FN) mistakes, i.e. confusing foreground for background and vice-versa. This follows from the fact that:

$$\begin{aligned} PC(I) &= \sum_{\substack{t=0, \\ p_B(t) \geq p_F(t)}}^{255} 255 \cdot [p_B(t) - p_F(t)] = \\ &= 255 - \sum_{\substack{t=0, \\ p_B(t) < p_F(t)}}^{255} 255 \cdot p_B(t) - \sum_{\substack{t=0, \\ p_B(t) \geq p_F(t)}}^{255} 255 \cdot p_F(t) = 255 \cdot (1 - FP - FN) \end{aligned}$$

In the case of perfect separability of populations, the PC would be maximal, i.e. 255.

Note: this is the case in Figs. 3.2b, 3.4c and 3.4e.

Symmetry between Foreground and Background

The last property proves that if we replace the foreground sampled histogram with the background sampled histogram and vice-versa, the result of the PC measure is the same. On the other hand, the respective PC-binarizations would be each other's negatives.

Proposition II (Invariance with Respect to Invertible g)

Given an image I , and an invertible $g \in G$, $PC(I) = PC(g \circ I)$.

Proof:

g is invertible, therefore $\exists g^{-1} \in G$. $g^{-1} \circ g = \text{identity}$. Thus:

$$PC(I) = CMI(g_I^{opt} \circ I) = CMI(g_I^{opt} \circ g^{-1} \circ g \circ I)$$

Denoting: $h := g_I^{opt} \circ g^{-1}$ and $J := g \circ I$:

$$PC(I) = CMI(h \circ J)$$

Assuming $h \neq g_J^{opt}$, then:

$$PC(I) = CMI(h \circ J) < PC(J) = CMI(g_J^{opt} \circ J) = CMI(g_J^{opt} \circ g \circ I)$$

A contradiction to the optimality of g_I^{opt} . Therefore, $PC(I) = PC(g \circ I)$.

■

Remark: This defines the following equivalence relation between two images:

$$I_1 \sim I_2 \Leftrightarrow \exists g \text{ invertible s.t. } I_1 = g \circ I_2$$

The invariance property of the PC, with respect to the images of Fig. 3.1, is demonstrated in Table 3.2. This supplements and contrasts with the results in Table 3.1.

Table 3.2 PC measure based on Fig. 3.1.

Image		PC
(a) Original	I	255.00
(b) Brightened	$I+70$	255.00
(c) Rescaled	$I \cdot 1.3$	255.00
(d) Equalized	$Eq(I)$	254.98

3.5 Automated Foreground/Background Selection

As stated above, the selection of foreground and background largely depends on the specific task and usage scenario. Nevertheless, one generic approach would be to utilize one of the existing saliency estimation techniques. Fortunately, a useful and enlightening comparison of the leading saliency methods is presented in (Bylinskii et al. 2016). Surprisingly, among the “leading” saliency methods is a simple saliency map dependent on the distance of each pixel from the center of the image. In this estimation, 255 (the most salient value) is assigned to the central pixels, while 0 (the least salient value) is assigned to its corners. The empirical success of this unsophisticated technique probably has to do with either conscious or unconscious preference of human photographers for images centered on the object of their interest.

Despite (Bylinskii et al. 2016) claim of using a Gaussian model in this estimation, a reverse-engineering of their saliency image reveals a replacement of the Gaussian with a second-order polynomial approximation. In particular, given an image $I(x, y) : [1, L] \times [1, M] \rightarrow [0, 255]$, the saliency (i.e. foreground) map $S(x, y) : [1, L] \times [1, M] \rightarrow [0, 255]$ is constructed via the following formula:

$$S(x, y) = 255 \cdot \left(1 - \frac{1}{2} \left(\left(\frac{x - L/2}{L/2} \right)^2 + \left(\frac{y - M/2}{M/2} \right)^2 \right) \right). \quad (3.10)$$

It is easy to see that this formula satisfies $S(0, 0) = S(L, 0) = S(0, M) = S(L, M) = 0$, $S(L/2, M/2) = 255$, as well as $0 \leq S(x, y) \leq 255$. Examples of such a saliency map used for the foreground, as well as its complimentary $255 - S(x, y)$ used for the background, can be seen in Fig. 3.6.

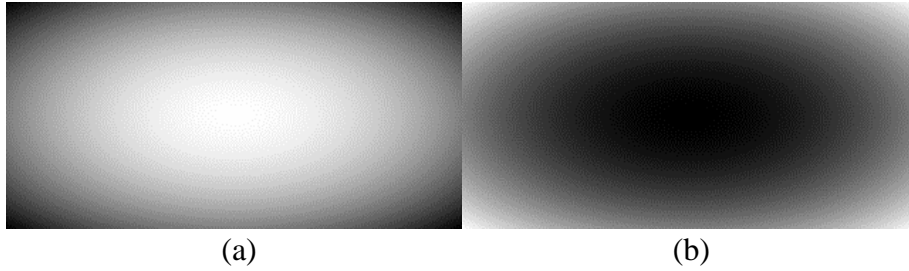


Figure 3.6 An example of automatically created saliency-based foreground (a) and background (b) maps.

Naturally, utilization of such continuous maps comes with the small price of adapting the measures. Indeed, apart from RMS (which does not rely on either the foreground or the background), all the measures utilize “crisp” definitions of the foreground and background populations. Fortunately, the measures’ definitions can be easily adapted for a “fuzzy” case, in which each pixel belongs to both the foreground and the background with a certain probability (in fact $S(x, y) / 255$ for foreground and $(255 - S(x, y)) / 255$ for background). E.g., μ_F and μ_B now become weighted means, while $\{P_F(t)\}_{t=0}^{255}$ and $\{P_B(t)\}_{t=0}^{255}$ represent weighted histograms over the entire image – maintaining the properties of the PC measure.

3.6 Experimental Results

The purpose of the following experiments is to empirically validate the behavior of the various contrast metrics including the Potential Contrast, with an emphasis on their invariant properties. The experiment consisted of the following steps:

1. The input for the experiments were images belonging to the popular GRAZ-02 data set, containing natural images (Opelt 2006). This included all images under the categories “bike”, “car” and “person”, which possessed a ground truth. With 300 files in each category, this resulted in 900 files.

2. If needed, each image was converted to grayscale by averaging its channels (e.g., $I(x, y) = (R(x, y) + G(x, y) + B(x, y)) / 3$). The histogram of the result was rescaled between 25 and 230 (maintaining the full dynamic range in transformations applied in the next step). This rescaled image is denoted herein as “initial” image.
3. Various gray-level transformations were applied to the “initial” image. This resulted in 6 additional images for each “initial” image. The transformations in use were: negative of an image, addition of 25, subtraction of 25, multiplication by 1.1, histogram stretching (from 0 to 255), and histogram equalization (from 0 to 255). In total, further $900 \times 6 = 5400$ images were obtained.
4. Five contrast measures (Weber, Michelson, RMS, CMI and PC) were applied on all the images (“initial” and transformations). The calculation used either marked background and foreground (utilizing ground truths from Opelt 2006), or an automated foreground and background selection scheme, as described above (the results for these two types of experiments are presented separately below).
5. For a given measure, the result for each transformation was divided by the result of the “initial” image, in order to obtain a “ratio of change” (e.g., if a given measure results in 2.718 on “initial” image, and in 3.14 on a transformed one, the division produces a ratio of 1.1557).
6. Ratios within the range of [0.99,1.01] were marked as indicating “invariance” of the measure with respect to a particular transformation, while others were counted as “non-invariant” outcomes. The percentage of the “invariant” ratios was calculated.

Experiment Results for Manual Foreground and Background Selection

The results in Table 3.3 were achieved by using existing ground truths, marking foreground and background. As expected, the most invariant and well-behaving metrics

are RMS and Potential Contrast. However, only the latter holds an almost-perfect invariance on histogram equalization transformation, whose non-linearity breaks the RMS record.

Table 3.3 Manual foreground and background selection: Ratios between the measures of transformed images with respect to “initial” image (predicted invariance marked in red).

Transformation	Statistics	Weber	Michelson	RMS	CMI	PC
Negative	Minimum	-2.8741	-1.7535	1	-1	1
	Maximum	-0.1468	-0.1842	1	-1	1
	Average	-0.8913	-0.7291	1	-1	1
	Invariance %	0.0%	0.0%	100.0%	0.0%	100.0%
+25	Minimum	0.5663	0.6134	1	1	1
	Maximum	0.8833	0.8666	1	1	1
	Average	0.8167	0.8032	1	1	1
	Invariance %	0.0%	0.0%	100.0%	100.0%	100.0%
-25	Minimum	1.1523	1.1820	1	1	1
	Maximum	4.2715	2.7046	1	1	1
	Average	1.3132	1.3391	1	1	1
	Invariance %	0.0%	0.0%	100.0%	100.0%	100.0%
×1.1	Minimum	1	1	1	1.1	1
	Maximum	1	1	1	1.1	1
	Average	1	1	1	1.1	1
	Invariance %	100.0%	100.0%	100.0%	0.0%	100.0%
Histogram stretching	Minimum	1.1523	1.1820	1	1.2439	1
	Maximum	4.2715	2.7046	1	1.2439	1
	Average	1.3132	1.3391	1	1.2439	1
	Invariance %	0.0%	0.0%	100.0%	0.0%	100.0%
Histogram equalization	Minimum	-99.4991	-102.5043	0.7581	-100.0948	0.9727
	Maximum	20.0625	19.6348	4.5870	19.2820	1.0000
	Average	1.3029	1.4134	1.5294	1.5560	0.9983
	Invariance %	0.7%	0.8%	1.1%	0.6%	98.7%

Experiment Results for Automated Foreground and Background Estimation

The results, which can be seen in Table 3.4, were achieved by using automated foreground and background estimation. Since this experiment is based on an estimated foreground and background, which may be quite far from a clear-cut partition of an image, the outcomes are expected to be less numerically stable. Indeed, the results for many transformations are much more spread-out. Nevertheless, yet again, the challenging histogram equalization provides a clear winner. In fact, it doesn't seem that

the stability of Potential Contrast was significantly hampered by the inaccuracy and fuzziness in the foreground and background selection.

Table 3.4 Automatic foreground and background estimation: Ratios between the measures of transformed images with respect to “initial” image (predicted invariance marked in red).

Transformation	Statistics	Weber	Michelson	RMS	CMI	PC
Negative	Minimum	-2.0264	-1.7467	1	-1	1
	Maximum	-0.1561	-0.1679	1	-1	1
	Average	-0.8588	-0.8406	1	-1	1
	Invariance %	0.0%	0.0%	100.0%	0.0%	100.0%
+25	Minimum	0.5794	0.5945	1	1	1
	Maximum	0.8723	0.8664	1	1	1
	Average	0.8143	0.8138	1	1	1
	Invariance %	0.0%	0.0%	100.0%	100.0%	100.0%
-25	Minimum	1.1715	1.1823	1	1	1
	Maximum	3.6483	3.1459	1	1	1
	Average	1.3161	1.3144	1	1	1
	Invariance %	0.0%	0.0%	100.0%	100.0%	100.0%
x1.1	Minimum	0.2481	0.2481	0.9993	0.2730	1
	Maximum	1.0399	1.0399	1.0039	1.1443	1
	Average	0.9980	0.9980	1.0023	1.0983	1
	Invariance %	96.7%	96.7%	100.0%	0.0%	100.0%
Histogram stretching	Minimum	0.5342	0.5342	0.9993	0.5426	1
	Maximum	3.6398	3.1516	1.0033	3.0089	1
	Average	1.3175	1.3158	1.0001	1.2452	1
	Invariance %	0.0%	0.0%	100.0%	0.0%	100.0%
Histogram equalization	Minimum	-2977.8	-2740.18	0.7597	-2664.94	0.9718
	Maximum	351.1975	336.03	4.5821	326.7109	1
	Average	-0.8685	-0.6326	1.5308	-0.3027	0.9983
	Invariance %	0.1%	0.1%	1.0%	0.4%	99.1%

3.7 Application of the Methodology

The PC measure received extensive real-world usage, applied on multispectral imagery of large corpora of ancient inscriptions. The first problem included a selection of optimal wavelengths for multispectral imagery of Second Temple Period Dead Sea Scrolls (Dead Sea Scrolls 2016). See Fig. 3.7a for an example of such a scroll, with a correct channel automatically selected and binarized in Fig. 3.7b.

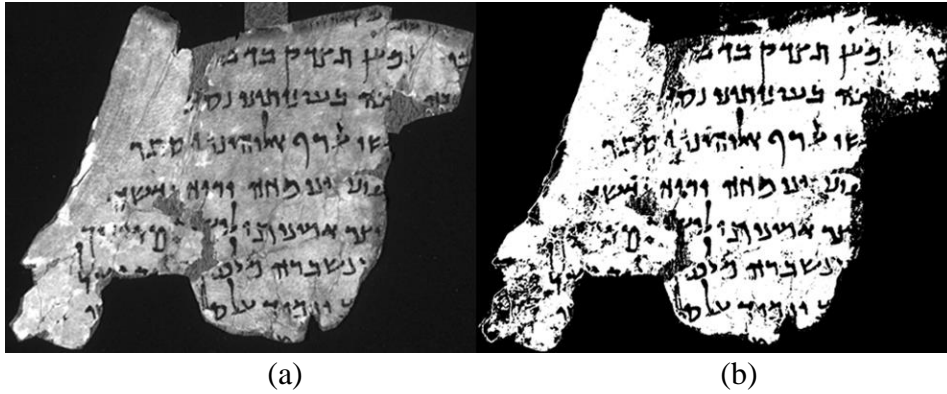


Figure 3.7 Section of Dead Sea scroll No. 124, fragment 001 (Dead Sea Scrolls 2016).

(a) Image of a scroll; (b) PC-binarization of (a).

Another test for our technique had to do with First Temple Period Hebrew, as well as Late Bronze Hieratic (cursive Egyptian) ink-on-clay inscriptions. These were unearthed during the excavations of Horvat Radum and Horvat Uza (Beit-Arieh 2007, Sober et al. 2014; see Figs. 3.8, 3.9), Tel Malhata (Beit-Arieh and Freud 2015, Faigenbaum et al. 2015), Qubur el-Walaydah (Faigenbaum et al. 2014; see Fig. 3.10), Jerusalem (Faigenbaum-Golovin et al. 2015), Arad (Faigenbaum-Golovin et al. 2017; Mendel-Geberovich et al. 2017; see Fig. 3.11) and Nahal Yarmut (Mendel-Geberovich, et al. forthcoming). The difficult and noisy medium of the ink written on pottery sherds presented a good opportunity to test the new methodology. Again, our task was to automatically select the “potentially” most contrasted image out of a spectral cube, in order to allow further analysis by human scholars. See Figs. 3.8-3.11 for examples of ostraca handled by our method, in order to find an optimal imaging wavelength. An elaboration of our experiments pertaining to this particular use case appears in (Faigenbaum et al. 2012).

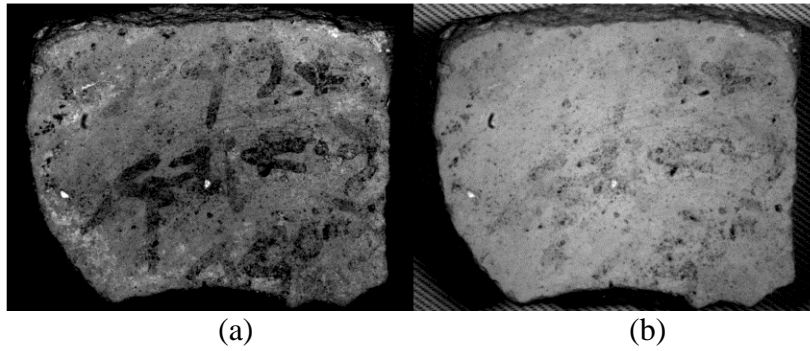


Figure 3.8 Images of Horvat Radum ostracon No. 1 (Beit-Arieh 2007, Sober et al. 2014). (a) optimal image at $\lambda=620$ nm, selected by our method; (b) sub-optimal image at $\lambda=950$ nm.

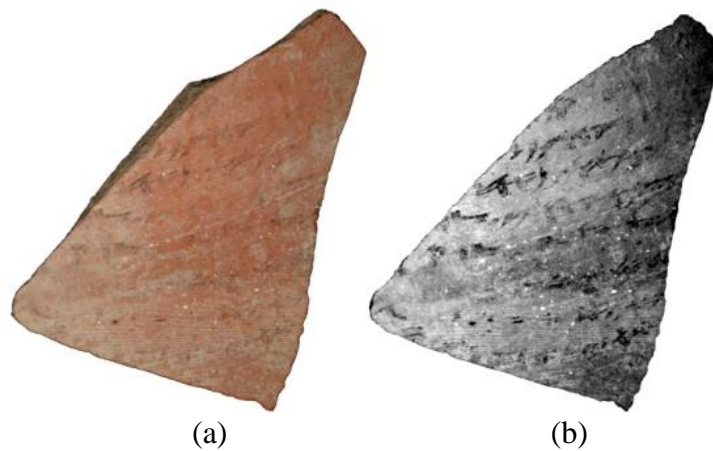


Figure 3.9 Images of Horvat Uza ostracon No. 3 (Beit-Arieh 2007, Sober et al. 2014). (a) RGB image; (b) multispectral image taken at $\lambda=660$ nm, selected by our method.

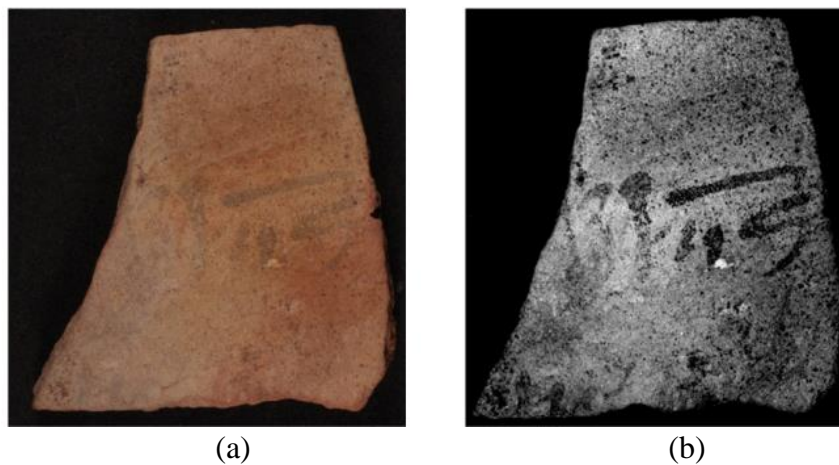


Figure 3.10 Images of ostracon No. 13.056-01-S01 from Qubur el-Walaydah (Faigenbaum et al. 2014). (a) RGB image; (b) multispectral image taken at $\lambda=690$ nm, selected by our method.

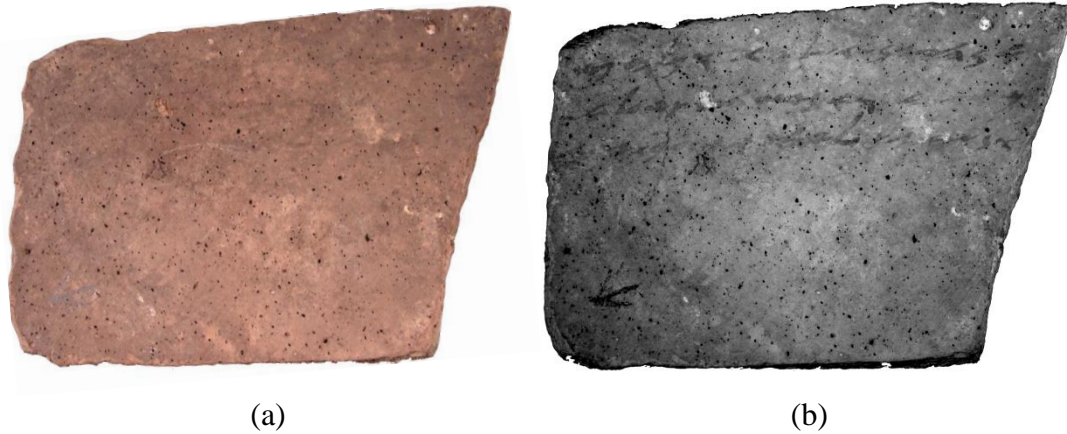


Figure 3.11 *Verso* of Arad Ostrakon 16. (a) current color image; (b) 890 nm image taken via our multi-spectral imaging system.

3.8 Summary

The current section presents a new approach for contrast estimation, necessitated by the need to choose the best ostraca multispectral band. Using available Image Processing software, an image can undergo various grayscale transformations, often improving its contrast. The common contrast evaluation methods, surveyed above, do not take this possibility into account.

Contrastingly, the Potential Contrast measure encompasses an analytic solution to the problem of finding the most contrasting grayscale transformation. The properties of the Potential Contrast were tested and compared to other measures on a large data set of 900 images, in two scenarios of foreground and background selection. The results indicate the invariance and the stability of the measure with respect to various grayscale transformations. The technique was applied on ancient inscriptions from various corpora with impressive results.

4. Binarization via Registration-based Scheme

4.1 Introduction

As previously noted, the discipline of Iron Age epigraphy relies heavily on manually-drawn, and thus conceivably biased, facsimiles of ostraca. Despite their importance, little attention has thus far been devoted to automatic creation of binarizations for Hebrew Iron Age ostraca.

In this section, we first survey the performance of several known computerized binarization techniques, either general-purpose (Otsu 1979; Bernsen 1986; and Niblack 1986), or specifically adapted for document analysis (White and Rohrer 1983; Sauvola and Pietikainen 2000; and Gatos et al. 2004). The resulting binarizations are found to be of insufficient quality for our purposes. 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 in Section 2), on both large and small scales. The performance of the new binarization will also be tested. The overall method was first presented in (Shaus et al. 2012b).

4.2 Prior Art

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 prominent binarization algorithms are considered. These include three general-purpose binarization algorithms with wide acceptance: Otsu (Otsu 1979), Bernsen (Bernsen 1986) and Niblack (Niblack 1986), as well as the White (White and Rohrer 1983),

Sauvola (Sauvola and Pietikainen 2000) and Gatos (Gatos et al. 2004) methods, which focus on the domain of document analysis, in particular in a low quality (e.g. historical) setting. The algorithms of White and Gatos were implemented via the Gamera toolkit (Droettboom et al. 2012).

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 (Sezgin and Sankur 2004; He et al. 2005; Gatos et al. 2009; Stathis et al. 2008a).

Otsu (Otsu 1979) maximizes the between-class variance criteria:

$$\omega_0\omega_1(\mu_1 - \mu_0)^2, \quad (4.1)$$

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

Bernsen's method (Bernsen 1986) 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; Trier and Taxt 1995 recommend a value of $l = 15$), the local population is assumed to be homogeneous, and is marked as background. Otherwise, the threshold is:

$$T(x, y) = (z_{high} + z_{low}) / 2, \quad (4.2)$$

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

The Niblack (Niblack 1986) binarization uses the threshold:

$$T(x, y) = m(x, y) + k \cdot s(x, y), \quad (4.3)$$

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 $k < 0$, it is guaranteed that $T(x, y) < m(x, y)$. 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 (White and Rohrer 1983).

The Sauvola method (Sauvola and Pietikainen 2000) 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:

$$T(x, y) = m(x, y) \cdot \left[1 + k \cdot \left(\frac{s(x, y)}{R} - 1 \right) \right], \quad (4.4)$$

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 (Gatos et al. 2004), 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.

Some additional binarization algorithms in historic documents setting (e.g. Bar-Yosef et al. 2007; Ben Messaoud et al. 2012) were also examined, yet found to be unsuitable for our needs (e.g. found to be relatively “tailor-made” for specific domains).

Binarization Results for Existing Algorithms

The experiments presented below were performed on three images of different ostraca, Lachish ostracon No. 3 (verso; Torczyner et al. 1938), Arad ostracon No. 1 (Aharoni 1981; both Lachish No. 3 and Arad No. 1 contain ancient Hebrew writing), and Arad ostracon No. 34 (Aharoni 1981; 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. 4.1, 4.2 and 4.3.

The experiments demonstrate 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 sub-section, an alternative binarization scheme, taking into account information from the facsimile itself, will be presented.

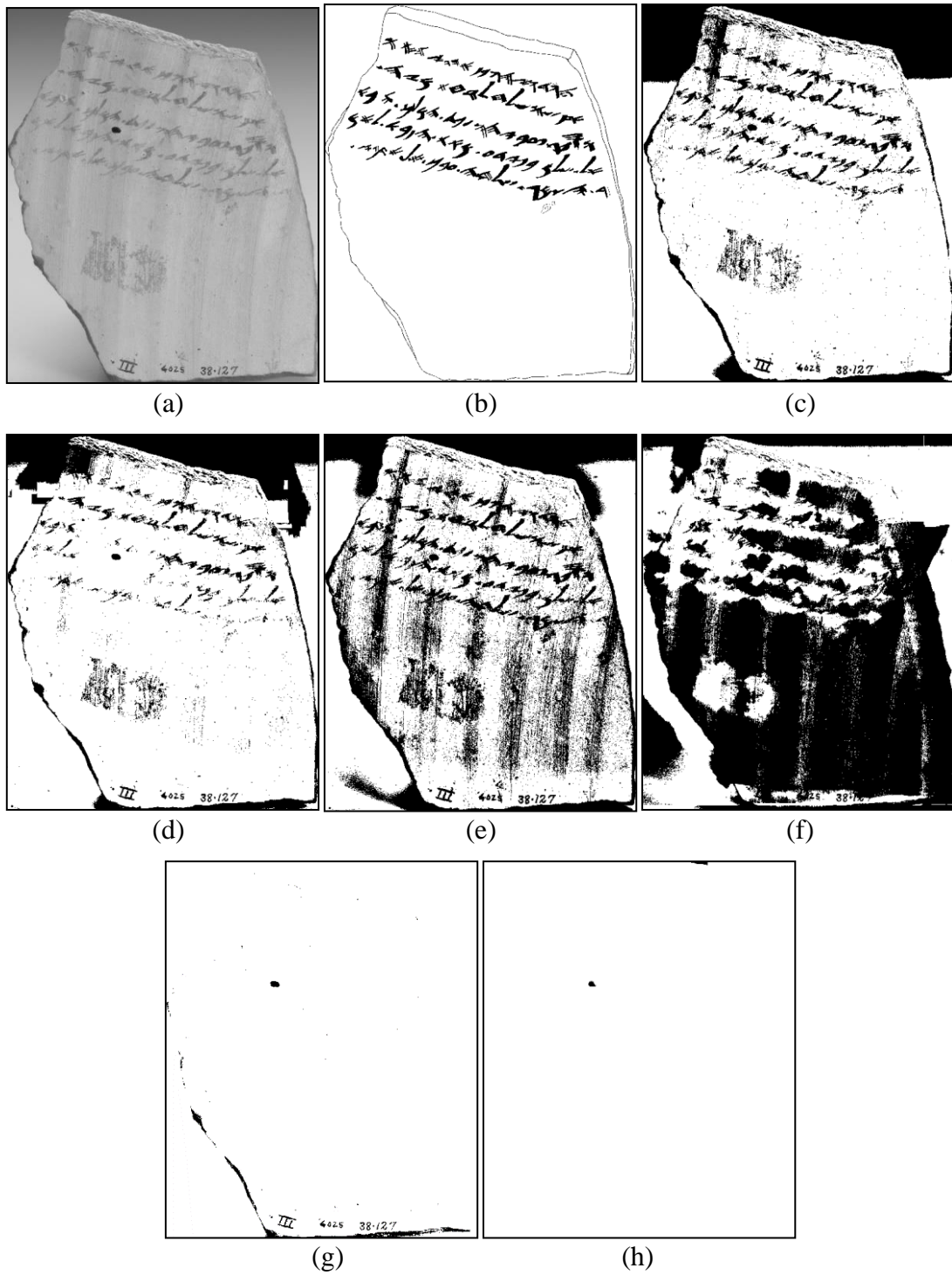


Figure 4.1 Lachish No. 3 experiment: (a) ostracon image; (b) manual facsimile. Results of: (c) Otsu; (d) Bernsen; (e) Niblack; (f) White; (g) Sauvola; (h) Gatos.

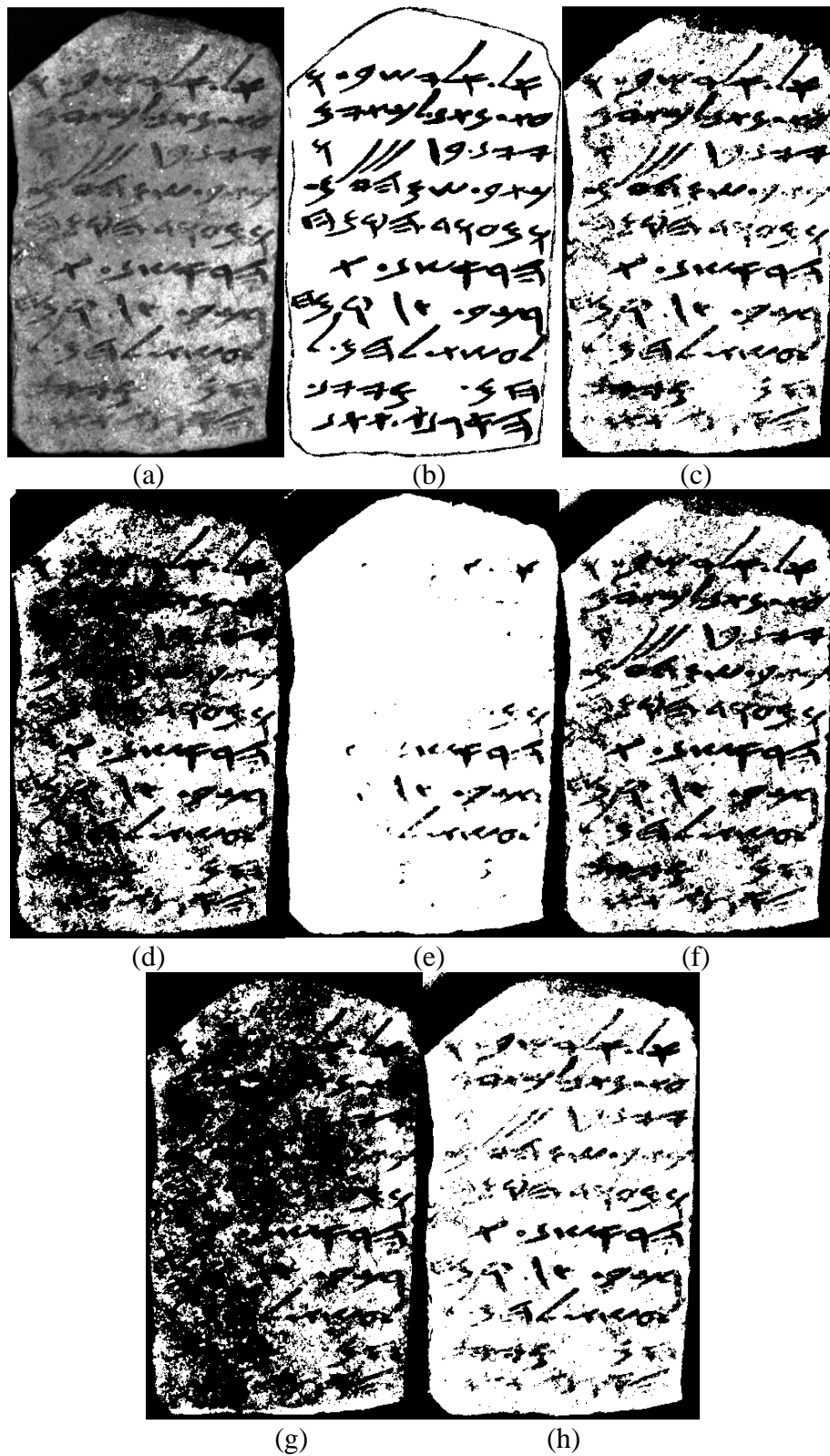


Figure 4.2 Arad No. 1 experiment: (a) ostracon image; (b) manual facsimile. Results of: (c) Otsu; (d) Bernsen; (e) Niblack; (f) White; (g) Sauvola; (h) Gatos.

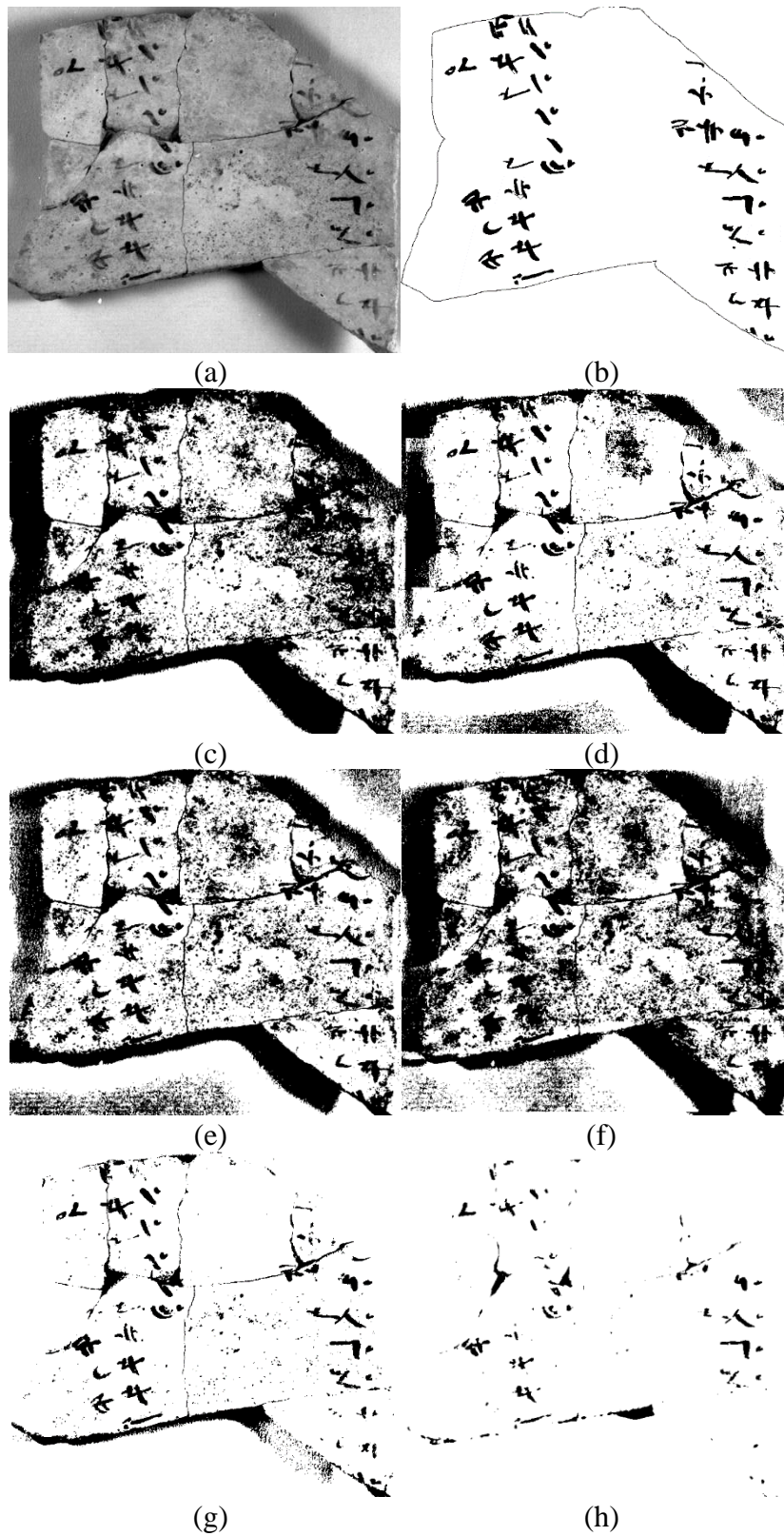


Figure 4.3 Arad No. 34 experiment: (a) ostracon image; (b) manual facsimile. Results of: (c) Otsu; (d) Bernsen; (e) Niblack; (f) White; (g) Sauvola; (h) Gatos.

4.3 Proposed Algorithm's 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 Zitová and Flusser 2003). 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 ostraca and the facsimile images (Aharoni 1981).

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 ostraca image dimensions. The target function for this, and all the subsequent stages, is:

$$CMI(F, O) = \mu_C - \mu_I, \quad (4.5)$$

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). For additional details regarding this measure and its properties, see previous Sections 2 and 3. Fig. 4.4

illustrates the results of the registration on superimposed facsimile and ostraca 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.

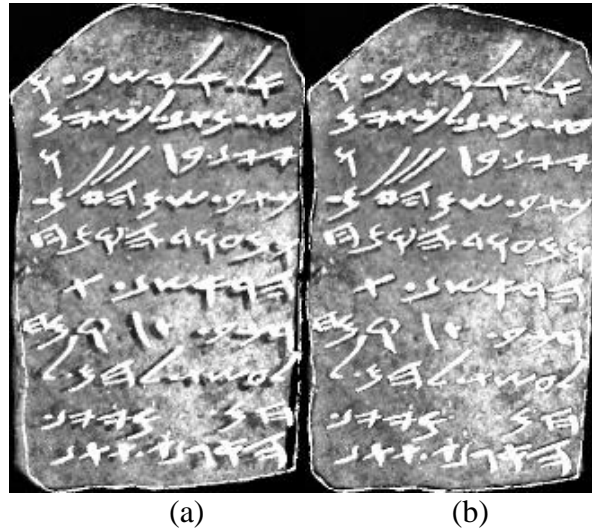


Figure 4.4 Example of ostracum-facsimile correspondence before (a) and after (b) the registration.

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. 4.5. 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 4.5 An example of ostracon-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 in (b), 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. 4.6 illustrates the CC's movements before and after the application of median filter and re-registration. Fig. 4.7 shows the improvement in the ostraca-facsimile correspondence.

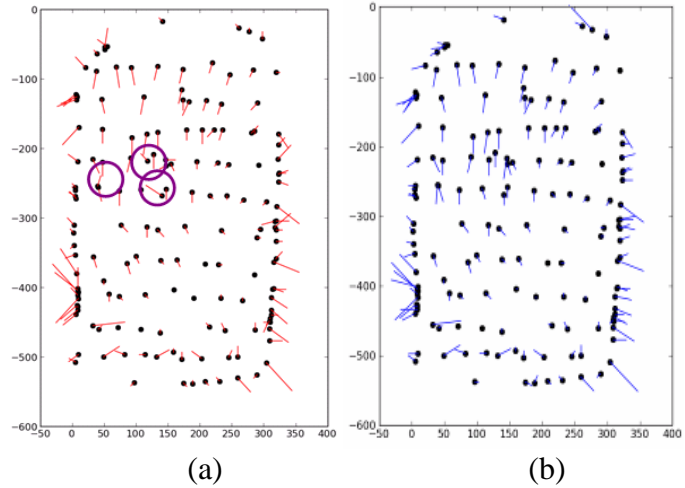


Figure 4.6 An example of per-CC movement (in pixels) before (a) and after (b) median filter and re-registration. Note the disappearance of the old misalignments, marked by violet color in (a).

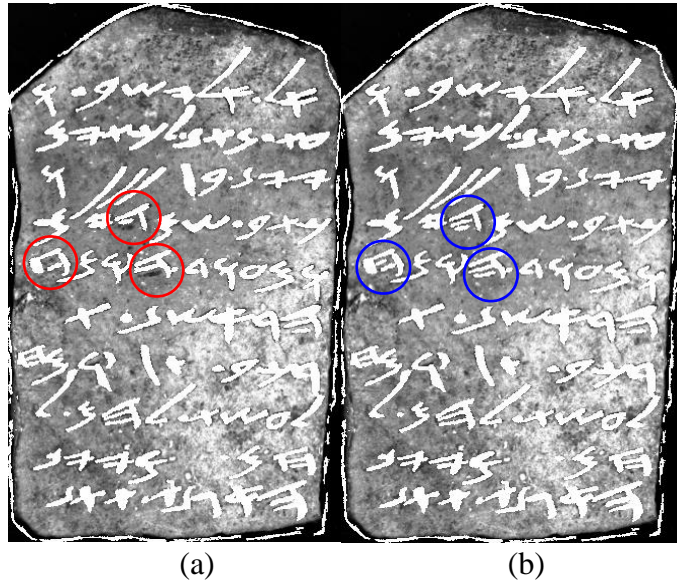


Figure 4.7 An example of improvement between the second (a) and third (b) registration stages. Note the reappearance of the missing CC's.

4) Proportional Binarization

The last stage utilizes the current registration in order to achieve a satisfying binarization of the ostraca 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 ostracon 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 ostracon image. The classical algorithms mentioned in previous sub-sections, 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. Therefore, some of the methods will either continue to be stuck in sub-optimal maxima, or will have to be adapted by ad-hoc modification of their parameters' tuning.

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 ($RF(p)$):

$$IP_j = \frac{\#\{p \mid p \in BO_j \wedge RF(p) = 0\}}{\#\{p \mid p \in BO_j\}}. \quad (4.6)$$

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

$$\frac{\#\{p \mid p \in BO_j \wedge O(p) < T_j\}}{\#\{p \mid p \in BO_j\}} \cong IP_j. \quad (4.7)$$

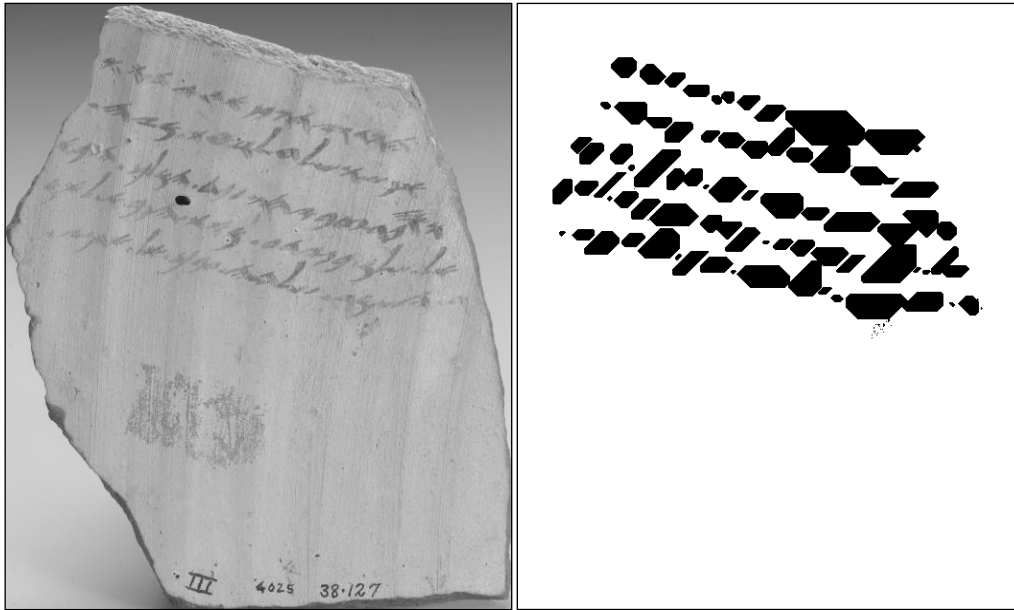
Finally, every BO_j within $O(p)$ is thresholded according to the T_j in Eq. 4.7. In addition, small denoising procedures (e.g. morphological operations) can be performed, either within each BO, or on a global scale. In what follows, we present

results without denoising, as well as results with simple stain (CC below certain size) removal.

4.4 Proposed Algorithm's Results

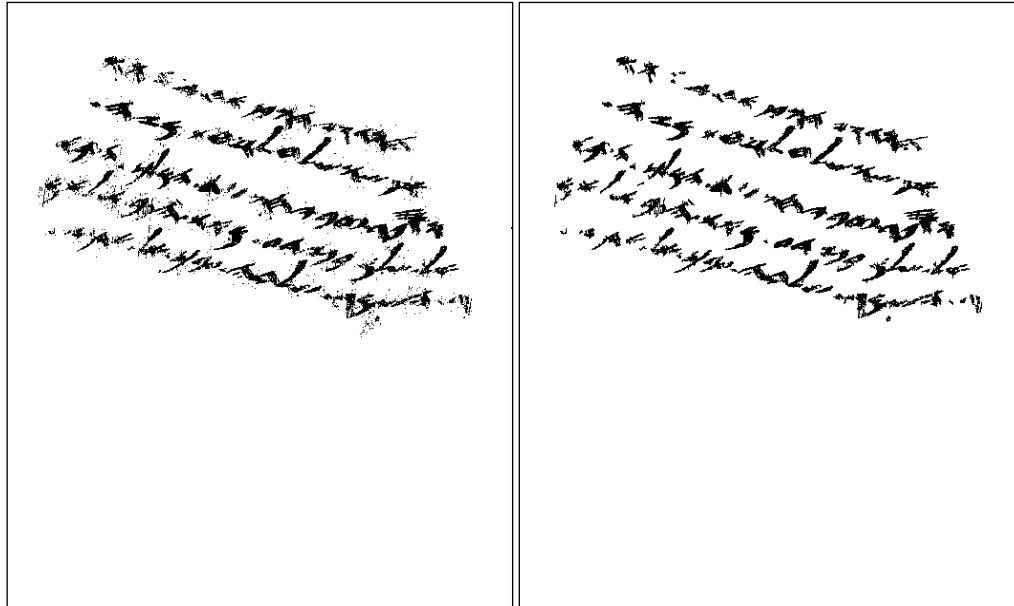
Following the previous experimental setting, the ostraca and facsimile images of Lachish No. 3, Arad No. 1 and Arad No. 34 were analyzed. The results of the new registration and binarization algorithm can be seen respectively in Figs. 4.8, 4.9 and 4.10. In all cases, the ostraca border pixels were removed.

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.



(a)

(b)



(c)

(d)

Figure 4.8 Lachish No. 3: (a) ostracon image; (b) bounding octagons; (c) binarization result; (d) binarization result with stain removal.

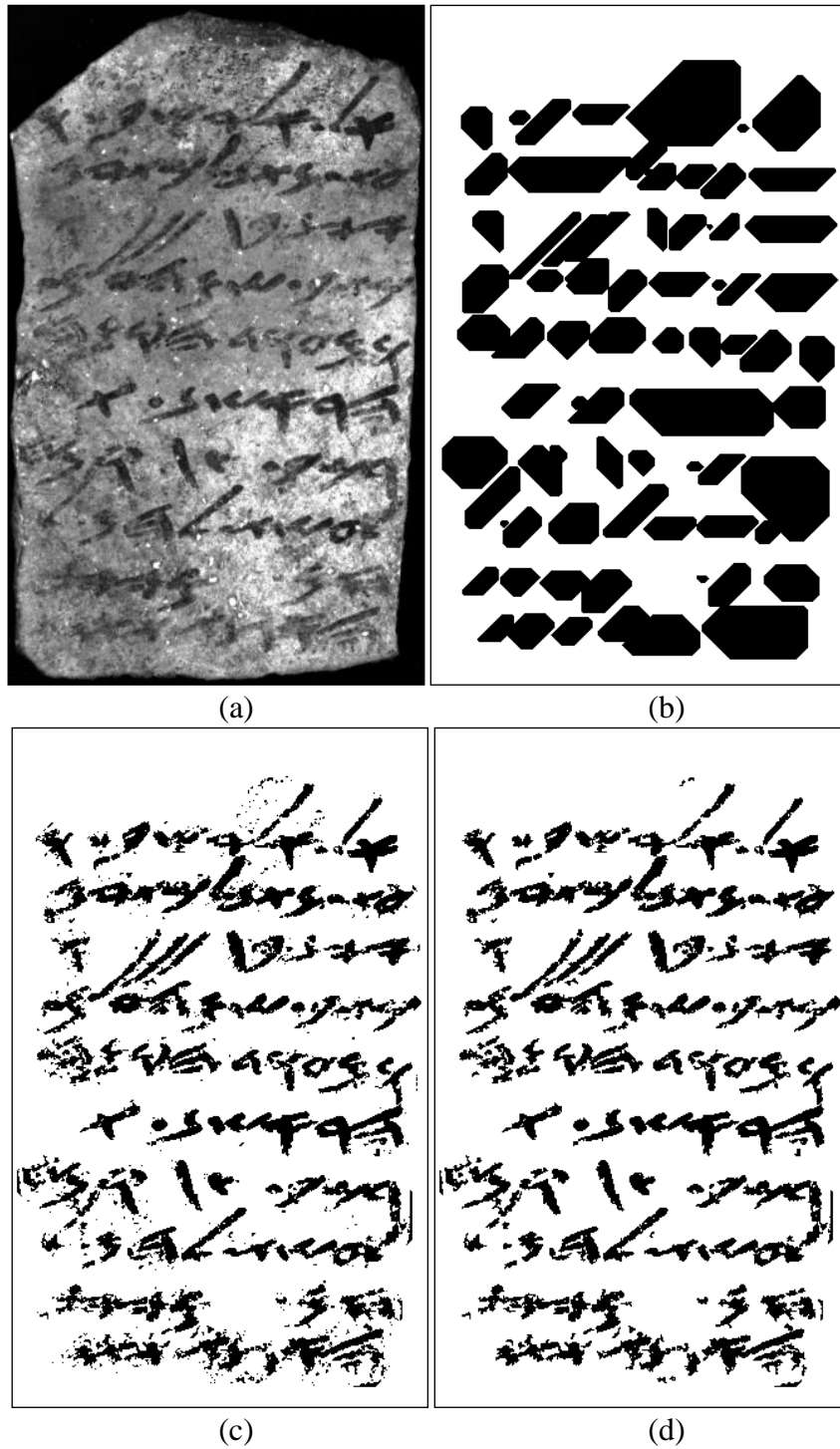


Figure 4.9 Arad No. 1: (a) ostracon image; (b) bounding octagons; (c) binarization result; (d) binarization result with stain removal.

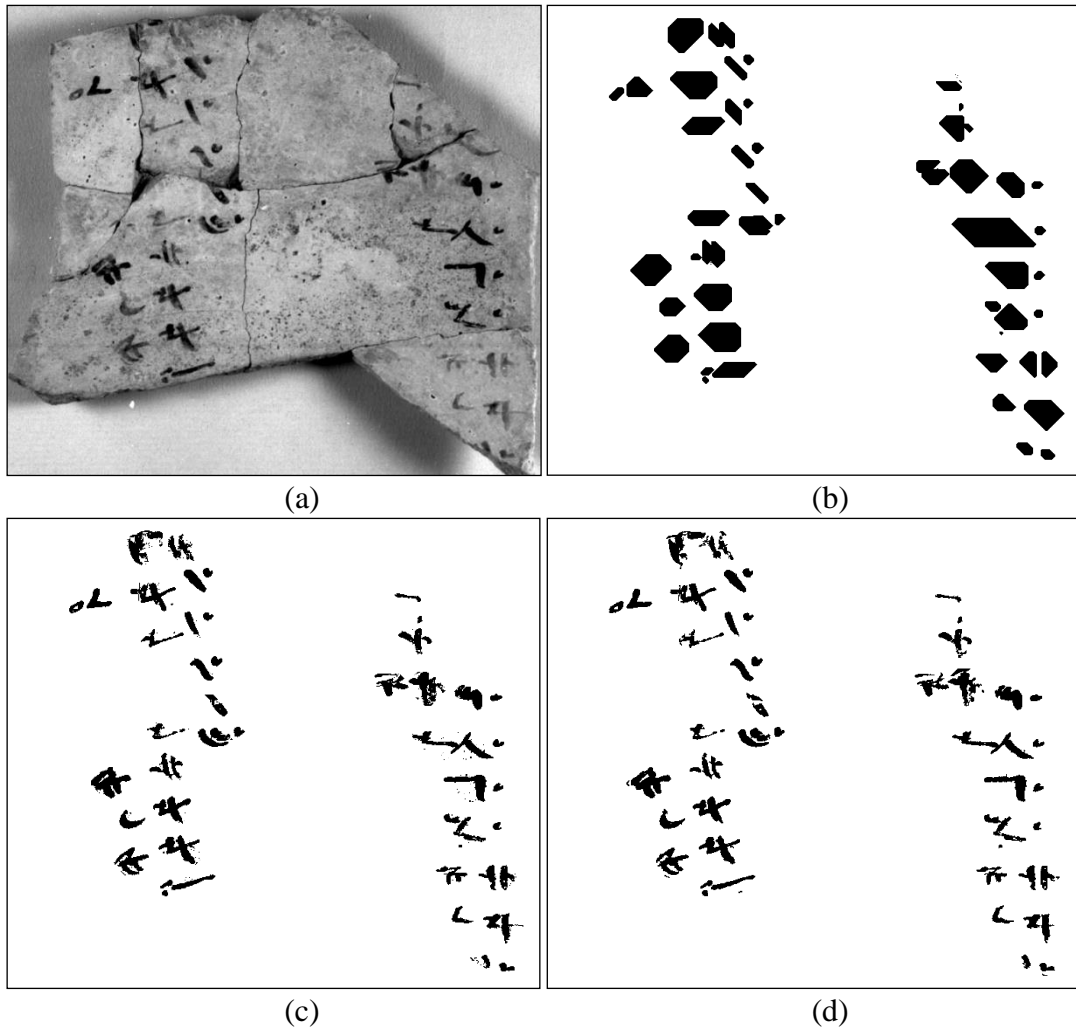


Figure 4.10 Arad No. 34: (a) ostracon image; (b) bounding octagons; (c) binarization result; (d) binarization result with stain removal.

4.5 Summary

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 (containing an approximate depiction of all the characters) to an ostracon image, was also tested, with superior results. It can therefore be concluded that the proposed method is sound. This technique will be further improved in Section 5.

5. Binarization Improvement via Sparse Dictionary Model

5.1 Problem Statement

In the previous section, a binarization algorithm for diverse types of ostraca was presented and tested. The resulting binarizations are of superior quality comparing to several other prominent algorithms. Nevertheless, in our view, this quality can still be improved via modern dictionary-based denoising method. Our approach was first demonstrated in (Shaus et al. 2013).

In the last decade, there has been a rapid development in the field of sparse coding methods, for various Image Processing tasks. We explore the possibility of using similar techniques in order to produce an improved inscription binarization.

Let $D \in \mathbb{R}^{n \times K}$ be an over-complete dictionary that contains K atoms $\{d_j\}_{j=1}^K$ for columns. A signal $y \in \mathbb{R}^n$ can be represented or approximated by a sparse linear combination of these atoms, $y \approx Dx$, $x \in \mathbb{R}^K$. The approximation is chosen in the sense that $\|y - Dx\|_p \leq \varepsilon$ (p is commonly selected to be 1, 2 or ∞), with the sparsity of x minimized by the l_0 norm, counting the number of nonzero coefficients. In other words:

$$\min_x \|x\|_0 \quad s.t. \quad \|y - Dx\|_p \leq \varepsilon . \quad (5.1)$$

Since an exact determination of the sparsest representation is an NP-hard problem (Davis et al. 1997), there arises a need for reducing the size of the dictionary D . This can either be performed prior to solving the minimization problem in Eq. 5.1, or in parallel. The most prominent methods for dictionary training and quantization are k-means (Gersho and Gray 1991), ML - Maximum Likelihood methods (Olshausen and Field 1996; Lewicki and Olshausen 1999), MOD - Method of Optimal Directions (Engan et al. 1999), MAP - Maximum A-Posteriori Probability (Kreutz-Delgado and

Rao 2000), UONB - Union of Orthonormal Bases (Lesage et al. 2000) and the highly popular k-SVD (Aharon et al. 2006).

In order to produce a binarization, we would like to use a dictionary containing black and white patches. In addition, the representation should avoid any combinations of such atoms in order to maintain the binary property of the resulting approximation. Only one atom d_j from the dictionary, with a corresponding coefficient of $x_j = 1$ should be used for each approximated patch, while $\forall i \neq j \quad x_i = 0$. Therefore, the problem is slightly changed: we set the l_0 norm of x exactly to 1, while we wish to find the best approximation of y . Though it is tempting to approximate the inscription image by itself (using it as a source for y), empirically its imperfect binarized images perform much better in this role.

Thus, using the above-mentioned formalism, our problem is composed of two steps:

1. Learn an over-complete **binary** dictionary $D \in \{0, 255\}^{n \times K}$, representing black and white patches.
2. For each patch y in an existing **imperfect binarization**, find $\min_x \|y - Dx\|_p$ subject to $\|x\|_0 = 1$.

5.2 Proposed Solution

The most demanding task is the first step of dictionary learning. It would seem that after obtaining a large database of patches, one would be able to plug-in any of the above mentioned off-the-shelf solutions. However, this is not the case. Almost all of

these methods can be essentially interpreted as generalizations of the k-means algorithm. Thus, the constructed dictionary would almost certainly result in gray level, rather than black and white values. Moreover, most of the methods assume that d_j are part of a linear space (possibly coupled with an inner product), which is untrue in our case.

One possible solution may be applying “extensive” methods, avoiding the need for quantization altogether and using a large patches database as a dictionary. This may not always be feasible. A more elegant suggestion would be the utilization of the k-medians (Jain and Dubes 1981) or k-medoids (Kaufman and Rousseeuw 1987) methods, which results in K atoms with appropriate values of 0 or 255 (the uncommon case of 127.5 can be assigned to either one of them). In what follows, the k-medians and k-medoids algorithms were implemented via the Pycluster toolkit (De Hoon et al. 2004).

The formal description of the algorithm, for each inscription, is:

1. Collect a large database of clean black and white patches. We use the above-mentioned hand-made facsimiles as the primary source.
2. Learn a dictionary based on the database, using either k-medians or k-medoids method. Alternatively (if allowed computationally), use the whole database as a dictionary.
3. For each patch y in the existing imperfect binarization, find the most suitable replacement $d_j \in D$, chosen by the solution of $\min_{\|x\|_0=1} \|y - Dx\|_p$. If the patches y overlap, construct the binarization by prioritizing the patches with a better score.

A collection of patches, illustrating stage1 and 2, can be seen at Fig. 5.1. Illustration of stage 3, with reconstructed images' fragments, can be seen on Figs. 5.2 and 5.3, demonstrating respectively the results for Arad No. 1 and Arad No. 34 ostraca (Aharoni 1981).



Figure 5.1 A collection of patches, illustrating stages 1 and 2 of the algorithm.

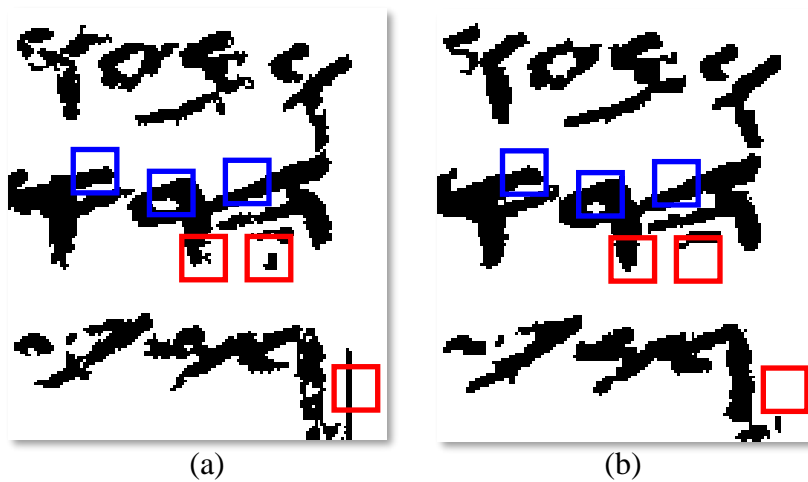


Figure 5.2 Fragment of Arad #1: (a) binarization from Section 4 – in **blue** good patches reflecting the writing practice, in **red** non-representative “noisy” patches; (b) binarization improvement, with representative patches maintained with minimal changes, while non-representative patches replaced.



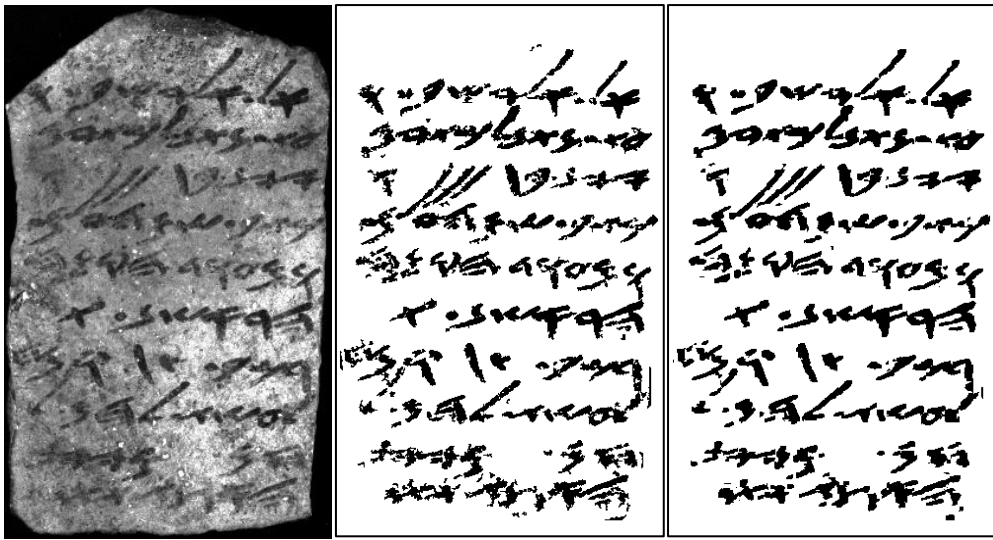
Figure 5.3 Fragment of Arad No. 34: (a) binarization from Section 4, in **red** non-representative “noisy” patches; (b) binarization improvement, with non-representative patches replaced.

5.3 Experimental Results

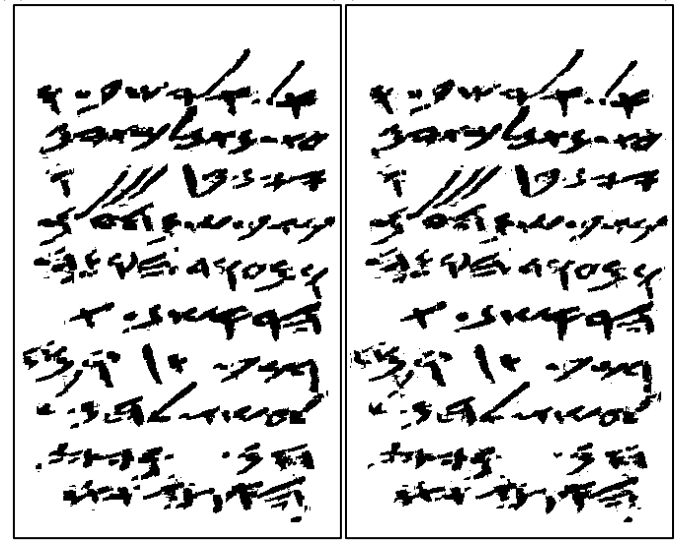
Our experiments tested the soundness and performance of the technique with respect to different algorithm parameters and various ostraca inscriptions. The following parameters were kept constant: patch size = 11×11 pixels (in the initial database, patches are sampled on 3 pixels' grid, with at most 73% overlap), dictionary size = 100 atoms (except for the extensive dictionary solution, where a typical number of atoms was 1000 up to 30000), number of repeated random initializations for k-medians and k-medoids = 100.

The first experiment tested the relationship between the best binary images available for our medium (see previous Section), and the improved binarization obtained by k-medians, k-medoids and extensive dictionary methods. The results for the ostrakon of Arad No. 1 (Aharoni 1981) can be seen in Fig. 5.4, while the results for the ostrakon of Arad No. 34 (Aharoni 1981) can be seen on Fig. 5.5.

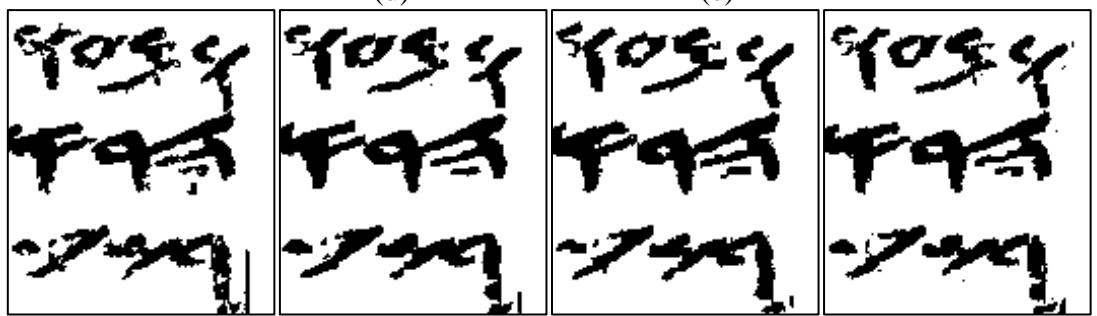
The results show that the performance of the sparse models rivals that of the best binarization. In fact, when looking on fine-grained details like strokes continuity, deviations from straight line, edge noise and the presence of stains, k-medians and k-medoids outcomes are superior to the available binarization, though not by a far margin. We note that despite its heavy computational burden, the extensive dictionary solution does not surpass the k-medians and k-medoids in both cases. It may be that the optimally fitting patches of the extensive dictionary result lack the robustness of the k-medians and the k-medoids solutions.



(a) (b) (c)



(d) (e)



(f) (g) (h) (i)

Figure 5.4 Arad No. 1: (a) ostracon image; (b) binarization from Section 4; (c) k-medians result; (d) k-medoids result; (e) extensive dictionary result. Zoom on right-center: (f) binarization from Section 4; (g) k-medians result; (h) k-medoids result; (i) extensive dictionary result.

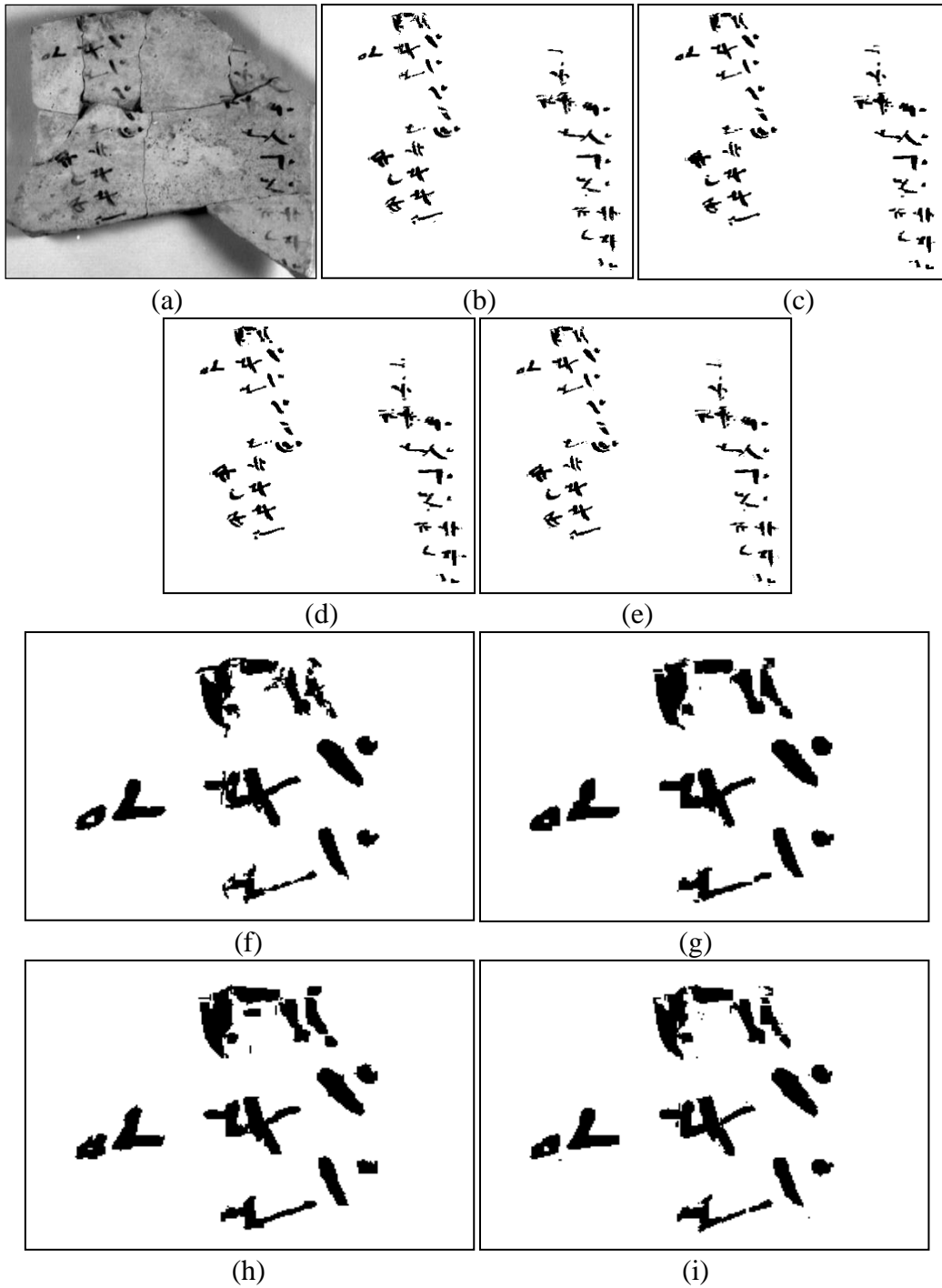


Figure 5.5 Arad No. 34: (a) ostracon image; (b) binarization from Section 4; (c) k-medians result; (d) k-medoids result; (e) extensive dictionary result. Zoom on top-left: (f) binarization from Section 4; (g) k-medians result; (h) k-medoids result; (i) extensive dictionary result.

The robustness of the different methods was put to a test in the second experiment. The initial patches database was reduced by a factor of 3 by removing duplicate patches (in a non-robust scenario, this may bias the selection of the dictionary

atoms). It was then further reduced by 9 and by 25 by changing the sampling ratio. The results of this test can be seen on Fig. 5.6.



Figure 5.6 Arad No. 1: experiment testing the robustness of k-medians, initial DB size reduced by a factor of: (a) 3 (b) 21 (c) 75; experiment testing the robustness of k-medoids, initial DB size reduced by a factor of: (d) 3 (e) 21 (f) 75.

The results demonstrate that the k-medians algorithm has an impressively robust behaviour, even under relatively strenuous initial database shrinkage. On the other hand, the performance of k-medoids is less robust and hard to predict. It may be that the medoids are prone to be altered upon changes in database (since medoids are database members) or in the random initialization.

5.4 Summary

We presented a method to improve an already existing unsatisfactory binarization utilizing a sparse model. A database of black and white patches was created from a clean source. Existing dictionary learning methods were found to be unsuitable for our needs. Therefore, a dictionary was created via k-medians, k-medoids and extensive dictionary techniques. The results of k-medians and k-medoids were found to be sound, with fine-grained details superior to the available binarization, though not by a far margin. Further tests revealed that k-medians algorithm is more robust to initial database shrinkage than k-medoids.

6. Quality Evaluation of Binarizations

6.1 Introduction

The established methodology of document binarization assessment relies upon ground truth (GT) images (see binarization competitions results in Gatos et al. 2009; Pratikakis et al. 2010, 2011, 2012, 2013; Ntirogiannis et al. 2014). This is motivated by the need for binarization quality criteria. A manually created GT image is presumed to be a close approximation to the binarization ideal. Consequently, the different binarized images are scored according to their adherence to the GT image. The entire evaluation process, depicted in Fig. 6.1, consists of the following stages:

- Preliminary step: A black and white **GT** is created manually, based upon a gray-scale **document image**. This process is driven by human-operated tools (e.g. Ntirogiannis et al. 2008; Saund et al. 2009; Fischer et al. 2010; Clausner et al. 2011; Biller et al. 2013).
- Algorithms application: The same **document image** serves as an input for the various binarization algorithms, resulting in **binary images** (herein: binarizations).
- Algorithms evaluation: These **binarizations** are judged against the **GT**, using quality assessment metrics (such as F-measure, pseudo F-measure, PSNR, Negative Rate Metric, Distance Reciprocal Distortion Metric and Misclassification Penalty Metric; see Gatos et al. 2009; Pratikakis et al. 2010, 2011, 2012, 2013 and Ntirogiannis et al. 2014 for details).

Due to certain drawbacks in this methodology (detailed below), we present two alternative solutions. The first suggestion is an **evaluation of the binarizations directly versus the document image**, avoiding the use of GT altogether. The second option is strengthening the existing methodology by **assessing the GT quality** prior to

its usage. Both solutions rely on an identical mechanism and we therefore consider them together.

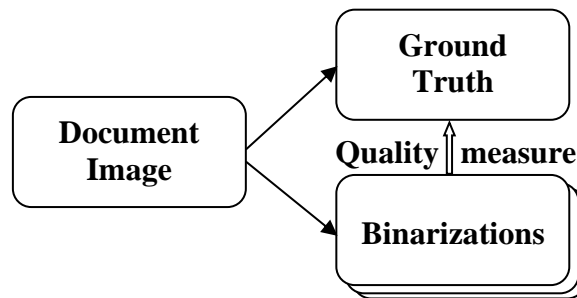


Figure 6.1 Standard binarization quality evaluation process. The document image is gray-scale, while the binarization and the ground truth are black and white images. The quality metric measures the adherence of the binarization to the ground truth.

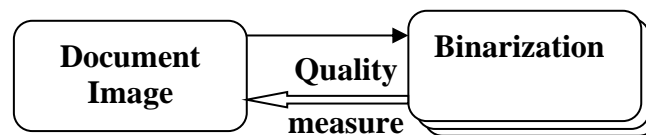


Figure 6.2 Proposed binarization quality evaluation process. The quality of binarization or ground truth is assessed by measuring their adherence to the document image.

The main contribution of the current section is the suggestion of several new measures, enabling the **assessment of the accuracy of black and white depictions of a document** (binarizations or GT) **directly vs. the document image itself** (see the proposed framework in Fig. 6.2). The original presentation was made in (Shaus et al. 2016b).

6.2 Methodological Pitfalls

Several papers deal with the deficiencies of the existing methodology. All of them emphasize the subjectivity and the inherent inconsistency of the GT creation process.

In (Barney Smith 2010), the variability of five binarization algorithms was compared to that of different manual GTs. Significant irregularities in the GTs of the same document were found. Surprisingly, the results revealed that the variance between the binarizations was smaller than the variance between the GTs, created by diverse human operators.

The research presented in (Shaus et al. 2012a; see Section 2) deals with GTs of First Temple period Hebrew inscriptions, created by several experts. Their GTs were shown to be of markedly different quality.

The study in (Barney Smith and An 2012) performed a binarization classifier training, based on three variants of GT. The performance of the classifiers varied significantly with respect to the underlying GT.

Therefore, existing evidence demonstrates that the GT is inherently subjective, with large deviations between different human operators and creation techniques, influencing the performance of the algorithms “downstream”. This problem was noted already in (Brown et al. 1988), where automatic systems were found to be more reliable than the human “ground truther”.

6.3 Existing Solutions

The aforementioned methodological pitfalls were addressed by some articles in the past. This sub-section provides a brief survey of these proposed solutions which are found to be inadequate in certain scenarios.

The research in (Ntirogiannis et al. 2008) aims at presenting an objective evaluation methodology for document image binarization, performed in the following fashion:

- Preliminary steps: A skeleton of GT is created via the algorithms (Kamel and Zhao 1993; Lee and Chen 1992), and **corrected manually**. The document image edges are extracted by the Canny's method (Canny 1986).
- Algorithms evaluation: The GT skeleton is dilated **within each binarization**, until 50% of the edges inside each connected component are covered. This results in a new, "evaluated GT".

This approach has several shortcomings. First, it includes a manual stage. According to our tests, the impact of this stage is not negligible. Second, the method constructs a different "evaluated GT" for each binarization. Therefore, every binarization is judged against its own GT, with no common ground for comparison. Finally, no justification is given for preferring the proposed intricate scheme to the current methodology. The similarity of the outcomes in (Ntirogiannis et al. 2008), as well as Occam's razor principle, suggest that the existing simpler methodology should be favored. A later article (Ntirogiannis et al. 2012) made attempts to improve upon (Ntirogiannis et al. 2008), yet hasn't avoided the manually performed stages (e.g. "The user shall verify that at least one dilation marker exists within the borders of each ground truth component"; "the user shall close any edge disconnections", etc.).

Another approach presented in (Ben Messaoud et al. 2011) is an elaboration on the same theme. The main changes are dropping the manual correction phase, and dilating with respect to binarizations created by methods (Sauvola and Pietikainen 2000; Gatos et al. 2006; Lu et al. 2010). This circumvents a creation of different GT for

each binarization and the potential for human error. However, this approach merely creates another, albeit sophisticated, binarization procedure. Though this is certainly an “objective” way to handle the binarization evaluation, in fact it pre-supposes that the presented procedure creates the perfect binarization for all scenarios, which is not proved by the authors.

A different proposal, specified in (Stathis et al. 2008b; Paredes and Kavallieratou 2010), is to create an algorithms’ evaluation strategy evading the manual GT creation step. A clean, binary image of a document is marked as GT. This image is combined with any desired type of noise, in order to create a **synthetic document image**. The evaluated binarization algorithms are activated on the synthetic document image and are judged against the perfect GT. This elegant technique avoids the need for the creation of GT images. On the other hand, it cannot evaluate binarizations of already existing degraded documents. In addition, if no clean version of a given type of handwriting or typeface exists (e.g. in case of ancient inscriptions), or if the noise model cannot be adequately deduced, the method is also inapplicable.

Yet another, “goal-directed” approach (Trier and Jain 1995), also avoids ground-truthing. The results of different binarization techniques are used as inputs for other algorithms (e.g. OCR systems), whose outputs are the ones being evaluated. However, with any sufficiently complicated goal, the tuning of the parameters “downstream” may have a major influence on the outcomes. In certain cases (e.g. historical documents), the binarization may also be the desired end product, with no further processing required.

6.4 Preliminary Definitions and Assumptions

This section proposes **several new metrics** assessing either the binarization or the GT. A first step in that direction was undertaken in (Shaus et al. 2012a; see Section 2), where different GTs of the same historical inscription were compared. The technique superimposed the GTs over the document image. The quality of the fit was used in order to rank the different GTs. In similar fashion, other metrics can be used in order to evaluate the quality of either the binarizations themselves (bypassing the GT), or the accompanying GT (therefore, adding a verification step to the existing scheme).

In what follows, we assume:

1. A **black and white image** $BW(x, y)$ ($BW : [1, M] \times [1, N] \rightarrow \{0, 255\}$) which can be either a **binarization** or a **GT**, is superimposed over a gray-scale **document image** $D(x, y)$ of the same dimensions (if needed, a preliminary registration is performed, see Section 2).
2. A measure m , taking into account certain correspondences between BW and D , is used in order to evaluate the quality of BW .

In the considered situation, the correspondence between the BW and D images defines the **foreground** and **background** sets of pixels: $F = \{(x, y) \mid BW(x, y) = 0\}$ and $B = \{(x, y) \mid BW(x, y) = 255\}$, respectively (with $\#F + \#B = MN$). The measure m may take into account the properties of these two populations **within** D .

We use the following notations:

- μ_F and μ_B are the foreground and background mean values within the D image,

$$\text{i.e. } \mu_S = \left(\sum_{(x,y) \in S} D(x, y) \right) / \#S \text{ for } S = F, B$$

- σ_F and σ_B are their standard deviations, defined in a similar fashion.
- $n_F = \frac{\#F}{\#F + \#B}$ and $n_B = \frac{\#B}{\#F + \#B}$ are respectively the proportions of the foreground and the background pixels.
- $f_i = \frac{\#\{(x, y) \in F \mid D(x, y) = i\}}{\#F}$ and $b_i = \frac{\#\{(x, y) \in B \mid D(x, y) = i\}}{\#B}$, $i = 0 \dots 255$, are the empirical distributions (histograms) of foreground and background pixels within D .

6.5 Proposed Measures

We consider the following measures:

- **Adapted Otsu:** The article (Otsu 1979) used a thresholding criterion minimizing the intra-class variance for background-foreground separation. A similar measure can be used in order to assess the intra-class variance, dropping the requirement of hard-thresholding. Thus:

$$m_{Otsu} = n_F \cdot \sigma_F^2 + n_B \cdot \sigma_B^2. \quad (6.1)$$

It is assumed that smaller values of m_{Otsu} reflect better quality of BW .

- **Adapted Kapur:** The paper (Kapur et al. 1985) used an entropy-based thresholding criterion for binarization, maximizing the sum of entropies of background and foreground populations. Again, dropping the requirement for a threshold, we obtain:

$$m_{Kapur} = \sum_{i=0}^{255} f_i \log(f_i) + \sum_{j=0}^{255} b_j \log(b_j), \quad (6.2)$$

with $x \log(x)$ considered zero at $x=0$. Our expectation is that larger values of m_{Kapur} indicate a better *BW*.

- **Adapted Kittler-Illingworth (KI):** The study (Kittler and Illingworth 1986) presumed a normally distributed foreground and background pixel populations. The derived criterion function tries to reduce the classification error rate under this supposition. Again, we shall use a similar measure, with no hard-thresholding:

$$m_{KI} = 1 + 2 \cdot [n_B \log(\sigma_B) + n_F \log(\sigma_F)] - 2 \cdot [n_B \log(n_B) + n_F \log(n_F)]. \quad (6.3)$$

Our expectation is that smaller m_{KI} values reflect better *BW*.

- **CMI:** The measure deals with the quality assessment-related tasks in historical inscriptions settings. It was defined and employed in Sections 2 and 4. As such, this is not an adapted method, but a measure developed directly in order to handle similar issues. It would be reminded, that the measure was defined as:

$$m_{CMI} = \mu_B - \mu_F. \quad (6.4)$$

- **Potential Contrast (PC):** This concept was presented in Section 3, for the purpose of assessment of multispectral images. The rationale behind this measure is an optimization of m_{CMI} under all possible gray-level transformations of the **document** image. It can be shown that this is achieved by:

$$m_{PC} = 255 \cdot \sum_{i: f_i \leq b_i} (b_i - f_i). \quad (6.5)$$

As in the case of m_{CMI} , it is assumed that better *BW* is indicated by larger m_{PC} .

Remark: As seen above, different approaches prefer either small or large measure values. For the sake of consistency, in the experimental sub-section (below)

we **negate** the Otsu and the KI measures. Thus, it is assumed that the better BW always corresponds to a higher value of a given measure.

Additional “classical” measures for image (or matrix in stacked column vector format) comparison can be also utilized for our purpose, in particular L_1 , L_2 and PSNR measures.

- **L_1** : Defined by:

$$m_{L_1} = \sum_{(x,y)} |D(x, y) - BW(x, y)|. \quad (6.6)$$

- **L_2** : Defined by:

$$m_{L_2} = \sqrt{\sum_{(x,y)} (D(x, y) - BW(x, y))^2}. \quad (6.7)$$

Again, consistency-wise, these two measures ought to be negated.

- **PSNR**: Defined by:

$$m_{PSNR} = 10 \cdot \log_{10} \left(255^2 / \left(\frac{m_{L_2}^2}{MN} \right) \right). \quad (6.8)$$

Definition of Measures' Equivalence

Two given measures m_1 and m_2 are denoted as equivalent, $m_1 \sim m_2$, if for a constant D and different BW and BW^* the monotonicity is maintained jointly, i.e.:

$$m_1(BW, D) > m_1(BW^*, D) \Leftrightarrow m_2(BW, D) > m_2(BW^*, D). \quad (6.9)$$

Proposition I (Equivalence of PSNR and $-L_2$):

The PSNR measure is equivalent to the negated L_2 , i.e. $m_{PSNR} \sim -m_{L_2}$.

Proof:

Indeed, due to strictly increasing monotonicity of $C \cdot x$ ($0 < C \in \mathbb{R}$), $\log_{10}(x)$,

$-1/x$ and x^2 (for $x \geq 0$):

$$m_{PSNR} = 10 \cdot \log_{10} \left(\frac{255^2 MN}{m_{L_2}^2} \right) \sim \frac{255^2 MN}{m_{L_2}^2} \sim -m_{L_2}^{-2} \sim -m_{L_2}. \quad (6.10)$$

■

Proposition II (Equivalence of L_1 and L_2):

If $BW(x, y) \in \{0, 255\}$ (like in our setting), then $m_{L_1} \sim m_{L_2}$.

Proof:

The norms are influenced by the foreground and the background populations, induced by BW . Indeed, on the one side:

$$m_{L_1} = \sum_{(x,y)} |D(x, y) - BW(x, y)| = \sum_{(x,y) \in F} D(x, y) + \sum_{(x,y) \in B} (255 - D(x, y)) \quad (6.11)$$

Subtracting a constant (sum over the unvarying $D(x, y)$) would result in equivalent measure, therefore:

$$\sim \sum_F D(x, y) + \sum_B (255 - D(x, y)) - \sum_{(x,y)} D(x, y) = \sum_B (255 - 2 \cdot D(x, y)) \quad (6.12)$$

On the other side:

$$m_{L_2} = \sqrt{\sum_{(x,y)} (D(x, y) - BW(x, y))^2} \sim \sum_{(x,y)} (D(x, y) - BW(x, y))^2 \quad (6.13)$$

And moreover:

$$\begin{aligned}
&= \sum_{(x,y) \in F} D(x,y)^2 + \sum_{(x,y) \in B} (255 - D(x,y))^2 = & (6.14) \\
&= \sum_{(x,y) \in F \cup B} D(x,y)^2 + 255 \sum_B (255 - 2D(x,y))
\end{aligned}$$

Since the first term is constant, and as a multiplicative non-zero constant results in equivalent measure, we obtain:

$$\sim \sum_B (255 - 2 \cdot D(x,y)) \tag{6.15}$$

■

From Propositions I and II it follows that despite the seeming dissimilarity of the last three measures, they would in fact yield the same binarizations' ranking. Therefore, in subsequent sub-section, we would only use the m_{PSNR} measure.

6.6 Experimental Setting and Results

This section compares the performance of the six quality measures described above. We begin with the experimental settings, continuing with the results.

Experimental Setting

Goal: The goal of this experiment is to compare the performance of the measures under controlled deterioration of high-quality binarizations of various documents. We **require the measures to maintain a monotonic decrease with respect to the increasing worsening of the binarizations.** This may be seen as an “axiomatic” (and certainly reasonable) **requirement for the measures.** We stress that **in this experiment, the elements under examination are the different measures,** and not the binarizations.

Methodology: We tested the measures on purposely engineered binary images with gradually diminishing quality. For each document image, its corresponding high-quality binarization was used in order to obtain a sequence of progressively inferior black and white images. Three different types of deteriorations were pursued:

1. An addition of increasing levels of random **salt and pepper (S&P) noise** (1%, 2%, etc., stopping at 10%), imitating isolated artifacts of the binarization process (e.g. stains, see Sections 4, 5 for examples and methods for their handling). In order to ensure the significance of the results, each noise level was added independently 25 times (thus 25 different binary images were created with 1% noise, 25 more with 2% noise, etc.).
2. A continuing **morphological dilation of the foreground** (4-connectivity; dilations of 1 up to 10 pixels), emulating a binarization algorithm prone to False Positive errors near the edge (e.g. due to miscalculated threshold), or an operator with a preference for wide strokes creating the GT.
3. A continuing **morphological erosion of the foreground** (4-connectivity; erosions of 1 up to 3 pixels), mimicking a binarization algorithm prone to False Negative errors near the edge (e.g. due to miscalculated threshold), or an operator with a preference for narrow strokes creating the GT.

As already stated, our expectation was a constantly declining score, with the continuing deterioration of the engineered binarizations.

Datasets: Openly available data from several past binarization competitions were used, in particular DIBCO 2009 (Gatos et al. 2009; 5 handwritten and 5 printed documents), H-DIBCO 2010 (Pratikakis et al. 2010; 10 handwritten documents), DIBCO 2011 (Pratikakis et al. 2011; 8 handwritten and 8 printed documents), H-

DIBCO 2012 (Pratikakis et al. 2012; 14 handwritten documents), DIBCO 2013 (Pratikakis et al. 2013; 8 handwritten and 8 printed documents), and H-DIBCO 2014 (Ntirogiannis et al. 2014; 10 handwritten documents); a total of 76 documents. As the measures require a grayscale document image, in case RGB document images were provided, they were converted to grayscale by channel averaging.

Within the datasets, each document image was accompanied by its corresponding GT. The GTs were taken as a high-quality basis for our deterioration procedures, resulting in 2064 different binarizations tested.

Success criterion (for each image, each type of deterioration and each measure): **Monotonic decrease of the scores sequence** (e.g., maximal score for the original binary image, the next for 1% S&P noise, etc.). A non-observance of correct monotonic behavior between two consecutive deteriorated binarizations (e.g. the score increasing between 3% and 4% of S&P noise) was counted as a “**break of monotonicity**”.

Note: The abovementioned setting ensures the significance and the reproducibility of our results.

Experimental Results

Summaries of the results for distinct types of deterioration are presented in Tables 6.1, 6.2 and 6.3.

Table 6.1 presents the results of the S&P noising experiment. It can be seen that Otsu, KI, CMI and PC measures perform perfectly in this setting, with 0% ordering mistakes in all the sequences.

Table 6.1 Results for Salt and Pepper Deterioration

Dataset ^a	#Files	% of Breaks of Monotonicity					
		<i>Otsu</i>	<i>Kapur</i>	<i>KI</i>	<i>CMI</i>	<i>PC</i>	<i>PSNR</i>
DIBCO2009 H	5	0%	26%	0%	0%	0%	0%
DIBCO2009 P	5	0%	82%	0%	0%	0%	0%
H-DIBCO2010 H	10	0%	22%	0%	0%	0%	0%
DIBCO2011 H	8	0%	41%	0%	0%	0%	13%
DIBCO2011 P	8	0%	71%	0%	0%	0%	13%
H-DIBCO2012 H	14	0%	30%	0%	0%	0%	0%
DIBCO2013 H	8	0%	26%	0%	0%	0%	0%
DIBCO2013 P	8	0%	80%	0%	0%	0%	0%
H-DIBCO2014 H	10	0%	37%	0%	0%	0%	0%
Mean		0%	43.4%	0%	0%	0%	2.6%

a. H=Handwritten, P=Printed.

The PSNR measure also behaves nicely in most cases. Unfortunately, it shows 2.6% of monotonicity break. On in-depth inspection, these cases correlate with the existence of bright stripes across the document. In such cases, the PSNR (and consequently the equivalent L_1 and L_2 measures) might “prefer” a presence of foreground pixels mistaken for background, which may indeed happen in this type of noise.

Finally, the Kapur measure (with 43.4% mistakes) is unreliable in this experiment. Moreover, we do not consider this measure as well-founded, as it ignores the gray-level values altogether (a permutation of the histogram results in the same score).

Table 6.2 Results for Dilation of the Foreground

Dataset ^a	#Files	% of Breaks of Monotonicity					
		<i>Otsu</i>	<i>Kapur</i>	<i>KI</i>	<i>CMI</i>	<i>PC</i>	<i>PSNR</i>
DIBCO2009 H	5	24%	26%	4%	0%	0%	0%
DIBCO2009 P	5	0%	20%	2%	0%	0%	0%
H-DIBCO2010 H	10	0%	12%	6%	0%	0%	0%
DIBCO2011 H	8	0%	20%	1%	0%	0%	13%
DIBCO2011 P	8	0%	29%	0%	0%	0%	15%
H-DIBCO2012 H	14	0%	19%	6%	0%	0%	0%
DIBCO2013 H	8	0%	20%	3%	0%	0%	0%
DIBCO2013 P	8	0%	25%	0%	0%	0%	0%
H-DIBCO2014 H	10	0%	11%	4%	0%	0%	0%
Mean		1.6%	19.5%	3.2%	0%	0%	2.9%

a. H=Handwritten, P=Printed.

Table 6.2 shows the results of morphological dilation experiment. The CMI and PC measures still perform perfectly, with 0% mistakes. Otsu (1.6% breaks of monotonicity, all in a single dataset), PSNR (2.9% mistakes) and KI (3.2% mistakes) also exhibit satisfactory performance. A close examination shows that all the Otsu mistakes are attributed to the presence of dark stains, covering a large part of the document. In such a case, the Otsu metric may “prefer” a relocation of some B pixels to F , in order to reduce the variance σ_B^2 . As before, the Kapur metric does not show a reliable behavior.

Table 6.3 Results for Erosion of the Foreground

Dataset ^a	#Files	% of Breaks of Monotonicity					
		<i>Otsu</i>	<i>Kapur</i>	<i>KI</i>	<i>CMI</i>	<i>PC</i>	<i>PSNR</i>
DIBCO2009 H	5	0%	7%	20%	100%	60%	7%
DIBCO2009 P	5	0%	7%	0%	73%	20%	0%
H-DIBCO2010 H	10	0%	37%	0%	80%	47%	47%
DIBCO2011 H	8	0%	13%	21%	88%	71%	4%
DIBCO2011 P	8	0%	4%	0%	75%	46%	13%
H-DIBCO2012 H	14	0%	31%	7%	71%	50%	24%
DIBCO2013 H	8	4%	25%	0%	75%	46%	21%
DIBCO2013 P	8	0%	17%	21%	75%	46%	25%
H-DIBCO2014 H	10	0%	20%	0%	70%	37%	37%
Mean		0.4%	20%	7%	77%	47%	22%

a. H=Handwritten, P=Printed.

Table 6.3 documents a relatively small-scale morphological erosion experiment, limited to 3 erosions (as 4 erosion would result in a complete elimination of the foreground in some binary images). The almost perfectly performing Otsu measure is followed by KI, with 7% mistakes. Most of KI's mistakes were made on 1-pixel erosion stage, surely within the limits of the original GTs reliability. Kapur, PSNR, and particularly PC and CMI measures were confused by this setting. It is noticeable that the CMI and the PC measures do not take into account the information regarding the size of F and B . Subsequently, a preference for “thinning” the characters (limiting the foreground to only the most certain “skeleton” pixels, with only minor penalty to the background statistics) might be observed in these measures.

6.7 Summary

We presented several measures, which quantify the adherence of a binary image to its grayscale document image. The binary document can either be a product of a binarization algorithm, or a GT. Both cases are treated in the same fashion. In order to check the adequacy of the proposed measures, an experimental framework was constructed utilizing a clean binary document with specifically engineered increasing deterioration of the binarization.

The results indicate that the adapted Otsu and KI measures present the best overall performance for binarizations evaluation purposes. The PSNR, PC and CMI measures can probably be useful in scenarios with adequate stroke width. The adapted Kapur measure is not a viable option for a quality measure.

We note that it is not incidental that some of the measures mentioned in the current section are adaptations of global binarization techniques. Indeed, in our view,

assessing a binarization “looking back” at the document image can be considered as a dual problem to the task of arriving at the binarization itself.

Finally, we may be tempted to eliminate the reliance not only on the GT, but also on the document image itself. This may be possible utilizing the intrinsic properties of individual characters’ binarization, as proposed in the next section.

7. Quality Evaluation of Individual Characters' Binarizations

7.1 Introduction

This section continues the endeavor of the previous one, in providing a GT-free method for evaluating a binarization. This time, the effort will be based on analyzing the intrinsic properties of the binarizations. In what follows, we concentrate on the scale of individual characters. Our method lacks direct predecessors, with the possible exception of (Trier and Taxt 1995). The former paper proposed a technique somewhat reminiscent of the one specified herein, yet it was performed *manually* upon *visual* inspection of binarizations, and not via a computational approach.

In our scheme, the document binarizations are judged automatically, based on the intrinsic properties of their characters. Four estimates are introduced: stroke width consistency, proportion of stains, average edge curvature, and proportion of edge noise. In certain scenarios, these may be utilized on their own right. Alternatively, these measures can be combined in order to provide the relative ranking of the binarizations. Producing such a model may involve a train-test procedure, depending on the task under consideration (human epigraphic analysis, alphabet reconstruction, OCR, etc.). The current section is a corrected and expanded version of (Faigenbaum, Shaus, Sober et al. 2013).

7.2 Suggested Character Measures

We start by defining independent binarization quality measures, correlating to common human perception. Four measures, pertaining to distinct aspects of binarized images, are proposed and formalized. We will work on small binarized images, each

containing a single character. The challenging problem of characters' segmentation, along with its related topics of concern such as broken strokes and touching characters, is outside the scope of this thesis, and can be handled by methods such as (Casey and Lecolinet 1996; Breuel 2001; Shaus et al. 2012b – see Section 4). The foreground (valued at 0) and the background (valued at 255) areas will be denoted respectively as F and B , with $p = (x, y)$ a pixel coordinate.

Stroke Width Consistency Measure

The local scale consistency of a character stroke width is closely related to the quality of the binarized character. Indeed, partially erased letters, or the presence of stains may introduce discontinuities in stroke width. An example of such behavior can be seen at Fig. 7.1.



Figure 7.1 Example of local-scale stroke width discontinuity due to stains and letter erosion (discontinuities marked in **red**).

The idea is not simply to measure the width of a stroke at every point, but to assess the smoothness of its change between adjacent pixels. The measure is defined by the following algorithm (though devised independently, our first step is reminiscent of Epshtein et al. 2010, while steps 2 and 3 are original).

Step 1 - Evaluate the stroke width $SW(p)$ for each $p \in F$:

- For each angle $\alpha \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$, examine the line segments with inclination α passing through p and restricted to F . Among these, denote the *longest*

segment (i.e. the one running from one edge of the character to another, as opposed to its sub-segments) as $seg(p, \alpha)$.

- Define:

$$SW(p) = \min_{\alpha} \|seg(p, \alpha)\|_2 \quad (7.1)$$

In other words, after Step 1, each pixel $p \in F$ possesses an associated stroke width $SW(p)$; see illustration on Fig. 7.2.

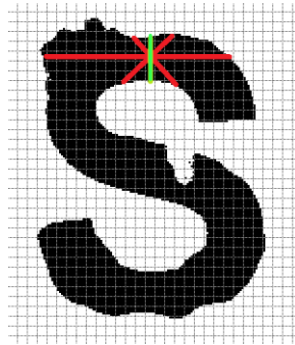


Figure 7.2 A demonstration of shortest stroke width = segment selection for a particular foreground pixel (in **green** – the shortest segment, in **red** – other segments considered).

Step 2 – Calculate the stroke width gradient magnitude $G(p)$:

- Calculate an approximation of directional derivatives $SW_x(p)$ and $SW_y(p)$ by subtracting adjacent pixels along the x and the y axes.
- Define the gradient magnitude with respect to L_{∞} norm:

$$G(p) = \max(|SW_x(p)|, |SW_y(p)|) \quad (7.2)$$

Step 3 – Apply the measure:

$$M_{SWC} = \underset{p \in F}{\text{mean}}(G(p)) \quad (7.3)$$

Note that given a clean binarization with gradually changing stroke widths, $G(p)$ yields low values, resulting in a small M_{SWC} .

Stains Proportion Measure

The existence of black spots within a white background, or vice versa, is an indication of either an imperfect binarization or the presence of noise. In what follows, we will consider *the stains relative area in pixels*, denoted below as $\|\dots\|$. While stains count may be used instead, according to our experiments, this measure performs poorly.

The image is partitioned into a set of Connected Components $CC = \{cc_i\}_{i=1}^N$; these belong to either F or B . The set of *Stain* CCs is defined as: $SCC = \{cc_i \in CC \mid \|cc_i\| \leq Thr\}$. Throughout our experiments, the value of the threshold Thr was set to 0.5% of the character image size. The measure definition is:

$$M_{SP} = \frac{\sum_{cc_j \in SCC} \|cc_j\|}{\sum_{cc_i \in CC} \|cc_i\|} \quad (7.4)$$

Average Edge Curvature Measure

The “ideal” letter is expected to possess a smooth edge. This is tightly related to the average edge curvature (herein, we use its absolute value):

$$\kappa = \left| \frac{dT}{ds} \right| = \left| \frac{d\theta}{ds} \right| \cong \left| \frac{\Delta\theta}{\Delta s} \right| \quad (7.5)$$

where T is the normalized tangent of the edge curve; θ is the tangent angle; and S is the arclength parameter. The computation of the average edge curvature is as follows:

Step 1 – Find the edge via 4-connectivity erosion of F :

$$E = F \setminus erosion(F) \quad (7.6)$$

Step 2 – Calculate the local angle:

- For each pixel $p \in E$, and for each pair of its neighboring pixels $p_1, p_2 \in E$ (assuming 8-connectivity), define the unit vectors $v_k(p) = (p_k - p) / \|p_k - p\|_2$ for $k = 1, 2$. Next, we find $\psi(p)$, the angle between $v_1(p)$ and $v_2(p)$:

$$\psi(p) = \arccos \langle v_1(p), v_2(p) \rangle \quad (7.7)$$

- The angle $\Delta\theta(p)$, used for the curvature definition, is:

$$\Delta\theta(p) = \pi - \psi(p) \quad (7.8)$$

Due to the definition of \arccos , $\psi(p) \in [0, \pi]$ and $\Delta\theta(p) \in [0, \pi]$. See an illustration in Fig. 7.3.

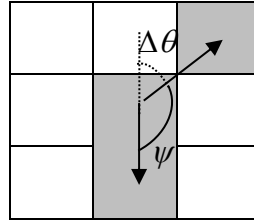


Figure 7.3 An illustration of Step 2 in average edge curvature measure computation.

Step 3 – Approximate the local curvature:

Using Eq. 7.5, and plugging-in Eqs. 7.7 and 7.8, the curvature is defined as:

$$\kappa(p) \cong \frac{\Delta\theta(p)}{\Delta s(p)} = \frac{\pi - \arccos \langle v_1(p), v_2(p) \rangle}{\sum_{k=1}^2 \|p_k - p\|_2} \quad (7.9)$$

Step 4 – Apply the measure:

$$M_{AEC} = \underset{p \in E}{\text{mean}}(\kappa(p)) = \frac{\sum_{p \in E} \Delta s(p) \kappa(p)}{\sum_{p \in E} \Delta s(p)} \quad (7.10)$$

It should be also stated that in certain cases, $p \in E$ might possess more than two neighboring pixels. In such a case, we account for all possible neighboring pairs in Steps 2-4. An example with 3 neighboring pixels is illustrated in Figure 4.

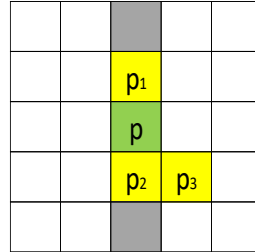


Figure 7.4 An example of edge pixel p , possessing three neighbors p_1 , p_2 and p_3 . This requires an adjustment in M_{AEC} calculations.

Edge Noise Proportion Measure

Another suggested property is the presence of typical edge noise, see (McGillivray et al. 2009) for details. This type of noise is assumed to correlate with the overall quality of the binarization. The current parameter is calculated via a simplified procedure. In what follows, we perform common morphological operations assuming 4-connectivity.

Step 1 – Define the edge utilizing dilation and erosion of F :

$$\bar{E} = dilation(F) \setminus erosion(F) \quad (7.11)$$

(Note \bar{E} is different than E in Eq. 7.6)

Step 2 – Calculate a noise estimate:

$$N = (closure(F) \setminus F) \cup (F \setminus opening(F)) = closure(F) \setminus opening(F) \quad (7.12)$$

In other words, the closure handles isolated white pixels assumed to be “salt” noise, by attaching them to F . Similarly, the opening removes secluded F pixels, assumed to be “pepper” type noise. N provides a set of all estimated noise pixels.

Step 3 – Measure definition:

$$M_{ENP} = \frac{\|N\|}{\|\bar{E}\|} \quad (7.13)$$

Note: For all measures, the cases where an insufficient number of either F or B pixels exists within the character image, were detected and treated in the following fashion. If a dilation of F (assuming 4-connectivity and performed twice) leaves no B pixels, or if an erosion of F (assuming 4-connectivity and performed twice) leaves no F pixels, the image was declared as possessing “lacking information”. In such a case, all the measures discussed above were set to Inf value (we used $Inf = 32768$).

A small-scale example of the four measures applied on both clean and corrupted characters is shown in Fig. 7.5 and Table 7.1.

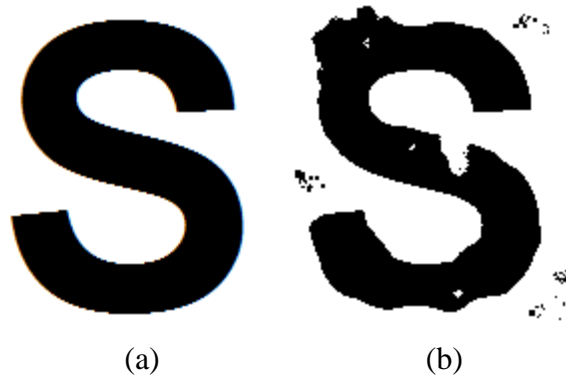


Figure 7.5 (a) Clean character, (b) Corrupted character.

Table 7.1 Comparison of quality measures, activated on clean (Fig. 7.5a) and corrupted (Fig. 7.5b) images.

Measure		Clean image Fig. 7.5a	Corrupted image Fig. 7.5b
Stroke width consistency	M_{SWC}	1.185	2.413
Stains proportion	M_{SP}	0	0.260
Average edge curvature	M_{AEC}	0.407	1.352
Edge noise proportion	M_{ENP}	0.004	0.646

As expected, the measures produce considerably smaller results for Fig 7.5a (clean image) than for Fig 7.5b (corrupted image).

Measures' Combinations

The measures presented above can be applied on their own right, each assessing a distinctive character feature, susceptible to noise. In fact, in certain settings, we have seen some of them producing judgments comparable to human appraisals. Conversely, these measures can be combined into a joint score or classifier, depending on the task under consideration. These may vary according to the type of writing in question (printed or handwritten), medium, corpora, noise characteristic, binarizations end goal (epigraphical research, character reconstruction, OCR), etc. Subsequently, we do not suggest that the combinations derived below to be the ultimate model in all conceivable cases. We do suggest a procedure to derive models for settings comparable to ours. With certain adjustments, these ideas may also be applicable for training binarization quality control apparatus for other tasks.

The combinations dealt with below are linear and tree models, used due to their simplicity. These models require training and testing phases, based on experts' estimations. Such a procedure is presented in the next sub-section. The trees were implemented via the tree package (Ripley 2016) of the R programming language (R Core Team 2012).

7.3 Experimental Design

Motivation

The motivation behind this research was an attempt at ranking binarizations according to their suitability for human and computer-based handwriting analysis. Visually appealing binarizations, faithful to the document images, were preferred.

Dataset

Our database consisted of segmented characters, along with their binarizations. We used characters originating from two different First Temple Period Hebrew inscriptions: 50 images (characters) were taken from Arad No. 1 (Aharoni 1981), while 47 images (characters) were obtained from Lachish No. 3 (Torczyner et al. 1938). The segmentation into individual characters was performed via the algorithm from Section 4. The state of preservation of these ink-over clay samples was poor, presenting a challenge for our methodology.

The 9 binarizations used were: Otsu (Otsu 1979), Bernsen (Bernsen 1986) with window sizes (in pixels) of $w = 50$ and $w = 200$, Niblack (Niblack 1986) with $w = 50$ and $w = 200$, Sauvola (Sauvola and Pietikainen 2000) with $w = 50$ and $w = 200$, as well as our own binarization (see Section 4) with or without unspeckle stage.

From the *97 original grayscale images*, a database of *873 (97 x 9) binary images* was constructed. Each set of 9 binarizations, denoted herein as a “binarization block”, was judged independently by three different experts. The experts’ rankings (from 1=high, up to 9=low) were based on their prior epigraphical knowledge. An example of a single expert’s opinion is presented in Fig. 7.6.

Constructing such a data set with manual ranking information for different binarization procedures is a highly labor-intensive procedure. This explains the relatively modest size of our database.

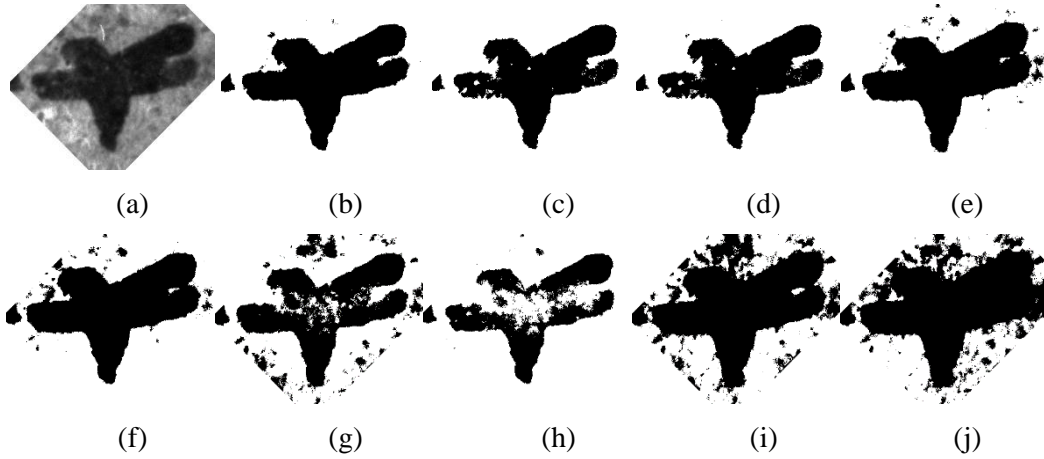


Figure 7.6 Expert's ranking of one character, in decreasing quality order.
 (a) Original image, (b) Sauvola $w = 200$, (c) Shaus et al. inc. unspeckle stage,
 (d) Shaus et al., (e) Otsu, (f) Niblack $w = 200$, (g) Niblack $w = 50$,
 (h) Sauvola $w = 50$, (i) Bernsen $w = 50$, (j) Bernsen $w = 200$.

Goal

The experiment attempted at creating a model matching the three experts' ranking. The model types under consideration were linear and tree-based regressions (Tree 2011). These models used the four *rankings* based on the measures M_{SWC} , M_{SP} , M_{AEC} and M_{ENP} . The utilization of rankings, rather than measure values, provides a common scale across different letters. The experiment consisted of model selection and model verification stages. Both necessitated the prerequisites specified in the next subsection.

Input Data

As stated previously, each binarization block (containing 9 binarizations) for each of the 97 letters, had 3 expert rankings. Resulting vectors of length $97 \times 3 = 873$, containing rankings of binarization blocks in a stacked manner (i.e. containing concatenated rankings of the 9 binarizations of the 1st image, then rankings of the 2nd image, and so forth), are denoted as R_1, R_2, R_3 (one for each expert). For training

purposes, a combined experts ranking $R_{experts}$ was derived. First, $R_{mean} = mean(R_1, R_2, R_3) = (R_1 + R_2 + R_3) / 3$, was calculated, possibly containing non-integer values. Then, a re-ranking of R_{mean} enforced scores of 1...9 within each binarization block, resulting in $R_{experts}$ (e.g., if the mean scores were 1.33, 2, 3.33, 3.67, 6.33, 5, 6.67, 8.67 and 8, the re-ranking results in 1, 2, 3, 4, 6, 5, 7, 9, 8). Such process is denoted below as “re-ranking procedure”. Independently, the 4 different measures produced their own rankings for every binarization block, yielding the corresponding vectors R_{SWC} , R_{SP} , R_{AEC} and R_{ENP} .

Models' Specifications

Both linear and tree-based regression models were considered for estimation purposes. The independent variables were R_{SWC} , R_{SP} , R_{AEC} and R_{ENP} , while the dependent variable was $R_{experts}$.

The linear regression models differed from each other by the presence or absence of independent variables (*15 possible combinations*). A presence or an absence of an intercept was meaningless, since the model's prediction was re-ranked.

The tree regression models differed from each other by the presence or absence of independent variables, as well as by their depths. 2 configurations were attempted: default setting of the library (Ripley 2016), as well as a “forced” tree with 9 leaves. This resulted in a total of 30 tree-based models under consideration.

Models' Score, Selection and Success Criteria

A model m was scored in the following fashion. A prediction produced by the model was re-ranked, resulting in R_m , which was then compared with the experts ranking via standard linear (COR) or Kendall (τ ; Kendall 1938) correlations:

$$c_m = \min_{i=1..3} (cor(R_i, R_m)) \quad (7.14)$$

$$\tau_m = \min_{i=1..3} (\tau(R_i, R_m)) \quad (7.15)$$

The model corresponding to the highest c_m and τ_m scores was selected. As will be seen, in this experiment, both scores resulted in the same selected model.

Since occasionally even human experts differ in their judgments, we did not expect the best model to perform flawlessly, but in a “human-like” fashion. Our golden standards were the *minimal* correlations between pairs of human experts, denoted as c_{expert} and τ_{expert} . Our optimal model was expected to adhere to the following pre-established success criteria:

$$c_m \leq 0.8 \cdot \min_{1 \leq i < j \leq 3} (cor(R_i, R_j)) = 0.8 \cdot c_{expert} \quad (7.16)$$

$$\tau_m \leq 0.8 \cdot \min_{1 \leq i < j \leq 3} (\tau(R_i, R_j)) = 0.8 \cdot \tau_{expert} \quad (7.17)$$

Selected Model

The selected model, for both c_m and τ_m scores, was a tree with 9 leaves, of depth 6. The tree used rankings from all 4 measures, with the most important one (used for the upper splits) being R_{ENP} , with $c_m = 0.678$ and $\tau_m = 0.543$. The resulting tree (a “forced” tree with 9 leaves) can be seen at Fig. 7.7, while an agreement with the

success criteria can be seen at Table 7.2. Since $c_{expert} = 0.768$ and $\tau_{expert} = 0.634$, both success criteria were met.

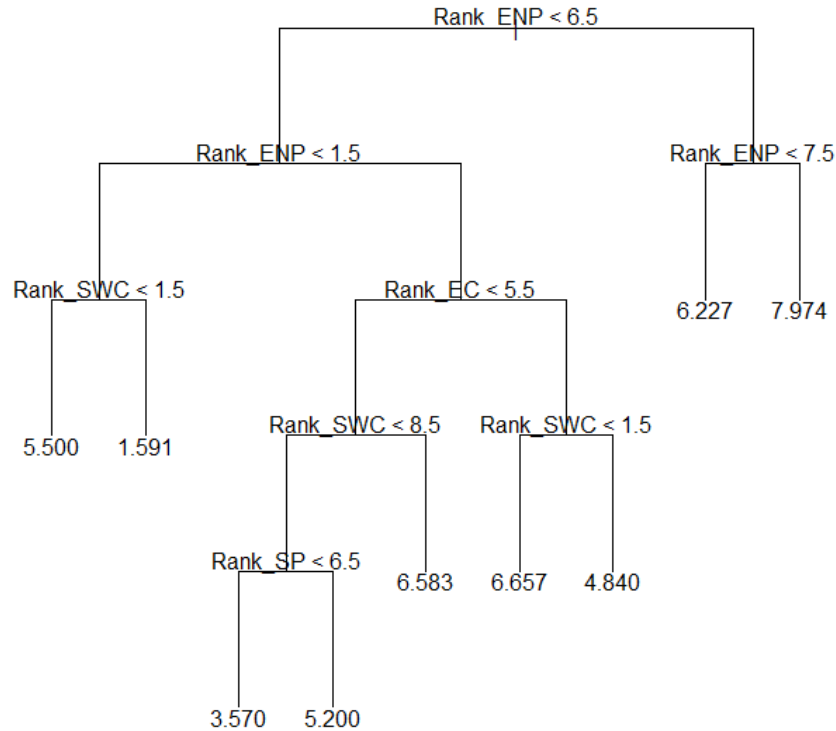


Figure 7.7 The selected regression model, a “forced” tree with 9 leaves. The leaves indicate the mean predicted rank (prior to re-ranking; after applying the ranking function 1.591 will become 1, 3.57 will become 2, etc.). Note that all four proposed measures are utilized by the selected model.

Table 7.2 Agreement with success criteria.

Minimal model correlation	Minimal experts' correlation	%
$c_m = 0.678$	$c_{expert} = 0.768$	$\frac{c_m}{c_{expert}} = 88.2\% > 80\%$
$\tau_m = 0.543$	$\tau_{expert} = 0.634$	$\frac{\tau_m}{\tau_{expert}} = 85.7\% > 80\%$

Selected Model Verification

The selected *model type* (a tree with 9 leaves and all independent variables) was bootstrapped in order to check its robustness. Each iteration performed a 50-50 test/train separation on the binary blocks level (thus, all the binarizations of a single character

were assigned either to train or to test data, avoiding possible bias). Subsequently, a *new model* was trained and tested.

The bootstrap included 1000 iterations, resulting in p -value=0.05 confidence intervals of [0.582, 0.74] for c_m , and [0.454, 0.610] for τ_m . These indicate the robustness of our model.

7.4 Summary

Following inherent obstacles in GT-based quality evaluation of binary images, we proposed a solution based on several intrinsic properties of individual binary characters. Four binarization quality measures were introduced: stroke width consistency, proportion of stains or edge noise, and average edge curvature. In certain scenarios, these may suffice on their own right. Alternatively, a combination of these scores can be trained for specific purposes, such as paleographical analysis, character reconstruction or OCR. For our uses, a tree-based model produced adequate and robust results.

8. Writers' Identification via a Combination of Features, with Historical Implications

8.1 Introduction

Based on biblical exegesis and historical considerations scholars debate whether the first major phase of compilation of biblical texts in Jerusalem took place before or after the destruction of the city by the Babylonians in 586 BCE (e.g., Schmid 2012). A related – and also disputed issue – is the level of literacy, that is, the basic ability to communicate in writing, especially in the Hebrew kingdoms – Israel and Judah (Rollston 2010). The best way to answer this question is to look at the material evidence – the corpus of inscriptions that originated from archaeological excavations (e.g., Ahituv 2008). Inscriptions citing biblical texts, or related to them, are rarely found (for two Jerusalem amulets possibly dating to this period, echoing the priestly blessing in Numbers 6: 23-26 see Barkay 1992; Barkay et al. 2004), probably because papyrus and parchment are not well preserved in the climate of the region. However, ostraca (inscriptions in ink on ceramic sherds) which deal with more mundane issues can also shed light on the volume and quality of writing and on the recognition of the power of the written word in the society.

In order to explore the degree of literacy and stage-setting for compilation of literary texts in monarchic Judah, we turned to Hebrew ostraca from the final days of the kingdom, prior to its destruction by Nebuchadnezzar in 586 BCE and the deportation of its elite to Babylonia. Several corpora of inscriptions exist for this period. We focused on the corpus of over 100 Hebrew ostraca found at the fortress of Arad, located in arid southern Judah, on the border of the kingdom with Edom (see Aharoni 1981 and Fig. 8.1 below). The inscriptions contain military commands regarding movement of troops and provision of supplies (wine, oil and flour) set against the

background of the stormy events of the final years before the fall of Judah. They include orders that came to the fortress of Arad from higher echelons in the Judahite military system, as well as correspondence with neighboring forts. One of the inscriptions mentions "the King of Judah" and another "the house of YHWH," referring to the Temple in Jerusalem. Most of the provision orders that mention the *Kittiyim* – apparently a Greek mercenary unit (Na'aman 2010) – were found on the floor of a single room. They are addressed to a person named Eliashib – the quartermaster in the fortress. It has been suggested that most of Eliashib's letters involve the registration of about one month's expenses (Lemaire 1977).

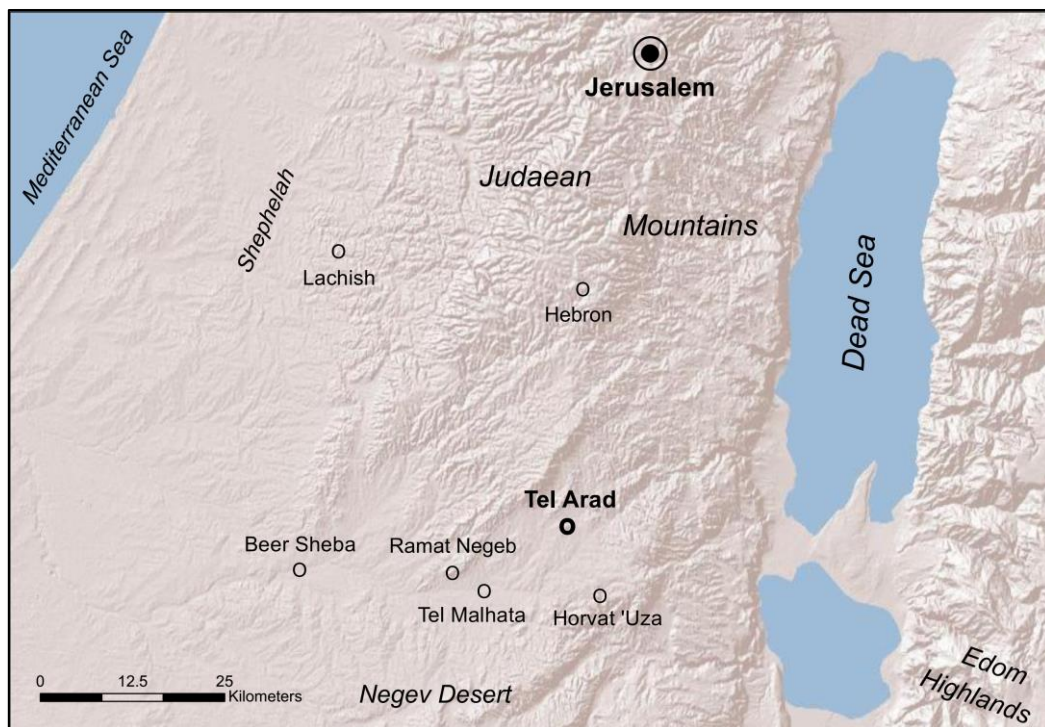


Figure 8.1 Main towns in Judah and sites in the Beer Sheba Valley mentioned in the current section.

Of all the corpora of Hebrew inscriptions, Arad provides the best set of data for exploring the question of literacy at the end of the First Temple period: A) the lion's share of the corpus represents a short time span of a few years ca. 600 BCE; B) it comes from a remote region of the kingdom, where the spread of literacy is more significant

than its dissemination in the capital; C) it is connected to Judah's military administration and hence bureaucratic apparatus. Identifying the number of "hands" (i.e., authors) involved in this corpus can shed light on the dissemination of writing, and consequently on the spread of literacy in Judah. The current section is a refinement of the material in (Faigenbaum-Golovin, Shaus, Sober et al. 2016), supplemented with information from (Shaus and Turkel 2017a).

8.2 Prior Art

The problem of computerized writer identification within historical documents exists in the literature for several decades. Several features and their combination methods have been proposed for that purpose. The paper (Dinstein and Shapira 1982) uses run-length histograms, combined via PCA (first two components). Article (Bulacu and Schomaker 2007) continues the use of run-lengths distributions, supplementing them with allographic features (grapheme codebook generated using self-organizing map); the feature fusion is performed via simple or weighted averaging distances due to the individual features. Similar allographic features (“fraglets”), optionally supplemented with edge-directional feature (“hinge”) are present in (Schomaker et al. 2007), with Hamming distance measures between the normalized features.

The paper (Bar-Yosef et al. 2007) presents another feature combination technique; extracting 8 types of features pertaining to various relations between foreground and background pixels of segmented characters, as well as their central moments. The features are selected via dimension reduction techniques such as sequential forward floating selection and linear discriminant analysis, classifying the reduced feature vectors via a linear Bayes classifier or K-nearest neighbors. Yet another

set of classifiers is based on grid microstructure, allograph-level and topological features, combined via weighting procedure, is presented in (Aiolli and Giollo 2011). Article (Fecker et al. 2014a) provides a wealth of contour-based, oriented basic image, as well as SIFT, features classified by a voting procedure and SVM; (Fecker et al. 2014b) uses a similar setup, adding HOG features. An adaptation of SIFT features is also used in (Fiel et al. 2014), with dimensionality reduced via PCA, resulting in a visual vocabulary. The features are clustered using a Gaussian Mixture Model and employing the Fisher kernel. A recent use of KDA in a setting involving both chain-code and edge-based directional features can be found in (Al-Maadeed et al. 2016).

A thoroughly different approach is demonstrated in (Panagopoulos et al. 2009), operating on a segmented character level, and treating them as realizations of estimated “Ideal Prototypes”. The identity or distinction among writers is made via several techniques, employing comparisons of the contours of the realizations to various ideals, and using heuristics and maximum likelihood estimations procedure combining information from different letters in order to find similar writers. A similar method is described in (Papaodysseus et al. 2014), with comparisons between character or ideal contours solved analytically.

A review of these papers, as well as surveys of the broader field of writer identification (Schomaker 2007; Sreeraj and Idicula 2011) demonstrate the common denominators of most of these algorithms: a series of features (e.g. based on edge, allograph or topological information, or using “classical” computer vision features such as SIFT, HOG and Gabor filters) is extracted. Optionally, the dimensionality is reduced (e.g. via weighting, LDA, PCA or KDA methods), followed by writer classification of the resulting feature vectors (e.g. by employing KNN, SVM, MLE or the Fisher Kernel). Usually, the question is whether a given document, according to some metric,

is written by the same author as the most closely matching document. Alternatively, several (e.g. 5 or 10) “closest” documents are fetched for the purpose of identifying at least one identical writer. The algorithm’s performance is checked based on an existing ground truth.

Although some of these methods perform reasonably well for their tasks and data-sets, their typical output is an abstract distance between two given inscriptions, or else a table indicating the distances between several inscriptions. However, these distances do not yield any probabilistic information. Thus, it is difficult to interpret such an output outside a well ground-truthed framework. In particular the distances, by themselves, are insufficient for the different task of analyzing a corpus of many inscriptions, with an unknown number of authors. The main contribution of the current research, detailed below, is a proposal of an entirely different approach, allowing for an **estimation of the minimal number of writers within the given corpus.**

8.3 Materials and Methods

This research was conducted on two datasets of written material. The foremost document assemblage was a corpus of 16 Hebrew ostraca inscriptions found at the Arad fortress (ca. 600 BCE). The research was performed on digital images of these inscriptions. A second dataset, used to validate the algorithm, contained handwriting samples collected from 18 present-day writers of Modern Hebrew.

The aim of our core algorithm was to differentiate between writers in a given set of texts. This algorithm consisted of several stages. In the first step, character restoration, the image of the inscription was segmented into (often noisy) characters that were restored via a semi-automatic reconstruction procedure. The method was

based on the representation of a character as a union of individual strokes that were treated independently and later recombined. The purpose of stroke restoration was to imitate a reed-pen's movement using several manually sampled key-points. An optimization of the pen's trajectory was performed for all intermediate sampled points. The restoration was conducted via the minimization of image energy functional, which took into account the adherence to the original image, the smoothness of the stroke, as well as certain properties of the reed radius. The minimization problem was solved by performing Gradient Descent iterations on a Cubic-Spline representation of the stroke. The end product of the reconstruction was a binary image of the character, incorporating all its strokes.

The second stage of the algorithm, letter comparison, relied on features extracted from the characters' binary images, utilized in order to automatically compare characters from different texts. Several features were adapted, referring to aspects such as the character's overall shape, the angles between strokes, the character's center of gravity, as well as its horizontal and vertical projections. The features in use were: SIFT (Lowe 2004), Zernike (Tahmasbi et al. 2011), DCT, Kd-tree (Sexton et al. 2000), Image projections (Trier et al. 1996), L_1 and CMI (see sections 2-4 and 6). Additionally, for each feature, a respective distance was defined. Later on, all these distances were combined into a single, generalized feature vector. This vector described each character by the degree of its proximity to all the characters, using all the features. Finally, a distance between any two characters was calculated according to the Euclidean distance between their generalized feature vectors.

The final stage of the algorithm addressed the main question: "What is the probability that two given texts were written by the same author?" This was achieved by posing an alternative null hypothesis H_0 ("both texts were written by the same

author”) and attempting to reject it by conducting a relevant experiment. If its outcome was unlikely ($P \leq 0.2$; a threshold established in advance by our collaborating archaeologists), we rejected the H_0 and concluded that the documents were written by two individuals. Alternatively, if the occurrence of H_0 was probable ($P > 0.2$), we remained agnostic. The experiment testing the H_0 performed a clustering on a set of letters from the two tested inscriptions (of specific type, e.g., *alep*), disregarding their affiliation to either of the inscriptions. The clustering results should have resembled the original inscriptions if two different writers were present, while being random if this was not the case. While this kind of test could have been performed on one specific letter, we could gain additional statistical significance if several different letters (e.g., *alep*, *he*, *waw*, etc.) were present in the compared documents.

Subsequently, several independent experiments were conducted (one for each letter), and their P values were combined via the well-established Fisher’s method (Fisher 1925). The combination represented the probability that H_0 was true based on all the available experimental data.

8.4 Algorithmic Apparatus

The main goal of the current research was to estimate the minimal number of authors involved in the scripting of the Arad corpus. In order to deal with this issue, we had to differentiate between authors of different inscriptions. Although relevant algorithms have been proposed in the past, none offered a systematic technique for establishing a minimal number of authors within the given corpus. In addition, the poor state of preservation of the Arad First Temple period ostraca, the conciseness of the inscriptions, and the high variance of their cursive texts of mundane nature, presented

difficulties that none of the available methods could overcome (see Fig. 8.2). Therefore, novel image processing and machine learning tools had to be developed.



Figure 8.2 Ostraca from Arad (Aharoni 1981): No. 5 (A), No. 24 (B) and No. 40 (C). The poor state of preservation, including stains, erased characters and blurred text, is evident.

The input for our system is the digital images of the inscriptions. The algorithm involves two preparatory stages, leading to a third step that estimates the probability that two given inscriptions were written by the same author. All the stages are fully automatic, with the exception of the first, semi-automatic, preparatory step. The basic steps of the algorithm are:

- A. **Restoring characters** via approximation of their composing strokes, represented as a spline-based structure, and estimated by an optimization procedure (for further details see Sober and Levin 2017).
- B. **Feature Extraction and Distance Calculation:** creation of feature vectors describing the characters' various aspects (e.g., angles between strokes and character profiles); calculating the distance (similarity) *between characters*.
- C. **Testing the hypothesis that two given inscriptions were written by the same author.** Upon obtaining a suitable P-value (the significance level of the test, denoted as P), we reject the hypothesis of a *single author* and accept the competing proposition of *two different authors*; otherwise we remain undecided.

As already stated, step A. is implemented via the technique elaborated in (Sober and Levin 2017). Below, we present an in-depth description of stages B and C.

Feature Extraction and Distance Calculation

Commonly, automatic comparison of characters relies upon features extracted from the characters' binary images. In this study, we adapted several well established features from the domains of Computer Vision and Document Analysis. These features refer to aspects such as the character's overall shape, the angles between strokes, the character's center of gravity, as well as its horizontal and vertical projections. Some of these features correspond to characteristics commonly employed in traditional paleography (Rollston 1999).

The feature extraction process includes a preliminary step of the characters' standardization. The steps involve rotating the characters according to their line inclination, resizing them according to a pre-defined scale, and fitting the results into a padded (at least 10% on each side) square of size $a_L \times a_L$ (with $L = 1, \dots, 22$ the index of the alphabet letter under consideration). On average, the resized characters were 300 by 300 pixels.

Subsequently, the proximity of two characters can be measured using each of the extracted features, representing various aspects of the characters. For each feature, a *different* distance function is defined (to be combined at a later stage).

Table 8.1 provides a list of the features and distances we employ, along with a description of their implementation details. Some of the adjustments (e.g., replacement of the L_2 norm with the L_1 norm) were required due to the large amount of noise present in our medium.

After the features are extracted, and the distances between the features are measured, there arises a challenge of combining the various distances. Several combination techniques (e.g.: AdaBoost, Freund and Schapire 1997; and Bag of Features, Sivic and Zisserman 2003) were considered. Unfortunately, boosting-related methods are unsuitable due to the lack of training statistics, while the Bag of Features performed poorly in preliminary experiments using a modern handwritten character dataset (see details regarding this dataset below). Hence, we developed a different approach for combining the distances.

Table 8.1 Features and distances used in our algorithm.

Feature (reference)	Feature implementation details	Distance implementation details
SIFT (Lowe 2004)	For each character j , we use the normalized SIFT descriptors $\vec{d}_i \in \mathbb{R}^{128}$ (with $\ \vec{d}_i\ _2 = 1$) and the spatial locators $\vec{l}_i \in [1, a_L]^2$ for at most 40 significant key points $k_i = (\vec{d}_i, \vec{l}_i)$, according to the original SIFT implementation. The resulting feature is a set $f_j^{SIFT} = \{k_i\}_{i=1}^{40}$.	The distance between f_1^{SIFT} and f_2^{SIFT} is determined as follows: 1. For each key point $k_i^1 \in f_1^{SIFT}$, find a matching key point $m_i^2 \in f_2^{SIFT}$ s. t. $m_i^2 = \arg \min_{(d_j^2, l_j^2) \in f_2^{SIFT}} \text{dist}(k_i^1, k_j^2)$; where $\text{dist}(k_i^1, k_j^2) = \arccos(\langle d_i^1, d_j^2 \rangle) \cdot \ \vec{l}_i^1 - \vec{l}_j^2\ _2$. Thus, our definition augments the original SIFT distance by adding spatial information. 2. The one-sided distance is $D_{SIFT}^{1,2} = \text{median}_i \{ \text{dist}(k_i^1, m_i^2) \}$. 3. The final distance is $D_{SIFT}(1, 2) = \frac{D_{SIFT}^{1,2} + D_{SIFT}^{2,1}}{2}$.
Zernike (Tahmasbi et al. 2011)	An off-the-shelf (Tahmasbi 2014) implementation was used. Zernike moments up to the 5 th order were calculated.	$D_{Zernike}$ is the L_1 distance between the Zernike feature vectors.
DCT	MATLAB (R2009a) default implementation was used.	D_{DCT} is the L_1 distance between the DCT feature vectors.
Kd-tree (Sexton et al. 2000)	An off-the-shelf (Armon 2012) implementation was used. Both orders of partitioning are employed (first height, then width and vice-versa)	$D_{Kd-tree}$ is the L_1 distance between the Kd-tree feature vectors.
Image projections (Trier et al. 1996)	The implementation results in cumulative distribution functions of the histogram on both axes.	D_{Proj} is the L_1 distance between the projections' feature vectors; this is similar to the Cramér–von Mises criterion (which uses L_2 distance).
L1	Existing character binarizations.	D_{L1} is the L_1 distance between the character images.
CMI	Existing character binarizations, with values in $\{0, 1\}$.	The CMI computes a difference between the averages of the foreground and the background pixels of \mathcal{J} , marked by a binary mask M , $CMI(M, \mathcal{J}) = \mu_1 - \mu_0$, where: $\mu_k = \text{mean}\{\mathcal{J}(p, q) \mid M(p, q) = k\} \quad k = 0, 1$ In our case, given character-binarizations B_1, B_2 , the one-sided distance is $D_{CMI}^{1,2} = 1 - CMI(B_1, B_2)$. The final distance is $D_{CMI}(1, 2) = \frac{D_{CMI}^{1,2} + D_{CMI}^{2,1}}{2}$.

Our main idea was to consider the distances of a given character from *all the other characters*, with respect to *all of the features* under consideration. I.e., two

characters closely resembling each other, ought to have similar distances from all other characters. Namely, they will both have small distances from similar characters, and large distances from dissimilar characters. This observation leads to a notion of a *generalized feature vector* (defined here for the first time).

The generalized feature vector is defined by the following procedure (for each letter $L = 1, \dots, 22$ in the alphabet). First, we define a *distance matrix* for each feature. For example, the SIFT distance matrix is:

$$U_{SIFT} = \begin{pmatrix} D_{SIFT}(1,1) & \cdots & D_{SIFT}(1, J_L) \\ \vdots & \ddots & \vdots \\ D_{SIFT}(J_L,1) & \cdots & D_{SIFT}(J_L, J_L) \end{pmatrix} = \begin{pmatrix} - & \vec{u}_{SIFT}^1 & - \\ & \vdots & \\ - & \vec{u}_{SIFT}^{J_L} & - \end{pmatrix}, \quad (8.1)$$

where J_L represents the total number of characters; $D_{SIFT}(i, j)$ is the SIFT distance between characters i and j ; while $\vec{u}_{SIFT}^i = (D_{SIFT}(i,1) \cdots D_{SIFT}(i, J_L))$ is the vector of SIFT distances between the character i and all the others.

In addition, we denote the standard deviation of the elements of the matrix U_{SIFT} by $\sigma_{SIFT} = std\{D_{SIFT}(i, j) | (i, j) \in \{1, \dots, J_L\} \times \{1, \dots, J_L\}\}$. Matrices of all the other features ($U_{Zernike}, U_{DCT}$, and so forth) and their respective standard deviations ($\sigma_{Zernike}, \sigma_{DCT}$, etc.) are calculated in a similar fashion.

Therefore, each character k is represented by the following vector (of size $7 \cdot J_L$), concatenating the respective normalized row vectors of the distance matrices:

$$\vec{u}_k = \left(\frac{\vec{u}_{SIFT}^k}{\sigma_{SIFT}} \parallel \frac{\vec{u}_{Zernike}^k}{\sigma_{Zernike}} \parallel \frac{\vec{u}_{DCT}^k}{\sigma_{DCT}} \parallel \frac{\vec{u}_{Kd-tree}^k}{\sigma_{Kd-tree}} \parallel \frac{\vec{u}_{Proj}^k}{\sigma_{Proj}} \parallel \frac{\vec{u}_{L1}^k}{\sigma_{L1}} \parallel \frac{\vec{u}_{CMI}^k}{\sigma_{CMI}} \right) \in \mathbb{R}^{7 \cdot J_L}. \quad (8.2)$$

In this fashion, each character is described by the degree of its kinship to all of the characters, using all the various features.

Finally, the distance between characters i and j is calculated according to the Euclidean distance between their generalized feature vectors:

$$\text{chardist}(i, j) = \|\tilde{u}_i - \tilde{u}_j\|_2. \quad (8.3)$$

The main purpose of this distance is to serve as a basis for clustering at the next stage of the analysis.

Hypothesis Testing

At this stage we address the key question raised above: “*What is the probability that two given texts were written by the same author?*” Commonly, similar questions are addressed by posing an *alternative* null hypothesis H_0 and attempting to reject it. In our case, for each pair of ostraca, the H_0 is: *both texts were written by the same author*. This is performed by conducting an experiment (detailed below) and calculating the probability ($P \in [0,1]$) of affirmative answer to H_0 . If this event is unlikely ($P \leq 0.2$), we conclude that the documents were written by two different individuals (i.e., reject H_0). On the other hand, if the occurrence of H_0 is probable ($P > 0.2$), we remain agnostic. We reiterate that in the latter case we cannot conclude that the two texts were in fact written by a single author.

The experiment, which is designed to test H_0 , is comprised of several sub-steps (illustrated in Fig. 8.3):

1. **Initialization:** We begin with two sets of characters of the same letter type (e.g., *alep*), denoted A and B , originating from two different texts (Fig. 8.3A).
2. **Character clustering:** The union $A \cup B$ is a new, unlabeled set (Fig. 8.3B). This set is clustered into two classes, labeled I and II , using a brute-force (and not

heuristic) implementation of k-means (k=2). The clustering utilizes the generalized feature vectors of the characters, and the distance *chardist*, defined above (Fig. 8.3C).

3. **Cluster labels consistency:** If $|I| > |II|$, their labels are swapped.
4. **Similarity to cluster *I*:** For each of the two original sets, *A* and *B*, the maximal proportion of their elements in class *I* (their “similarity” to class *I*) is defined as:

$$MP_I = \max \left\{ \frac{|A \cap I|}{|A|}, \frac{|B \cap I|}{|B|} \right\}. \quad (8.4)$$

5. **Counting valid combinations:** We consider all the possible divisions of $A \cup B$ into two classes *i* and *ii*, s.t. $|i| = |I|$. The number of such valid combinations is denoted by *NC*.
6. **Significance level calculation:** The P-value is calculated as:

$$P = \frac{|\{i \mid MP_i \geq MP_I\}|}{NC}. \quad (8.5)$$

I.e., *P* is the proportion of valid combinations with at least the same observational *MP*. This is analogous to integrating over a tail of a probability density function.

The rationale behind this calculation is based on the scenario of *two authors* (negation of H_0). In such a case, we expect the k-means clustering to provide a sound separation of their characters (Fig. 8.3D), i.e., *I* and *II* would closely resemble *A* and *B* (or *B* and *A*). This would result in MP_I being close to 1. Furthermore, the proportion of valid combinations with $MP_i \geq MP_I$ will be meager, resulting in a low *P*. In such a case, the H_0 hypothesis would be justifiably rejected.

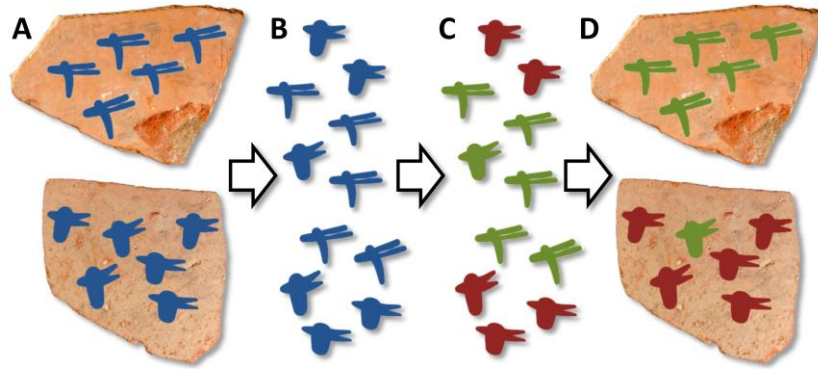


Figure 8.3 Artificial illustration of H_0 rejection experiment (containing only *alep* letters):
 (A) two compared documents; (B) unifying their sets of characters; (C) automatic clustering;
 (D) the clustering results vs. the original documents.

In the opposite scenario of a *single author*:

- If a sufficient number of characters is present, there is an arbitrary low probability of receiving clustering results resembling *A* and *B*. In a common case, the MP_I will be low, which will result in high P .
- Alternatively, if the number of characters is low, the clustering may result in a high MP_I by chance. However, in this case NC would be low, and the P will remain high.

Either way, in this scenario, we will not be able to reject the H_0 hypothesis.

Notes:

- We assume that each given text was written by a single author. If multiple authors wrote the text, both H_0 and its negation should be altered. We do not cover such a case.
- In sub-step 3, the swapping is performed for regularization purposes, since the measurement on sub-step 4 is not symmetric. Sub-step 3 verifies that I is a minority

class, and thus the value of $MP_l = 1$ is achieved only if the clustering resembles the original sets A and B .

- In cases where $|I| = |II|$ (sub-step 3), the results of sub-steps 4-6 can be affected by swapping the classes. To avoid such infrequent inconsistencies, we perform the calculations for both alternatives, and choose the lower P .
- Note that in any case, the definition of P in sub-step 6 results in $P > 0$.
- Not every text provides a sufficient amount of characters for every type of letter in the alphabet. In our case, we do not perform comparisons for sets A and B such that: $|A| = 1 \ \& \ |B| \leq 6$ or $|B| = 1 \ \& \ |A| \leq 6$ or $|A| = 2 \ \& \ |B| = 2$.

As specified, sub-steps 1-6 are applied to one specific letter of the alphabet (e.g., *alep*), present (in sufficient quantities) in the pair of texts under comparison. However, we can often gain additional statistical significance if several different letters (e.g., *alep*, *he*, *waw*, etc.) are present in the compared documents. In such circumstances, several independent experiments are conducted (one for each letter), resulting in corresponding P 's. We combine the different values into a single P via the well-established Fisher method (Fisher 1925; in case no comparison can be conducted for any letter in the alphabet, we assign $P=1$). Given p-values p_i ($i = 1, \dots, k$) stemming from k independent experiments, the method allows one to estimate a combined p-value, reflecting the entire wealth of evidence at our disposal. The technique utilizes the fact that

$$X_{2k}^2 \sim -2 \sum_{i=1}^k \ln(p_i), \text{ i.e. the sum produces a chi-squared distribution with } 2k \text{ degrees}$$

of freedom. This allows for a calculation of a single combined (“meta”) p-value. Intuitively, if several experiments produce low p-values (e.g. 0.1, 0.15 and 0.2), the probability for such an occurrence, by chance, is very small, and the combined p-value

will also be low (possibly even lower than the original p-values; 0.071 for the last example). The combined result represents the probability that H_0 is true based on all the acquired experimental data.

The end product of our algorithm is a table containing the P for a comparison of each pair of ostraca. Prior to implementing our methodology on the Arad corpus, it was thoroughly tested on modern Hebrew handwritings and found solid.

8.5 Results

Our experiments were conducted on two large datasets. The first is a set of samples collected from contemporary writers of Modern Hebrew (Modern Hebrew 2016). This dataset allowed us to test the soundness of our algorithm. It was not used for parameter-tuning purposes, however, as the algorithm was kept as parameter-free as possible. The second dataset contained information from various Arad Ancient Hebrew ostraca, dated to ca. 600 BCE, described in detail above (Ancient Hebrew 2016). Following are the specifications and the results of our experiments for both datasets.

Modern Hebrew experiment

The handwritings of 18 individuals $i = 1, \dots, 18$ were sampled. Each individual filled in a Modern Hebrew alphabet table consisting of ten occurrences of each letter, out of the 22 letters in the alphabet (the number of letters and their names are the same as in Ancient Hebrew; see Fig. 8.4 for a table example). These tables were scanned and their characters were segmented. For a complete data-set of the characters, see the supplementary Modern Hebrew characters dataset.

10	9	8	7	6	5	4	3	2	1	Letter
א	א	א	א	א	א	א	א	א	א	alep ^x
ב	ב	ב	ב	ב	ב	ב	ב	ב	ב	bet ^א
ג	ג	ג	ג	ג	ג	ג	ג	ג	ג	gimel ^ג
ד	ד	ד	ד	ד	ד	ד	ד	ד	ד	dalet ^ד
ה	ה	ה	ה	ה	ה	ה	ה	ה	ה	he ^ה
ו	ו	ו	ו	ו	ו	ו	ו	ו	ו	waw ^ו
ז	ז	ז	ז	ז	ז	ז	ז	ז	ז	zayin ^ז
ח	ח	ח	ח	ח	ח	ח	ח	ח	ח	het ^ח
ט	ט	ט	ט	ט	ט	ט	ט	ט	ט	tet ^ט
י	י	י	י	י	י	י	י	י	י	yod ^י
כ	כ	כ	כ	כ	כ	כ	כ	כ	כ	kap ^כ
ל	ל	ל	ל	ל	ל	ל	ל	ל	ל	lamed ^ל
מ	מ	מ	מ	מ	מ	מ	מ	מ	מ	mem ^מ
נ	נ	נ	נ	נ	נ	נ	נ	נ	נ	nun ^נ
ס	ס	ס	ס	ס	ס	ס	ס	ס	ס	samek ^ס
ע	ע	ע	ע	ע	ע	ע	ע	ע	ע	ayin ^ע
פ	פ	פ	פ	פ	פ	פ	פ	פ	פ	pe ^פ
צ	צ	צ	צ	צ	צ	צ	צ	צ	צ	sade ^צ
ק	ק	ק	ק	ק	ק	ק	ק	ק	ק	qop ^ק
ר	ר	ר	ר	ר	ר	ר	ר	ר	ר	resh ^ר
ש	ש	ש	ש	ש	ש	ש	ש	ש	ש	shin ^ש
ת	ת	ת	ת	ת	ת	ת	ת	ת	ת	taw ^ת

Figure 8.4 An example of a Modern Hebrew alphabet table, produced by a single writer (with 10 samples of each letter).

From this raw data, a series of “simulated” inscriptions were created. Due to the need to test both same-writer and different-writer scenarios, the data for each writer was split. Furthermore, in order to imitate a common situation in the Arad corpus, where the scarcity of data is prevalent (see Table 8.3), each simulated inscription used only 3 letters (i.e., 15 characters; 5 characters for each letter). In total, 252 inscriptions were “simulated” in the following manner:

All the letters of the alphabet except for *yod* (as it is too small to be considered by some of the features), were split randomly into 7 groups (3 letters in each group) $g = 1, \dots, 7$: *gimel, het, resh*; *bet, samek, shin*; *dalet, zayin, ayin*; *tet, lamed, mem*; *nun, sade, taw*; *he, pe, qop*; *alep, waw, kap*.

For each writer i , and each letter belonging to group g , 5 characters were assigned into simulated inscription $S_{i,g,1}$, with the rest assigned to $S_{i,g,2}$.

In this fashion, for constant i and g , we can test if our algorithm arrives at wrong rejection of H_0 for $S_{i,g,1}$ and $S_{i,g,2}$ (FP = “False Positive” error; 18 writers and 7 groups producing 126 tests in total). Additionally, for constant g , $1 \leq i \neq j \leq 18$, and $b, c \in \{1, 2\}$, we can test if our algorithm fails to correctly reject H_0 for $S_{i,g,b}$ and $S_{j,g,c}$ (FN = “False Negative” error; $\frac{18 \times 17}{2} \times 2 \times 2 = 4284$ tests in total).

The results of the Modern Hebrew experiment are summarized in Table 8.2. It can be seen that in modern context, the algorithm yields reliable results in ~98% of the cases (about 2% of both FP and FN errors). These results signify the soundness of our algorithmic sequence. The successful and significant results on the Modern Hebrew dataset paved the way for the algorithm’s application on the Arad Ancient Hebrew corpus.

Table 8.2 Results of the Modern Hebrew experiment.

Group of letters (corresponding to g -index of simulated inscriptions)	False Positive (FP out of all same-writer comparisons)	False Negative (FN out of all different- writer comparisons)	False Positive % (FP out of all same-writer comparisons)	False Negative % (FN out of all different- writer comparisons)
gimel, het, resh	0 / 18	8 / 612	0%	1.31%
bet, samek, shin	1 / 18	5 / 612	5.56%	0.82%
dalet, zayin, ayin	1 / 18	18 / 612	5.56%	2.94%
tet, lamed, mem	0 / 18	22 / 612	0%	3.59%
nun, sade, taw	0 / 18	3 / 612	0%	0.49%
he, pe, qop	0 / 18	16 / 612	0%	2.61%
alep, waw, kap	1 / 18	11 / 612	5.56%	1.80%
Total	3 / 126	83 / 4284	2.38%	1.94%

Arad Ancient Hebrew experiment

As specified above, the core experiment addresses ostraca from the Arad fortress, located on the southern frontier of the kingdom of Judah. These inscriptions belong to a short time span of a few years, ca. 600 BCE, and are comprised of army correspondence and documentation.

The texts under examination are sixteen Ostraca 1, 2, 3, 5, 7, 8, 16, 17, 18, 21, 24, 31, 38, 39, 40 and 111. These inscriptions are relatively legible and have a sufficient number of characters for investigation. Moreover, Ostraca 17 and 39 contain writing on both sides of the potshard, and were treated as separate texts (17a and 17b; 39a and 39b), resulting in eighteen texts under examination. As stated in the algorithm description, we assume that each text was written by a single author. A concise summary of the content of the texts can be seen in Table 8.4.

The seven letters we utilized were: *alep*, *he*, *waw*, *yod*, *lamed*, *shin* and *taw*, as they were the most prominent and simple to restore. In the abovementioned ostraca, out of the 670 deciphered characters of these types in the original publication (Aharoni 1981), 501 legible characters were restored, based upon computerized images of the inscriptions. These images were obtained by scanning the negatives taken by the Arad expedition (courtesy of the Israel Antiquities Authority and the Institute of Archaeology of Tel Aviv University). After performing a manual quality assurance procedure (verifying the adherence of the restored characters to the original image), 427 restored characters remained. The resulting letters' statistics for each text are summarized in Table 8.3. For a complete data-set of the characters, see the supplementary Arad Ancient Hebrew characters dataset. In addition, a comparison between several specimens of the letter *lamed* is provided in Fig. 8.5.

Table 8.3 Letter statistics for each text under comparison.

Text	Alphabet letters						
	<i>Alep</i>	<i>He</i>	<i>Waw</i>	<i>Yod</i>	<i>Lamed</i>	<i>Shin</i>	<i>Taw</i>
1	4	5	3	7	3	3	8
2	6	3	3	5	3	1	7
3	2	4	5	4	4	3	3
5	5	3	1	3	4	2	4
7	1	2	1	4	6	8	5
8	2	1	2	1	4	4	2
16	6	3	9	5	10	3	2
17a	2	4	2	2	2	1	2
17b		1		2	1	1	2
18	2	4	4	5	6	6	3
21	5	4	6	6	12	5	2
24	9	10	5	8	4	4	7
31	3	7	6	4	1	1	
38	1	1	2	2	2	1	
39a	3	3	3	5	2	1	1
39b	3	1	1	4	1		
40	4	5	3	4		3	2
111	4	3	3	3	1	3	2

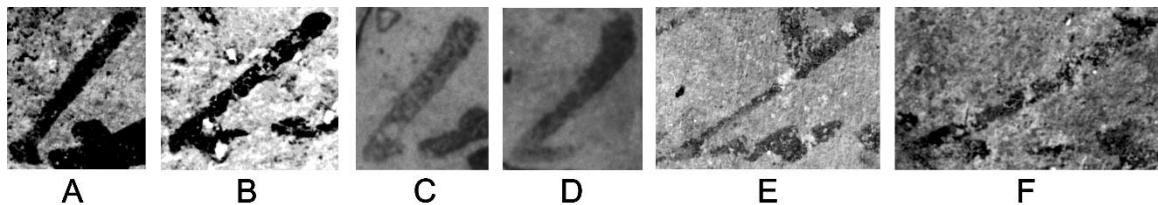


Figure 8.5 Comparison between several specimens of the letter *lamed*, stemming from: Arad 1 (A, B); Arad 7 (C, D) and Arad 18 (E, F). Note that our algorithm cannot distinguish between the author of Arad 1 and the author of Arad 7, or the authors of Arad 1 and Arad 18. On the other hand, Arad 7 and Arad 18 were probably written by different authors ($P=0.015$ for the letter *lamed* and $P=0.004$ for the whole inscription, combining information from different letters).

We reiterate that our algorithm requires a minimal number of characters in order to compare a pair of texts. For example, when we compared Ostraca 31 and 38, the letters in use were *he* (7:1 characters), *waw* (6:2 characters) and *yod* (4:2 characters). The three independent tests respectively yielded $P = 0.125$, $P = 0.25$ and $P = 1$. Their combination through Fisher's method resulted in the final value of $P = 0.327$, not passing the pre-established threshold. Therefore, in this case, we remain agnostic with respect to the question of common authorship. On the other hand, the comparison of

texts 1 and 24 used all possible letters: *alep, he, waw, yod, lamed, shin* and *taw*; resulting in P 's of 0.559, 0.00366, 0.375, 0.119, 0.0286, 0.429 and 0.0769, respectively. The combined result was $P = 0.003$, passing the threshold of 0.2 (again, this threshold was established in advance by our collaborating archaeologists). Therefore, in the latter case, we reject the H_0 hypothesis and conclude that these texts were written by two different individuals.

The complete comparison results are summarized in Table 8.4. The ostraca numbers head the rows and columns of the table, with the intersection cells providing the comparisons' P . The cells with $P \leq 0.2$ are marked in red, indicating that the two ostraca are considered to be written by different authors. We reiterate that when $P > 0.2$, we cannot claim that they were written by a single author.

We can observe six pair-wise distinct “quadruplets” of texts: I) **7, 17a, 24** and **40**; II) **5, 17a, 24** and **40**; III) **7, 18, 24** and **40**; IV) **5, 18, 24** and **40**; V) **7, 18, 24** and **31**; VI) **5, 18, 24** and **31**. The existence of no less than six such combinations indicates the high probability that the corpus indeed contains at least four different authors. It can be claimed that the results do not take into account the multiple comparisons taking place, necessitating an application of methods such as Bonferroni correction (Dunn 1961) or FDR (Benjamini and Hochberg 1995). However, a Monte-Carlo simulation demonstrates that given a random undirected graph of size 18 with an edge probability of 0.2, the probability for having at least 6 different cliques with at least 4 members is approximately 0.00021. Hence the high statistical significance of our results.

Table 8.4 Comparison between different Arad ostraca.

	Ostraca Content	1	2	3	5	7	8	16	17a	17b	18	21	24	31	38	39a	39b	40	111
1	Order to Eliashib, supply of provisions for the Kittiyim		0.64	0.50	0.91	0.30	0.64	0.51	0.98	0.78	0.53	0.24	0.003	0.10	0.27	0.41	0.06	0.23	0.79
2	Order to Eliashib, supply of provisions for the Kittiyim	0.64		1.00	1.00	0.72	1.00	0.39	0.85	0.78	0.31	0.75	0.79	0.06	0.38	0.98	0.70	0.11	0.96
3	Order to Eliashib mentioning Hananyahu, concerning provisions to Beer Sheba	0.50	1.00		0.23	0.06	0.55	0.36	1.00	0.77	0.27	0.94	0.72	0.16	0.61	0.96	0.84	0.22	0.79
5	Order to Eliashib, supply of provisions, probably for the Kittiyim	0.91	1.00	0.23		0.53	0.60	0.60	0.19	0.40	0.07	0.46	0.12	0.01	0.40	0.24	0.21	0.07	0.98
7	Order to Eliashib, supply of provisions for the Kittiyim	0.30	0.72	0.06	0.53		0.03	0.76	0.17	0.48	0.004	0.43	0.05	0.07	0.27	0.35	1.00	0.15	0.05
8	Order to Eliashib, supply of provisions for the Kittiyim	0.64	1.00	0.55	0.60	0.03		0.68	0.07	1.00	0.17	0.33	0.74	0.42	0.20	0.67	1.00	1.00	0.93
16	Letter to Eliashib from Hananyahu	0.51	0.39	0.36	0.60	0.76	0.68		0.33	1.00	0.03	0.80	0.13	0.38	0.38	0.41	0.40	0.72	0.68
17a	Order to Nahum to proceed to the house of Eliashib in order to collect provisions	0.98	0.85	1.00	0.19	0.17	0.07	0.33		1.00	0.92	0.36	0.13	0.41	1.00	0.68	1.00	0.17	0.68
17b	Note that Nahum provided provisions to the Kittiyim	0.78	0.78	0.77	0.40	0.48	1.00	1.00	1.00		1.00	0.35	0.40	0.47	1.00	1.00	0.33	0.20	0.40
18	Report to Eliashib from a subordinate fulfilling an order; mention of the Temple	0.53	0.31	0.27	0.07	0.004	0.17	0.03	0.92	1.00		3×10 ⁻⁴	0.02	0.20	0.32	0.94	0.86	0.04	0.73
21	Letter to Gedalyahu from a subordinate, Yehokal	0.24	0.75	0.94	0.46	0.43	0.33	0.80	0.36	0.35	3×10 ⁻⁴		0.35	0.04	0.23	0.71	0.21	0.31	0.90
24	A royal decree ordering the reinforcement of Ramat Negeb against Edom	0.003	0.79	0.72	0.12	0.05	0.74	0.13	0.13	0.40	0.02	0.35		0.01	0.05	0.73	0.38	0.002	0.92
31	List of names	0.10	0.06	0.16	0.01	0.07	0.42	0.38	0.41	0.47	0.20	0.04	0.01		0.33	0.16	0.11	0.35	0.57
38	List of names (inc. the son of Eliashib)	0.27	0.38	0.61	0.40	0.27	0.20	0.38	1.00	1.00	0.32	0.23	0.05	0.33		0.77	0.33	0.70	0.77
39a	List of names	0.41	0.98	0.96	0.24	0.35	0.67	0.41	0.68	1.00	0.94	0.71	0.73	0.16	0.77		1.00	0.04	0.75
39b	List of names	0.06	0.70	0.84	0.21	1.00	1.00	0.40	1.00	0.33	0.86	0.21	0.38	0.11	0.33	1.00		0.42	0.42
40	Gemaryahu & Nehemyahu report to Malkiyahu mentioning Edom and the king of Judah	0.23	0.11	0.22	0.07	0.15	1.00	0.72	0.17	0.20	0.04	0.31	0.002	0.35	0.70	0.04	0.42		0.67
111	Fragmentary, mentioning guard and horses	0.79	0.96	0.79	0.98	0.05	0.93	0.68	0.68	0.40	0.73	0.90	0.92	0.57	0.77	0.75	0.42	0.67	

Moreover, contextual considerations can raise the number of distinct writers up to at least six. Among these, the different authors of the prosaic lists of names in Ostraca 31 and 39¹ were most likely located at the tiny fort of Arad– as opposed to Ostraca 7,

¹ Contrary to the excavator's association of Ostraca 31 and 39 with Stratum VII (Aharoni 1981, also Herzog 2002) rather than VI where most of the examined ostraca were found, we agree with critics

18, 24, and 40, which were probably dispatched from other locations². As per the table, Ostracon 31 differs from *both* sides of Ostracon 39; we can thus conjecture an existence of two additional authors, totaling *at least six distinct writers*. Since a presence of two professional scribes in such a tiny fort is implausible, this implies the composition of Ostraca 31 and 39 by authors who were not professional scribes. For the full implications of our results, see below.

8.6 Discussion

Identifying the military ranks of the authors can provide information regarding the spread of literacy within the Judahite army. Our proposed reconstruction of the hierarchical relations between the signees and the addressees of the examined inscriptions is as follows³ (see Fig. 8.6):

(Mazar and Netzer 1986; Ussishkin 1988) that these strata are in fact one and the same. Note that Ostracon 31 was found in locus 779, alongside three seals of Eliashib (the addressee of Ostraca 1-16 and 18, from Strata VI).

² Ostraca 5, 7, 17a, 18 and 24 were most probably written in other locations (Aharoni 1981). Ostracon 40 may have been written by troop commanders Gemaryahu and Nehemyahu (see the following note) with some ties to Arad fortress; their names also appear at Ostracon 31. This renders the common authorship of Ostraca 31 and 40 unlikely. Furthermore, from Table 8.4, Ostraca 40 and 39a have different authors.

³ We conjecture that the status of the officers who commanded the supplies to the *Kittiyim* (the Greek or Cypriot mercenary unit), who wrote Ostraca 1-8 and 17a, was similar to that of Malkiyahu (the commander of the fortress at Arad), and in any case they were Eliashib's superiors. Also note that Gemaryahu and Nehemyahu (Ostracon 40) are Malkiyahu's subordinates, while Hananyahu (author of Ostracon 16, also mentioned in ostracon 3), is probably Eliashib's counterpart in Beer Sheba. The textual content of the ostraca also suggests differentiation between combatant and logistics-oriented officials (Fig. 8.6).

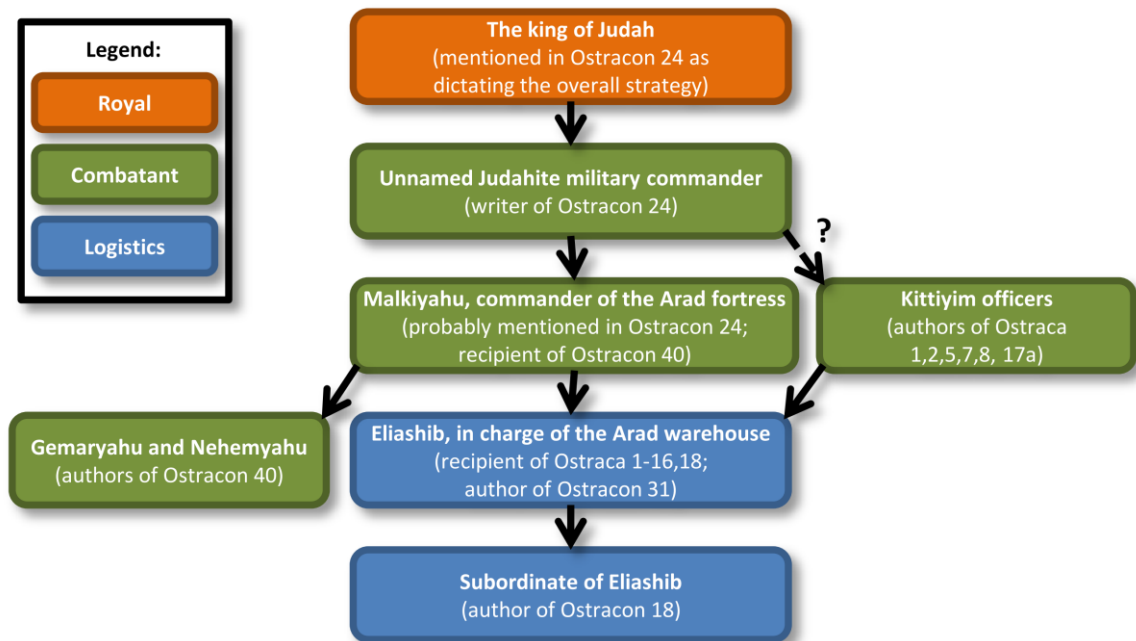


Figure 8.6 Reconstruction of the hierarchical relations between authors and recipients in the examined Arad inscriptions; also indicated is the differentiation between combatant and logistics officials.

1. **The King of Judah:** mentioned in Ostracon 24 as dictating the overall military strategy
2. **An unnamed military commander:** the author of Ostracon 24
3. **Malkiyahu, the commander of the Arad fortress:** mentioned in Ostracon 24 and the recipient of Ostracon 40⁴
4. **Eliashib, the quartermaster of the Arad fortress:** the addressee of Ostraca 1-16 and 18; mentioned in Ostracon 17a; the writer of Ostracon 31
5. **Eliashib's subordinate:** addressing Eliashib as "my lord" in Ostracon 18

Following this reconstruction, it is reasonable to deduce the proliferation of literacy among the Judahite army ranks ca. 600 BCE. A contending claim that the

⁴ Contrary to the excavator's dating of Ostracon 40 to Stratum VIII of the late 8th century (Aharoni 1981, also Ussishkin 1988), it should probably be placed a century later, along with Ostracon 24 (see Na'aman 2003 for details). Note that a conflict between the vassal kingdoms of Judah and Edom, seemingly hinted at in this inscription, is unlikely under the strong rule of the Assyrian empire in the region (ca. 730-630 BCE), especially along the vitally important Arabian trade routes.

ostraca were written by professional scribes can be dismissed with two arguments: First, the existence of two distinct writers in the tiny fortress of Arad (authors of Ostraca 31 and 39); second, the textual content of the inscriptions: Ostrakon 1 orders the recipient (Eliashib) “*write the name of the day*”; Ostrakon 7 commands “*and write it before you...*”; and in Ostrakon 40 (reconstructions Aharoni 1981; Na’aman 2003), the author mentions that he had written the letter. Thus, rather than implying the existence of scribes accompanying every Judahite official, the written evidence suggests a high degree of literacy in the entire Judahite chain of command.

The dissemination of writing within the Judahite army around 600 BCE is also confirmed by the existence of other military-related corpora of ostraca, at Horvat ‘Uza (Beit-Arieh 2007) and Tel Malḥata (Beit-Arieh and Freud 2015) in the vicinity of Arad, and at Lachish⁵ in the Shephelah (summary in Ahituv 2008) – all located on the borders of Judah (Fig. 8.1). We assume that in all these locations the situation was similar to Arad, with even the most mundane orders written down occasionally. In other words, the entire army apparatus, from high-ranking officials to humble vice-quartermasters of small, far from the center desert outposts, was literate, in the sense of the ability to communicate in writing.

In order to support this bureaucratic apparatus, an appropriate educational system must have existed in Judah at the end of the First Temple period (Lemaire 1981; Rollston 1999, 2006, 2010). Additional evidence supporting writing awareness by the lowest echelons of society seems to come from the Mezad Hashavyahu ostrakon (Naveh

⁵ In fact, Lachish Ostrakon 3, also containing military correspondence, represents the most unambiguous evidence of a writing officer. The author seems offended by a suggestion that he is assisted by a scribe. See detail, including discussion regarding the literacy of army personnel in (2).

1960) – which contains a complaint by a corvée worker against one of his overseers (most scholars agree that it was composed with the aid of a scribe).

Extrapolating the minimum of six authors in 16 Arad ostraca to the entire Arad corpus, to the whole military system in the southern Judahite frontier, to military posts in other sectors of the kingdom, to central administration towns such as Lachish, and to the capital Jerusalem, a significant number of literate individuals can be assumed to have lived in Judah ca. 600 BCE.

The spread of literacy in late-monarchic Judah provides a possible stage-setting for the compilation of literary works. True, biblical texts could have been written by a few and kept in seclusion in the Jerusalem Temple, and the illiterate populace could have been informed about them in public readings and verbal messages by these few (e.g., 2 Kings 23:2, referring to the period discussed here). However, widespread literacy offers a better background for the composition of ambitious works such as the Book of Deuteronomy and the history of Ancient Israel in the Books of Joshua-to-Kings (known as the Deuteronomistic History), which formed the platform for Judahite ideology and theology (e.g., Na'aman 2002). Ideally, in order to deduce from literacy on the composition of literary (to differ from mundane) texts, we should have conducted comparative research on the centuries after the destruction of Jerusalem, a period when other biblical texts were written in both Jerusalem and Babylonia according to current textual research (e.g., Schmid 2012; Albertz 2003). Yet, in the Babylonian, Persian, and early Hellenistic periods, Jerusalem and the southern highlands show almost no evidence in the form of Hebrew inscriptions. In fact, not a single securely-dated Hebrew inscription has been found in this territory for the period between 586 and ca. 350 BCE⁶

⁶ A few coins with Hebrew characters do appear between ca. 350 and 200 BCE.

– not an ostrakon, nor a seal, nor a seal impression nor a bulla (the little that we know of this period is in Aramaic, the script of the newly-present Persian empire; see Lipschits and Vanderhooft 2011). This should come as no surprise, as the destruction of Judah brought about the collapse of the kingdom's bureaucracy and deportation of many of the literati. Still, for the centuries between ca. 600 and 200 BCE, the tension between current biblical exegesis (arguing for massive composition of texts) and the negative archaeological evidence remains unresolved.

9. Writers' Identification via Binary Pixel Patterns and Kolmogorov-Smirnov Test

9.1 Introduction

In this research, we advance the ideas of the previous section to the next level. The writer identification analysis is performed independently, not only on a level of a single letter, but also on the level of a single feature, unleashing the full statistical power of multiple experiments. The main changes are: an entirely different, and much larger set of features (using 512 different binary pixel patterns instead of a combination of 7 features); a two-step experimental process, working on both individual feature (by comparing the feature distributions via Kolmogorov-Smirnov test), as well as individual letter level in order to deduce the p-values, later to be combined via Fisher's method (potentially, thousands of experiments, equaling the number of letters multiplied by the number of features, are conducted!); and an improvement in the significance level of the results by lowering the p-value threshold. All these allow us to establish a robust platform for analyzing corpora of many inscriptions, with an unknown number of authors, while arriving at meaningful and statistically highly significant outcomes. This approach was first suggested in (Shaus and Turkel 2017a).

A schematic comparison of the various handwriting analysis schemes is presented in Fig. 9.1.

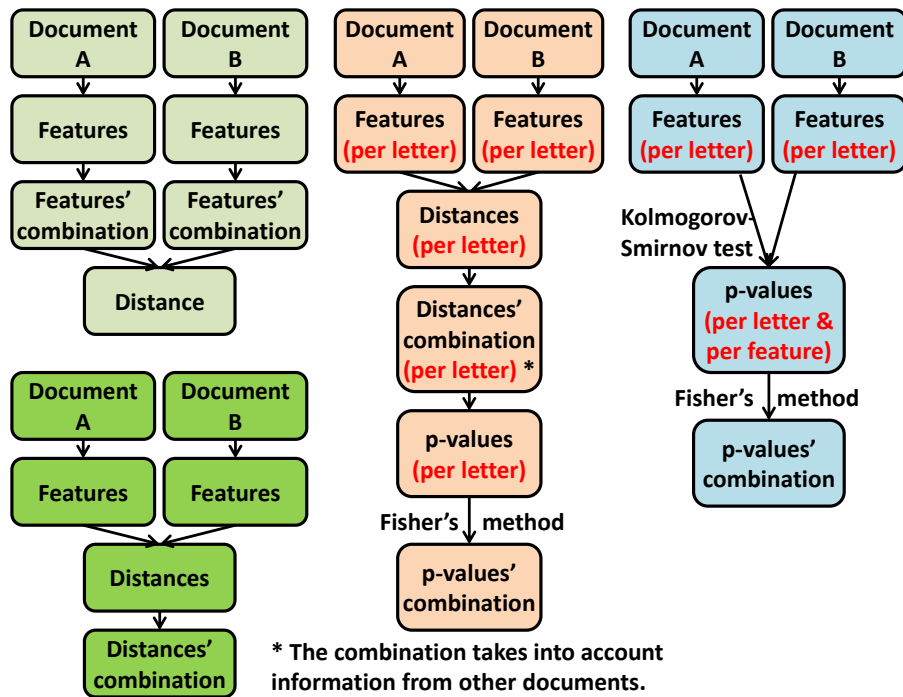


Figure 9.1 A comparison of handwriting analysis schemes. Left: common frameworks, producing an abstract distance between the documents as a final output. Center: the method of Section 8, performing the analysis on per-letter basis, yielding (number of letters) experimental p-values to be combined via Fisher’s method. Right: the current technique, performing Kolmogorov-Smirnov tests for each feature and each letter, yielding (num. of features) x (num. of letters) experimental p-values to be combined via Fisher’s method.

9.2 Algorithm’s Description

Preliminary Remarks

As in Section 8, we use the common statistical convention of defining a “null hypothesis” H_0 and trying to *disprove* it. In our case, H_0 is “two given inscriptions were written by the same author”. The probability for this event is the p-value, which will be estimated via the algorithm. If the p-value is *lower* than a pre-defined threshold, H_0 is rejected, and the competing hypothesis of “two different authors” is declared valid. On the other hand, an inability to reject the null hypothesis does not indicate its

validity. In such a case we remain agnostic, not being able to say anything regarding the documents' authors.

The estimation of the p-value involves an activation of the Kolmogorov-Smirnov (KS) test, a classical nonparametric test, allowing for a comparison of two samples, not necessarily of the same size (Corder and Foreman 2014). The main idea of KS is a comparison of *the empirical* distribution functions F_1 and F_2 (produced from the two samples), in order to calculate the observed statistic $D = \sup_x |F_1(x) - F_2(x)|$. The p-value of this statistic, under the hypothesis that the two samples stem from the same distribution, can be either calculated directly (via permutations) or approximated (our research utilizes the SciPy 2001 implementation). For example, if the samples' sizes are large enough, and all the values within the first sample are smaller than the values of the second sample, the p-value should be low. A previous usage of Kolmogorov-Smirnov test in a signature verification setting can be seen in (Griechisch et al. 2014).

Another well-established technique used by the algorithm is Fisher's method for p-values combinations (Fisher 1925; see similar utilization in Section 8). It allows for a calculation of a single combined ("meta") p-value, representing all the experimental evidence at our disposal. However, in the current section, the p-values of multiple experiments (stemming from different characters and features) are not necessarily independent (as assumed by Fisher's method), but are expected to be positively correlated. Thus, we're "over-confident" in the combined evidence against H_0 . A widespread remedy to this problem is to demand more significant results, by substituting T with $T \cdot (k+1) / 2k$ (k is the number of experiments) - a common modification representing a mean of false discovery rates (Benjamini and Hochberg

1995). In our case, this demand can be satisfied simply by lowering the threshold p-value τ from 0.2 (as in Section 8) to 0.1 or even 0.05.

Prior Assumptions

We begin with two images of different inscriptions, denoted as I and J . The algorithm operates based on information derived at a character level. Herein, by a character we denote a particular instance of a given letter (e.g. there may be many characters, which are all instances of a letter *alep*). As remarked above, we assume that the inscriptions' characters are binarized and segmented into images $I_{i_l}^l$ ($i_l = 1, \dots, M_l$, representing the instances of the letter l within I); and $J_{j_l}^l$ ($j_l = 1, \dots, N_l$, representing the instances of the same letter l within J), belonging to appropriate letters ($l = 1, \dots, L$). In the current research, the binarization and segmentation was performed automatically for Modern Hebrew, and in semi-manual fashion for Ancient Hebrew documents (Sober and Levin 2017). The resulting characters' images were padded with a 1-pixel white border on each side.






















Histogram Creation for each Character

Our features are the 3x3 binary pixel patterns, i.e. image patches of the individual characters. For additional information on pixel patterns, see the examples in (Akiyama et al. 1998; Ratnakar 1998); such patterns are a close, though less popular “relatives” of the local binary patterns (see Ojala et al. 1996; Ahonen et al. 2006; and notably Nicolaou et al. 2014 in Optical Font Recognition setting). There are $2^9=512$ optional patches of the requested size. All such possible patches are extracted from the images $I_{i_l}^l$ and $J_{j_l}^l$, in order to create normalized patches' histograms (counting

frequencies of patches' occurrences), $H_i^l(p)$ and $G_{j_i}^l(p)$, respectively (with $p = 1, \dots, 512$).

A simple, yet illustrative, example of two such images and their respective histograms is seen in Table 9.1. Remarkably, despite a similar overall shape of the character and only two pixels' difference in the character images, 16 out of 19 meaningful histogram entities are different.

Table 9.1 Example of character histograms.

Patches	Characters' images, patches' counts and frequencies ^a			
				
	1	0.083	1	0.083
	1	0.083	0	0
	0	0	1	0.083
	1	0.083	0	0
	2	0.167	2	0.167
	0	0	1	0.083
	1	0.083	0	0
	1	0.083	0	0
	1	0.083	0	0
	0	0	1	0.083
	0	0	1	0.083
	0	0	1	0.083
	1	0.083	0	0
	1	0.083	0	0
	1	0.083	0	0
	0	0	1	0.083
	0	0	1	0.083
	1	0.083	1	0.083
	0	0	1	0.083

a. Only the meaningful histogram entries are provided. In both cases, the remaining entries contain zeros. In **red** – discrepancies between the two histograms.

We stress that the histograms only serve normalization purposes. In the following, the histograms themselves will **not** be compared. Instead, the comparison will take place on an individual feature (patch) level, across different characters.

Same Writer Statistics Derivation

The experiments are performed in the following fashion: for given inscriptions' images I and J with $I \neq J$:

1. An empty *PVALS* array, to be utilized on a later stage, is initialized.
2. For each letter $l=1,\dots,L$, with sufficient character instances present ($M_l > 0$, $N_l > 0$, $M_l + N_l \geq 4$; we verify there is enough statistics for a meaningful comparison, slightly lowering the requirements in Section 8):

- 2.1. For each patch $p=1,\dots,512$, with at least one nonzero term present in the histogram (i.e. $\exists i_l.H_{i_l}^l(p) > 0$ OR $\exists j_l.G_{j_l}^l(p) > 0$), perform a Kolmogorov-Smirnov (KS) nonparametric test between the two samples $\{H_{i_l}^l(p)\}_{i_l=1}^{M_l}$ and

$$\{G_{j_l}^l(p)\}_{j_l=1}^{N_l} : pval_p^l = KS\left(\{H_{i_l}^l(p)\}_{i_l=1}^{M_l}, \{G_{j_l}^l(p)\}_{j_l=1}^{N_l}\right).$$

- 2.2. Append the resulting $pval_p^l$ to the *PVALS* array.

3. If the *PVALS* array is empty (i.e. no experiments were performed due to the scarcity of data), OR if $I = J$, set: $SameWriterP(I, J) = SameWriterP(J, I) = 1$.
4. Otherwise utilize the Fisher combination of all the *PVALS* instances, and set: $SameWriterP(I, J) = SameWriterP(J, I) = FisherMethod(PVALS)$

$SameWriterP(I, J)$ represents the deduced probability of having the same writer in both I and J (the H_0 hypothesis).

A toy-problem illustration of the whole scheme is shown in Fig. 9.2. In this demonstration, two *alep* and four *bet* letters are segmented from the first document, while three *alep* and two *bet* letters are segmented from the second document. As a first step, patches histograms are extracted from the two documents. For illustration purposes, it is assumed that in both cases, only the first two patches yield a non-zero count. Since two types of relevant features and two different letters are involved, $2 \times 2 = 4$ Kolmogorov-Smirnov tests are performed, yielding four p-values. These are combined into a single p-value via Fisher's method.

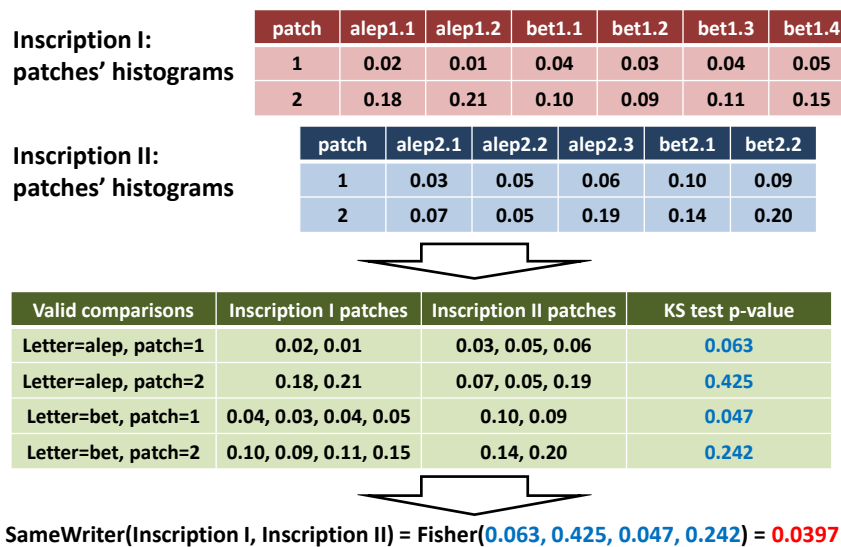


Figure 9.2 A toy example of the same writer statistics derivation for two hypothetical inscriptions. Inscription I consists of two instances of the letter *alep*, and four instances of the letter *bet*, while Inscription II consists of three instances of the letter *alep*, and two instances of the letter *bet*. The only patches with enough statistics are patches numbers 1 and 2. Four comparisons of appropriate samples (for each letter and each patch) are performed via Kolmogorov-Smirnov test, yielding four different p-values. These p-values are then combined via Fisher's method.

9.3 Modern Hebrew Experiment

The Basic Settings

This experiment closely follows the setting described in Section 8. The data-set (Modern Hebrew 2016) contains a sampling of 18 individuals. Each individual person filled in a Modern Hebrew alphabet table consisting of ten occurrences of each letter, out of the 22 letters in the alphabet (the number of letters and their names are the same as in the Ancient Hebrew in the next experiment; see Fig. 8.4 for a table example). These tables were scanned and thresholded in order to create black and white images. Then their characters were segmented utilizing their known bounding box location.

From this raw data, a series of “simulated” inscriptions were created. Due to the need to test both same-writer and different-writer scenarios, the data for each writer was split. Furthermore, in order to imitate a common situation in the Ancient Hebrew experiment, where the scarcity of data is prevalent (see below), each simulated inscription used only 3 letters (i.e., 15 characters; 5 characters for each letter), presenting a welcomed challenge for the new algorithm.

In total, 252 inscriptions were “simulated” in the following manner:

- All the letters of the alphabet except for *yod* (due to its small size), were split randomly into 7 groups (3 letters in each group), $g = 1, \dots, 7$: *gimel, het, resh*; *bet, samek, shin*; *dalet, zayin, ayin*; *tet, lamed, mem*; *nun, sade, taw*; *he, pe, qop*; *alep, waw, kap*.
- For each writer $k = 1, \dots, 18$, and each letter belonging to group g , 5 characters were assigned into simulated inscription $S_{i,g,1}$, with the rest assigned to $S_{i,g,2}$.

In this fashion, for constant k and g , we can test if our algorithm arrives at wrong rejection for $S_{i,g,1}$ and $S_{i,g,2}$ (FP = “False Positive” error; 18 writers and 7 groups producing 126 tests in total). Additionally, for constant g , writer q s.t. $q \neq k$, and $b, c \in \{1, 2\}$, we can test if our algorithm fails to correctly reject the “same writer” hypothesis for $S_{k,g,b}$ and $S_{q,g,c}$ (FN = “False Negative” error; 4284 tests in total).

Parameter Tuning and Robustness Verification

The algorithm described in the Algorithm’s Description sub-section provides an estimated probability for the H_0 hypothesis (“the two given inscriptions were written by the same writer”). However, two important parameters remain undecided. The first important parameter is the typical area of each character in pixels, leading to the optimal (or at least acceptable) performance. The second crucial parameter is the p-value threshold T , set for the purpose of rejecting the H_0 .

As is common in statistics, lowering T can result in fewer FP errors, unfortunately increasing the likelihood for FN errors. Conversely, raising T might result in the opposite outcome. In order to minimize the FP and FN errors, a set of simulations was performed. The simulations measured the behavior of the sum FP+FN, with respect to the area of the character’s image (ranging from 200 to 50,000 pixels), and to the chosen value of T (attempting the value 0.2 chosen in Section 8, as well as the values 0.1 and 0.05, as explained above).

The results of these simulations are shown in Fig. 9.3. Taking into account the performance of the algorithm described in Section 8 (FP+FN \approx 0.043), all the tested thresholds and all the areas between 1,000 and 40,000 pixels yield a reasonable and comparable performance (FP+FN<0.05). Among these, the results are slightly better in

the range of 8,000-20,000 pixels, with $T = 0.1$. This wide range for acceptable areas indicates an excellent robustness of the current algorithm (though it would probably result in better outcomes if the character images were of similar resolution). Since the mean area of the original character images was 17367 pixels, well within the reasonable limits of our analysis, we have chosen the typical area of each character to be 17000 pixels.

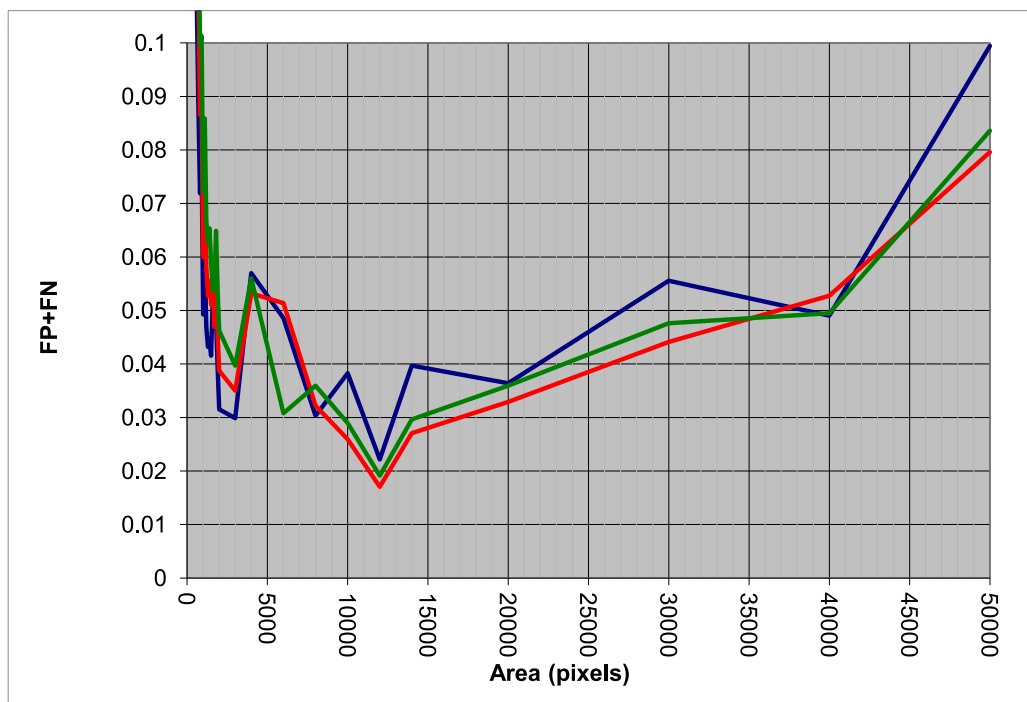


Figure 9.3 Testing the combined probability of FP+FN errors as a function of character area (in pixels) as well as different p-value thresholds: 0.2 in blue, 0.1 in red and 0.05 in green. Taking into account the performance of the algorithm in Section 8 ($FP+FN \approx 0.043$), all the tested thresholds and all the areas between 1000 and 40,000 pixels would yield reasonable and comparable performance. Slightly better results are achieved in the range of 8,000-20,000 pixels, with 0.1 threshold.

Experimental Results

The results of our configuration (for different values of T) are provided in Table 9.2. The results are certainly better than the results of Section 8 on the same data-set, with a much simpler configuration. As expected, FP error rate tends to zero as the

threshold is lowered, while the FN increases slightly. The threshold value of $T = 0.1$ produced better results, with a combined FP+FN error of less than 2%.

Confident in our newly obtained configuration (target area of $\sim 17,000$ pixels and $T = 0.1$), we proceed to the Ancient Hebrew experiment.

Table 9.2 Results of Modern Hebrew Experiment.

Configuration	Results		
	FP	FN	FP+FN
Results of Section 8, $T=0.2$	2.38%	1.94%	4.32%
Current setting, $T=0.2$	0.79%	1.70%	2.50%
Current setting, $T=0.1$	0.00	1.96%	1.96%
Current setting, $T=0.05$	0.00	2.12%	2.12%

9.4 Ancient Hebrew Experiment

The Basic Settings

As described in-length in the previous section, the data-set (Ancient Hebrew 2016) stems from the Judahite desert fortress of Arad, dated to the end of the First Temple period (Iron Age), ca. 600 BCE – the eve of Nebuchadnezzar’s destruction of Jerusalem. The fortress was unearthed half a century ago, with 100 ostraca (ink on clay) inscriptions found during the excavations (Aharoni 1981). The inscriptions represent the correspondence of the local military personnel. See Fig. 8.2 for examples of Arad ostraca.

Continuing a configuration of Section 8, we concentrate on 16 (relatively lengthy) Arad ostraca, two of them two-sided, which brings the total number of texts for analysis to 18. The scarcity of data in this situation is common for these ancient texts. Ostraca images were utilized in order to segment and restore the characters stroke-by-stroke via a variational procedure (detailed in Sober and Levin 2017) required

minimal human involvement. No further manipulation of the resulting characters' images (e.g. skeletonization, slant correction, etc.) was performed. Table 8.3 provides statistics of the most prominent letters, after reconstructing the legible characters. It can be seen, that even by the modest quantitative standards set in the current section, for some of the texts the comparisons are barely feasible.

Contrary to the situation in the modern context, now we do not possess any ground-truth, indicating the identity of the writers across different inscriptions. Moreover, the experiment's requirements do not impose a strict partition of the texts by their authors. The task is limited to detecting the minimal number of hands within this group of texts. Previous section demonstrated at least 4 different (pair-wise distinct) writers within the corpus (in fact 6 different "quadruplets" of texts), while bringing this number to 6 via textual considerations (not considered in the current research).

As already mentioned, following plausible results of the Modern Hebrew experiment, the characters were resized to 17,000 pixels, and the threshold was set to $T = 0.1$.

Experimental Results

The results of the experiment are presented in Table 9.3. The results indicate there are at least 5 different (pair-wise distinct) writers within this group of texts. In fact, a closer look at the table reveals that 3 such groups of 5 pair-wise distinct inscriptions exist, including the following texts: {1, 2, 18, 38, 40}, {1, 18, 24, 38, 40} and {5, 18, 24, 38, 40}. This can be contrasted with Section 8, where no such large pair-wise distinct groups were found, despite a higher threshold ($T = 0.2$). A simple simulation shows that given a random undirected graph of size 18 with an edge probability of 0.1, the probability for having at least 3 different cliques with at least 5

members is about 8×10^{-7} . Hence the high statistical significance of the results, substantially strengthening our confidence in the outcomes of Section 8.

9.5 Summary

The current research demonstrates a relatively simple and easily implementable algorithm for the purpose of writer identification. The algorithm demonstrates highly significant results in a setting including a minimal amount of letters. It is fast and robust with respect to both the typical area of the character images, and the evaluated p-value thresholds.

Our approach goes against the common wisdom of combining the different features or metrics before the documents' comparisons takes place. Instead, we propose to perform as many individual experiments as possible on both the letter and feature levels, combining their results only in the end. Although individual building blocks of our algorithm have occasionally been utilized in the literature, our specific amalgamation of binary pixel patterns, Kolmogorov-Smirnov test, and Fisher's method has not been described previously in the literature with regard to writer identification.

Table 9.3 Results of Ancient Hebrew Experiment, indicating separation of writers between texts (in red background).

Text	1	2	3	5	7	8	16	17a	17b	18	21	24	31	38	39a	39b	40	111
1	1	0.0001 04451	0.7940 37056	0.9972 86098	0.8355 55244	0.9999 86905	0.1451 23589	0.9999 94862	0.3456 85406	0.0127 74181	0.0125 35286	3.6592 5×10^{-12}	0.0160 73174	0.0781 82536	0.0387 94364	0.3086 94925	2.9710 3×10^{-06}	0.0001 19896
2	0.0001 04451	1	0.9999 99997	0.2308 10507	0.2310 3925	0.9998 37107	0.7828 82802	0.9993 77805	0.1210 18637	1.9698 9×10^{-08}	0.6776 87539	0.9070 18003	0.4606 83411	1.4592 1×10^{-07}	0.9999 99999	0.0047 77329	0.0117 57086	0.9185 37018
3	0.7940 37056	0.9999 99997	1	0.9999 9999	1	0.9999 99995	1	1	0.9999 17956	4.1385 8×10^{-07}	1	1	0.5188 2444	0.9608 18768	1	0.9999 99219	0.9831 16915	1
5	0.9972 86098	0.2308 10507	0.9999 9999	1	0.9999 99996	0.9999 99979	0.8296 55242	0.9552 54622	0.0937 82225	2.7843 6×10^{-17}	0.9968 39946	4.3890 7×10^{-10}	0.0261 92004	0.0267 97057	0.9429 02756	0.8672 49786	0.0408 87838	2.1339 $\times 10^{-05}$
7	0.8355 55244	0.2310 3925	1	0.9999 99996	1	0.9990 49432	0.9740 36343	0.9384 73936	0.5438 54905	6.1151 $\times 10^{-25}$	0.9532 18488	0.0144 8862	0.1514 6191	0.6513 69422	0.9732 78889	0.3812 86518	0.9996 02961	0.0720 03619
8	0.9999 86905	0.9998 37107	0.9999 99995	0.9999 99979	0.9990 49432	1	0.9968 67196	0.9999 99995	0.2648 1398	0.0001 60748	0.9999 9487	0.9956 23889	0.4905 01387	0.0289 88689	0.9999 33169	0.6108 34495	0.9999 79451	0.8184 31322
16	0.1451 23589	0.7828 82802	1	0.8296 55242	0.9740 36343	0.9968 67196	1	0.9143 22562	0.9895 23149	6.5381 9×10^{-37}	0.9999 9911	0.0501 44018	1.8251 3×10^{-08}	0.0175 41713	0.9999 99999	0.7529 72486	0.9861 54478	0.9999 98263
17a	0.9999 94862	0.9993 77805	1	0.9552 54622	0.9384 73936	0.9999 99995	0.9143 22562	1	0.9883 8179	0.9998 24883	0.9897 61454	0.9166 02116	0.9999 34002	0.9717 50814	0.9999 57681	0.9998 85898	0.7504 70353	0.9999 99995
17b	0.3456 85406	0.1210 18637	0.9999 17956	0.0937 82225	0.5438 54905	0.2648 1398	0.9895 23149	0.9883 8179	1	0.9894 32496	0.9873 54188	0.3970 38359	0.9979 36364	0.9946 74672	0.6903 49079	0.9894 49919	0.4693 71025	0.9763 81275
18	0.0127 74181	1.9698 9×10^{-08}	4.1385 8×10^{-07}	2.7843 6×10^{-17}	6.1151 $\times 10^{-25}$	0.0001 60748	6.5381 9×10^{-37}	0.9998 24883	0.9894 32496	1	1.7420 7×10^{-37}	2.8531 4×10^{-23}	0.8723 02919	6.4296 1×10^{-05}	0.0003 51619	3.9727 $\times 10^{-05}$	4.0665 6×10^{-14}	0.0162 51091
21	0.0125 35286	0.6776 87539	1	0.9968 39946	0.9532 18488	0.9999 9487	0.9999 9911	0.9897 61454	0.9873 54188	1.7420 7×10^{-37}	1	0.0049 05952	0.0027 46089	0.1227 37346	0.9999 99846	0.9825 66541	0.9578 07456	0.9998 09072
24	3.6592 5×10^{-12}	0.9070 18003	1	4.3890 7×10^{-10}	0.0144 8862	0.9956 23889	0.0501 44018	0.9166 02116	0.3970 38359	2.8531 4×10^{-23}	0.0049 05952	1	2.9342 2×10^{-09}	0.0092 50254	0.9999 99991	0.4051 06241	6.4106 $\times 10^{-12}$	0.6848 34031
31	0.0160 73174	0.4606 83411	0.5188 2444	0.0261 92004	0.1514 6191	0.4905 01387	1.8251 3×10^{-08}	0.9999 34002	0.9979 36364	0.8723 02919	0.0027 46089	2.9342 2×10^{-09}	1	0.8561 61514	0.9999 9999	0.9998 84872	1.7008 2×10^{-09}	0.7725 76882
38	0.0781 82536	1.4592 1×10^{-07}	0.9608 18768	0.0267 97057	0.6513 69422	0.0289 88689	0.0175 41713	0.9717 50814	0.9946 74672	6.4296 1×10^{-05}	0.1227 37346	0.0092 50254	0.8561 61514	1	0.6003 42228	0.1120 03203	0.0674 11667	0.0919 9445
39a	0.0387 94364	0.9999 99999	1	0.9429 02756	0.9732 78889	0.9999 33169	0.9999 99999	0.9999 57681	0.6903 49079	0.0003 51619	0.9999 99846	0.9999 99991	0.9999 9999	0.6003 42228	1	0.9999 85794	0.0464 50883	1
39b	0.3086 94925	0.0047 77329	0.9999 99219	0.8672 49786	0.3812 86518	0.6108 34495	0.7529 72486	0.9998 85898	0.9894 49919	3.9727 $\times 10^{-05}$	0.9825 66541	0.4051 06241	0.9998 84872	0.1120 03203	0.9999 85794	1	0.8390 63568	0.9075 6289
40	2.9710 3×10^{-06}	0.0117 57086	0.9831 16915	0.0408 87838	0.9996 02961	0.9999 79451	0.9861 54478	0.7504 70353	0.4693 71025	4.0665 6×10^{-14}	0.9578 07456	6.4106 $\times 10^{-12}$	1.7008 2×10^{-09}	0.0674 11667	0.0464 50883	0.8390 63568	1	0.1134 47052
111	0.0001 19896	0.9185 37018	1	2.1339 $\times 10^{-05}$	0.0720 03619	0.8184 31322	0.9999 98263	0.9999 99995	0.9763 81275	0.0162 51091	0.9998 09072	0.6848 34031	0.7725 76882	0.0919 9445	1	0.9075 6289	0.1134 47052	1

10. Segmentation via Morphologically-based Chan-Vese Framework

10.1 Introduction and Prior Art

Since its introduction at the beginning of this millennium, Chan-Vese (CV) segmentation (Chan and Vese 2001) has become one of the most widely used algorithms in the field of Computer Vision. In fact, currently, with more than 8,000 citations at Google Scholar, this method is almost twice as popular as the Mumford-Shah framework (Mumford and Shah 1989), upon which it is founded.

The power of CV technique lies within its ability to elegantly take into account the most important segmentation criteria. These include the length of the boundary curve between the segmented areas, the variance of gray-levels within each area, as well as the size of the “foreground” area. All these are handled within the scope of a single variational framework, leading to Euler-Lagrange equations, and thenceforth to numerical Gradient Descent PDE scheme. A straightforward extension of this theme to vector-valued (e.g. RGB) images (Chan et al. 2000, peculiarly published before Chan and Vese 2001), as well as a multi-phase level set framework (Vese and Chan 2002). These were proposed by the same authors, based on the same natural formulation.

Nonetheless, CV segmentation presents its own share of challenges. Among these are several “free” parameters of the algorithm (μ , ν , λ_1 , λ_2 , ε , h , Δt ; e.g., in experimental results of Chan and Vese 2001, μ ranges from 0.0000033×255^2 to 2×255^2 !), its initialization problem, as well as the intricate and sometimes computationally-intensive PDE scheme, based upon re-calculating the level set function on each step (an approach advanced by Osher and Sethian 1988). Although some of these hindrances might be handled by heuristic approaches (e.g. random re-

initializations, as proposed by Chan and Vese 2001), these are ad-hoc solutions, which add an overhead to the algorithm's implementation – with no guaranteed and sometimes difficult to forecast outcome.

Various approaches to these issues have been proposed. The method in (Xia et al. 2007) initializes using a modification of Canny edge detector (Canny 1986); (Solem et al. 2006) choose an initial level set via Gradient Descent over a thresholding criteria; (Pan et al. 2006) substitutes the level set formulation with curve evolution driven by Gaussian smoothing; (Wang et al. 2010; Liu and Peng 2012) replace the energy functional with different ones working on a local level; while (Brown et al. 2012) suggest another adjustment to the functional, possessing convexity properties.

We propose a new approach: a combination of an initialization based on Otsu's binarization method (Otsu 1979; it was proposed "heuristically" for CV initialization, yet not justified, in Xu and Wang 2008), supplemented by a morphological non-PDE energy minimization framework. Indeed, morphological methods have been suggested in the past for minimization of energy functionals pertaining to Computer Vision in general (Catté et al. 1995; Álvarez et al. 2010; Welk et al. 2011) and CV in particular. Among the latter are the techniques of (Jalba and Roerdink 2009), replacing the energy minimization with three compound morphological operators; (Anh et al. 2013), taking into consideration some pre-computed morphological data; (Fox et al. 2013b, 2013b), utilizing various structuring elements; (Oliveira et al. 2013), applying morphological filters *a-posteriori*; and (Kishore et al. 2015), adjusting CV energy functional by morphological gradient difference (MGD) term. The citation statistics of all these suggests such methods did not win wide acceptance, possibly due to their tendency to supplement one intricate solution with another.

The main contribution of the current section, first published in (Shaus and Turkel 2016), is an establishment of surprising relation between CV and Otsu's method, allowing for a simple initialization procedure. We also suggest a replacement of CV's PDE with a parameter-free morphological framework. The algorithm is to serve as an important building block in the next section.

10.2 The Chan-Vese algorithm

In their seminal paper, CV proposed the following segmentation energy functional:

$$F(c_1, c_2, C) = \mu \cdot \text{Length}(C) + \nu \cdot \text{Area}(\text{inside}(C)) + \lambda_1 \int_{\text{inside}(C)} |u_0(x, y) - c_1|^2 dx dy + \lambda_2 \int_{\text{outside}(C)} |u_0(x, y) - c_2|^2 dx dy, \quad (10.1)$$

where $u_0(x, y)$ is a given image; c_1, c_2 are constants; $C(s)$ is a parameterized curve partitioning the image domain Ω into disjoint $\text{inside}(C)$ and $\text{outside}(C)$ sets; while μ, ν, λ_1 and λ_2 are parameters. Eq. 10.1 is closely related to the energy functional of (Mumford and Shah 1989), which can be written as:

$$F^{MS}(u, C) = \mu \cdot \text{Length}(C) + \lambda \int_{\Omega} |u_0(x, y) - u(x, y)|^2 dx dy + \alpha \int_{\Omega \setminus C} |\nabla u(x, y)|^2 dx dy \quad (10.2)$$

where $u(x, y)$ is the estimated image, and α is a parameter. Assuming $\alpha \rightarrow \infty$, a piecewise-constant $u(x, y)$ is necessitated, eliminating the last term of F^{MS} . Assuming further that $u(x, y)$ has only two values, c_1 and c_2 , and adding the area term, we arrive at CV functional (Eq. 10.1). Using the level set formulation ϕ (zero on C , positive on

inside(C) and negative on *outside(C)*), the Heaviside function H and Dirac's function δ_0 :

$$H(z) = \begin{cases} 1 & 0 \leq z \\ 0 & z < 0 \end{cases}, \quad \delta_0(z) = \frac{d}{dz} H(z), \quad (10.3)$$

Eq. 10.1 can be reformulated in the following fashion:

$$\begin{aligned} F(c_1, c_2, \phi) = & \mu \int_{\Omega} \delta_0(\phi(x, y)) |\nabla \phi(x, y)| dx dy + \nu \int_{\Omega} H(\phi(x, y)) dx dy + \\ & + \lambda_1 \int_{\Omega} |u_0(x, y) - c_1|^2 H(\phi(x, y)) dx dy + \lambda_2 \int_{\Omega} |u_0(x, y) - c_2|^2 (1 - H(\phi(x, y))) dx dy. \end{aligned} \quad (10.4)$$

The first variation with respect to c_1 results in:

$$c_1(\phi) = \int_{\Omega} u_0(x, y) H(\phi(x, y)) dx dy / \int_{\Omega} H(\phi(x, y)) dx dy, \quad (10.5)$$

while the first variation with respect to c_2 yields:

$$c_2(\phi) = \int_{\Omega} u_0(x, y) (1 - H(\phi(x, y))) dx dy / \int_{\Omega} (1 - H(\phi(x, y))) dx dy. \quad (10.6)$$

On the other hand, the variation with respect to ϕ is less trivial. First, (Chan and Vese 2001) present an altered version of Eq. 10.4, introducing regularized H_ε and δ_ε functions:

$$\begin{aligned} F_\varepsilon(c_1, c_2, \phi) = & \mu \int_{\Omega} \delta_\varepsilon(\phi(x, y)) |\nabla \phi(x, y)| dx dy + \nu \int_{\Omega} H_\varepsilon(\phi(x, y)) dx dy + \\ & + \lambda_1 \int_{\Omega} |u_0(x, y) - c_1|^2 H_\varepsilon(\phi(x, y)) dx dy + \lambda_2 \int_{\Omega} |u_0(x, y) - c_2|^2 (1 - H_\varepsilon(\phi(x, y))) dx dy \end{aligned} \quad (10.7)$$

Next, an Euler–Lagrange equation for ϕ is derived and parameterized by an artificial time in the Gradient Descent direction:

$$\frac{\partial \phi}{\partial t} = \delta_\varepsilon(\phi) \left[\mu \cdot \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) - \nu - \lambda_1 (u_0 - c_1)^2 + \lambda_2 (u_0 - c_2)^2 \right]. \quad (10.8)$$

A numerical scheme for Eq. 10.8 is also suggested, for further details see (Chan and Vese 2001).

10.3 From Chan-Vese to Alternative Solution

We now analyze the CV (Chan and Vese 2001) algorithm, and suggest its restatement in alternative terms. In particular, we prefer not to use the level set framework. Similarly to CV, we strive to achieve a partition of the image domain Ω into two disjoint sets of pixels, denoted herein as A_1 and A_2 . Yet unlike CV, we have no prior assumptions and no limitations regarding their location within u_0 . An additional preference would be to avoid a *regularized* version of the algorithm, which tends to smooth some of the image features (cf. the criticism on Gaussian smoothing in Chan and Vese 2001).

Constants: CV noted that Eqs. 10.5, 10.6 represent the averages:

$$c_1(\varphi) = \text{average}(u_0) \text{ in } \{0 \leq \phi\}, \quad c_2(\varphi) = \text{average}(u_0) \text{ in } \{0 > \phi\}. \quad (10.9)$$

Our alternative (and symmetric) formulation retains the constants c_1 and c_1 , associated respectively with A_1 and A_2 , and calculates them in a similar fashion:

$$c_1(\varphi) = \text{average}(u_0) \text{ on } A_1, \quad c_2(\varphi) = \text{average}(u_0) \text{ on } A_2. \quad (10.10)$$

Localization: Eq. 10.8 defines the evolution of the level set, and subsequently the sets *inside*(C) and *outside*(C). We substitute this scheme with its morphological counterpart. We first consider the multiplicand $\delta_\varepsilon(\phi)$, where δ_ε is a regularization of

δ_0 . Since $\delta_0 \equiv 1$ at a zero-level $\{\phi = 0\}$, and $\delta_0 \equiv 0$ at $\{\phi \neq 0\}$, the term limits the evolution only to pixels belonging to C (optionally including their immediate neighbors for δ_ε). Agreeing with this strategy, we denote as “borderline” pixels the pixels of A_1 adjacent to at least one pixel in A_2 , or vice versa.

Curvature-driven evolution: We next consider the first term of the second multiplicand of Eq. 10.8, $\mu \cdot \text{div}(\nabla \phi / |\nabla \phi|) - \nu - \lambda_1 (u_0 - c_1)^2 + \lambda_2 (u_0 - c_2)^2$. As explained in (Vese and Chan 2002; Osher and Sethian 1988), $\kappa = \text{div}(\nabla \phi / |\nabla \phi|)$ is the curvature at zero level, which induces a minimization of the curve’s length. This theoretical construction may be supplemented by a low-level analysis. Assuming 4-connectivity (radius of 1 around the central pixel), and taking various symmetries into account, there exists only 5 possible neighborhoods of an A_1 borderline pixel (borderline pixels of A_2 admit similar analysis). These options are presented on Fig. 10.1. It can be seen, that given a non-negligible μ , only Figs. 10.1a and 10.1b necessitate a re-assignment of the center pixel to A_2 - in both of these cases the radius of the osculating circle is $r = 1$, hence $\kappa = (1/r) = 1$. Additionally, *ignoring the central pixel*, Figs. 10.1c and 10.1d present a symmetry between pixels assigned to A_1 and A_2 , thus no re-assignment is needed (otherwise an oscillatory behavior is expected), while Fig. 10.1e presents a case of clear A_1 majority. The morphological operator perfectly representing such pixel assignment is the *median filter*. While the presented analysis represents a radius of 1 around the central pixel, if some kind of regularization is desired, a different median filter radius can be chosen (cf. Catté et al. 1995; Fox et al. 2013a, 2013b, for the median filter in related contexts).

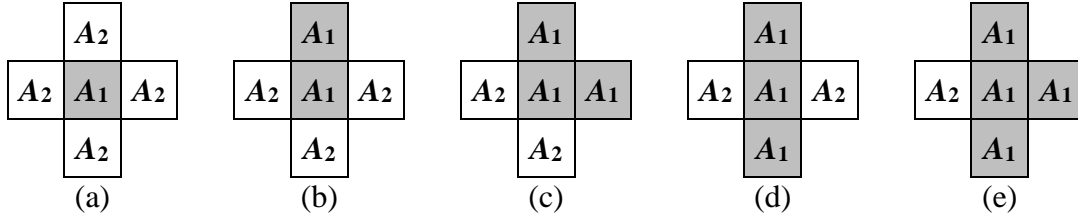


Figure 10.1 Five options of neighborhood of an A_1 borderline pixel. Only (a,b) require a re-assignment of the central pixel, due to a positive curvature.

Area-driven evolution: The next term to be analyzed within Eq. 10.8 is $-\nu$. If $0 < \nu$, this represents a constant reduction in the size of $inside(C)$, which is difficult to justify (unless a human operator fancies a specific result). If for some reason the initial sets $inside(C)$ and $outside(C)$ are switched, the dynamics is reversed, as $outside(C)$ is now expected to constantly grow, breaking the symmetry between the sets. Moreover, given images with small or zero curvature, and λ_1, λ_1 chosen to be small, the shrinking might continue until $inside(C)$ disappears completely! It seems that the dubious benefits of this term were understood by CV, since ν is mostly set to 0 in (Chan and Vese 2001), and the term is no longer mentioned in (Vese and Chan 2002). We also advise against using this term, but in case it is desired, its morphological substitution would be an erosion in case of $0 < \nu$, and dilation in case of $\nu < 0$, with A_1 or A_2 chosen as a target.

Fidelity-driven evolution: The last terms to be considered within Eq. 10.8 are $-\lambda_1(u_0 - c_1)^2 + \lambda_2(u_0 - c_2)^2$, presenting a balance between reducing the size of $inside(C)$ due to its gray-levels variance, and its enlargement due to the variance of gray-levels within $outside(C)$. Reversing our steps shows these terms originate from

$$\lambda_1 \int_{inside(C)} |u_0(x, y) - c_1|^2 dx dy + \lambda_2 \int_{outside(C)} |u_0(x, y) - c_2|^2 dx dy \quad \text{in Eq. 10.1, or}$$

$\lambda_1 \int_{A_1} |u_0(x, y) - c_1|^2 dx dy + \lambda_2 \int_{A_2} |u_0(x, y) - c_2|^2 dx dy$ in the current case. Using the

recommendation of $\lambda_1 = \lambda_2 = 1$ (Chan and Vese 2001), this has a surprising relation to

Otsu binarization method (Otsu 1979; also cf. Xu and Wang 2008). Otsu minimizes the

thresholding quality criterion $\omega_1 \sigma_1^2 + \omega_2 \sigma_2^2$, where $\sigma_1^2 = \sum_{i=1}^k (i - \mu_1)^2 \cdot p_i / \omega_1$;

$\sigma_2^2 = \sum_{i=k+1}^L (i - \mu_2)^2 \cdot p_i / \omega_2$; $\omega_1 = \sum_{i=1}^k p_i$; $\omega_2 = \sum_{i=k+1}^L p_i$ and p_i represents the value of the

gray-level $i \in [1, 2, \dots, L]$ within the normalized histogram of the image u_0 . This image

is thresholded by k and partitioned into disjoint sets A_1 and A_2 , respectively

containing gray-levels $[1, \dots, k]$ and $[k+1, \dots, L]$. Thus, translating Otsu's terms into

CV's terminology:

$$\omega_1 \sigma_1^2 + \omega_2 \sigma_2^2 = \int_{A_1} |u_0(x, y) - c_1|^2 dx dy + \int_{A_2} |u_0(x, y) - c_2|^2 dx dy. \quad (10.11)$$

Eq. 10.11 presents us with two opportunities. Firstly, it provides an excellent option for *initialization* of the algorithm, since Otsu's method efficiently handles the needed minimization of this energy functional, with only the curve length remaining to be optimized. Secondly, it offers an explanation of the inner machinery of the fidelity term. Indeed, if all the other terms are negligible, the fidelity term would "strive" to lower the energy until the minimum, corresponding to optimal Otsu's thresholding, is reached. Therefore, several options for fidelity-driven evolution strategies can be proposed:

1. The "original" rule: eroding A_1 if $-(u_0 - c_1)^2 + (u_0 - c_2)^2 < 0$ and dilating it if

$$-(u_0 - c_1)^2 + (u_0 - c_2)^2 > 0.$$

2. *The “Otsu-aware” rule*: At initialization, A_1 and A_2 are associated with their “optimal” partitioning (calculated only once). Even if changes in A_1 and A_2 occur due to other terms, it is still possible to immediately recognize the “misattributed” borderline pixels, which need to be re-assigned.
3. *The “no-rule” rule* (our preference): Since the initialization already used Otsu’s criterion in an optimal manner, it would be better to drop the fidelity from further consideration, allowing other factors to properly influence the calculations.

Proposed algorithm: Our recommendations are summarized in Table 10.1. Please note, that in the results below, a maximum, rather than Euclidean norm was utilized, for simplicity reasons. Thus, radius 1 neighborhood now includes 9, rather than 5 pixels.

Table 10.1 Description of the algorithm, including various options.

Step	Recommendation	Additional options
Initialization	Otsu’s method in order to partition u_0 into A_1 and A_2 .	
Evolution	Median filter with radius 1 on label (A_1 and A_2) map.	Median filters with other radii on label map, for regularization purposes.
	No area term ($\nu = 0$).	If desired, dilation/erosion of A_1 or A_2 (and vice versa).
	No fidelity term ($\lambda_1 = \lambda_2 = 0$).	<ul style="list-style-type: none"> • Dilation/erosion depending on fidelity term • Re-assigning “misattributed” Otsu borderline pixels
Stopping criterion	Convergence of A_1 and A_2 .	

10.4 Experimental Results

In the following experiments, a segmentation is demonstrated on non-trivial images, some of which resembling the ones used by (Chan and Vese 2001). Fig. 10.2 presents an object with a smooth contour, Fig. 10.3 shows satellite image of Europe

night-lights, Fig. 10.4 demonstrates a spiral art-work, while Figs. 10.5 and 10.6 represent noisy historical inscriptions. It can be observed that in general, the default or slightly regularized parameters produce high-quality segmentation, superior to Otsu with no curvature evolution. We omit comparisons with the CV algorithm, due to the high dependence of its results on the various parameters in use, as explained above.

10.5 Summary

A detailed analysis of the CV segmentation framework was presented. Among the main novelties of the article are the surprising relation between the Otsu binarization method and the fidelity terms of CV energy functional (which may explain the results of Brown et al. 2012, resembling binarization), allowing for intelligent initialization of the functional. This is accompanied by a suggestion of a very fast, parameter-free morphological framework, substituting the CV PDE-based segmentation method. The experimental results demonstrate the soundness of our approach, which will be utilized in the next section.

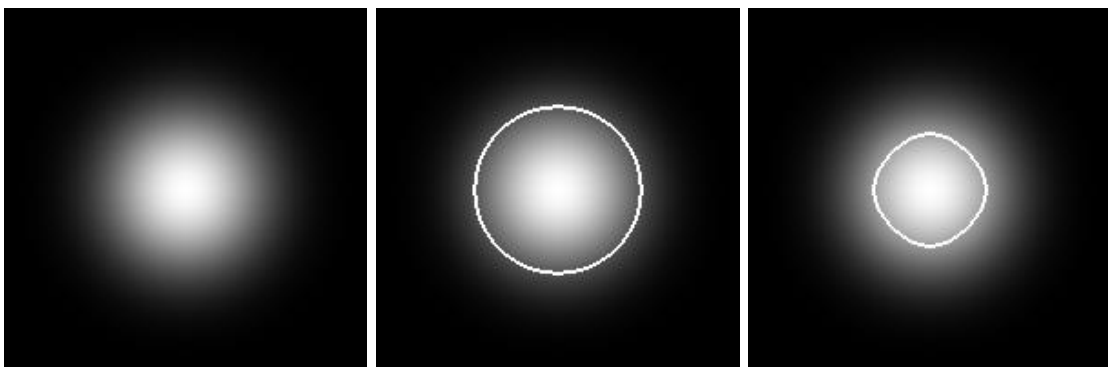


Figure 10.2 Segmentation of an object of smooth contour: original image (*left*), vs. result with default setting (*center*), vs. result with radius=11 (*right*).



Figure 10.3 Segmentation of a satellite image of Europe night-lights: original image (*left*), vs. Otsu binarization (*center*), vs. result with the default setting (*right*). Image courtesy NASA/Goddard Space Flight Center Scientific Visualization Studio, public domain.

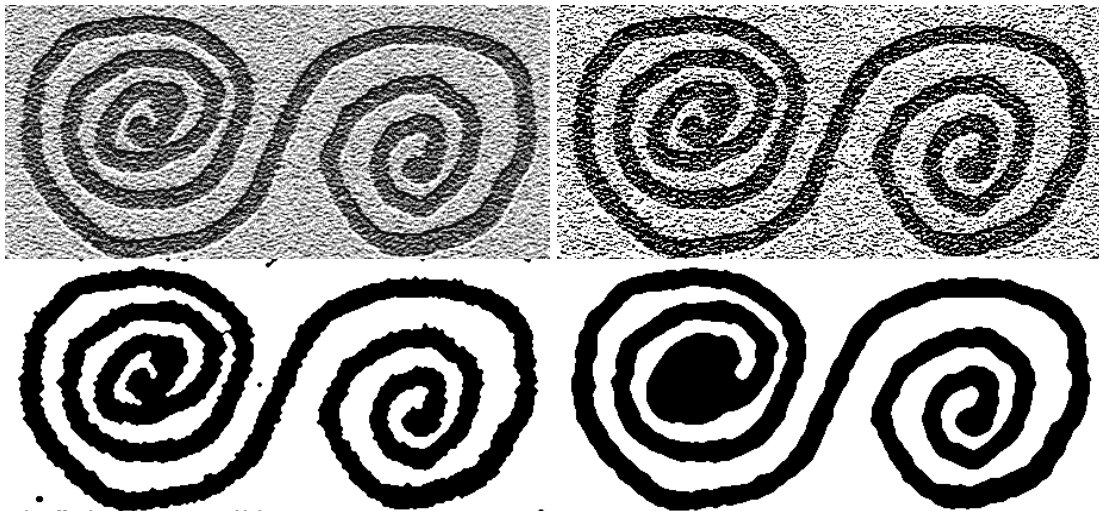


Figure 10.4 Segmentation of a spiral art-work: original image (*upper left*), vs. Otsu binarization (*upper right*), vs. result with the default setting (*lower left*), vs. result with radius=2 (*lower right*). Image courtesy José-Manuel Benito Álvarez, public domain.

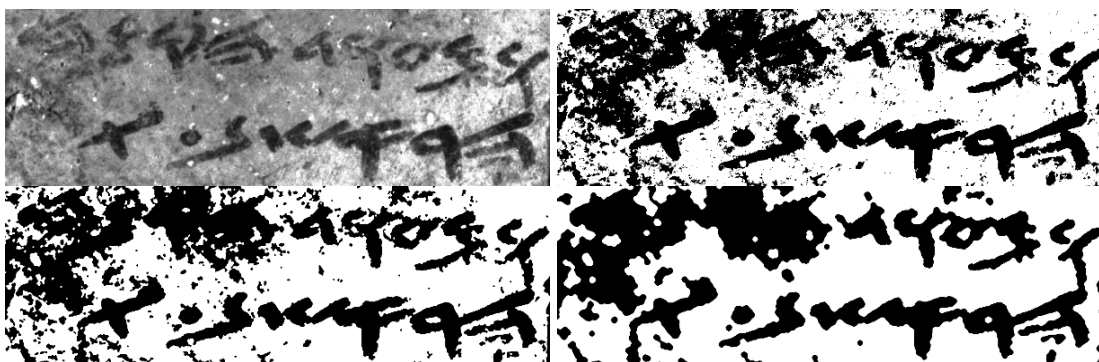


Figure 10.5 Segmentation of a fragment of Arad ostracon No. 1: original image (*upper left*), vs. Otsu binarization (*upper right*), vs. result with the default setting (*lower left*), vs. result with radius=2 (*lower right*).

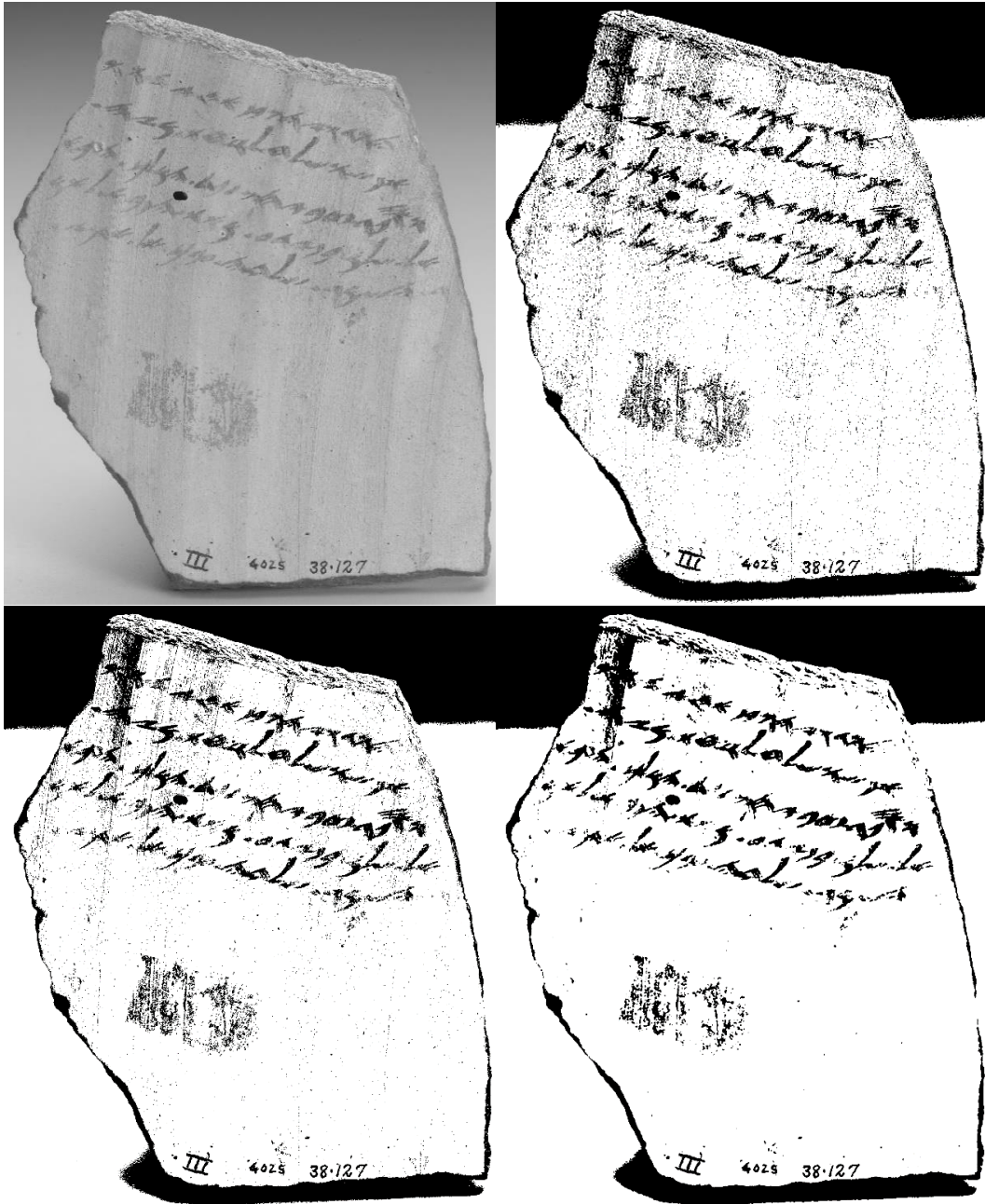


Figure 10.6 Segmentation of Lachish ostracon No. 3: original image (*upper left*), vs. Otsu binarization (*upper right*), vs. result with the default setting (*lower left*), vs. result with radius=3 (*lower right*).

11. Letter Shape Prior Estimation

11.1 Introduction

The problem of finding a prototype for typewritten or handwritten characters belongs to a broad type of “shape prior” determination problems, which has gathered substantial research interest during the last two decades. Despite that fact, articles deriving shape prior of handwritten or printed characters are relatively rare in both the Computer Vision (CV) and the OCR/Handwriting Recognition (HR) communities. The lack of interest of the CV scientists can be explained by the specificity of this potentially challenging problem. On the other hand, most of the HR studies focus on producing ever improving recognition engines – a related, yet not directly dependent problem. The relatively low interest in the subject resulted in diverse terms used by the existing publications. Among the related terms are “letter/handwriting prototypes”, “document-specific alphabet”, “reconstructed font”, “glyph extraction”, “character template estimation”, “character models”, “codebook generation”, “ideal/Platonic prototypes” and “letter shape priors”. In what follows, we shall use the last term, common in the CV community.

In the case of historical texts, the problem of deriving a character shape prior is closely related to the issue of creating the so called “paleographic tables” – a basic instrument in the toolbox of the historical epigrapher (an expert on ancient writings). Commonly, such tables contain one characteristic example of each letter type for each inscription in a given corpus; see example on Fig. 11.1. The tables are used in order to trace the similarities and the differences within the handwriting of different localities and time periods. This labor-intensive process joins other manually performed epigraphic tasks. Indeed, currently, the imaging, the creation of the facsimile (a black and white depiction of the inscription), the recognition of the letters, the transcription,

the creation of paleographic tables, as well as their analysis are all carried out manually by epigraphic experts. Such an effort is extremely time-consuming, producing results which may accidentally mix-up documentation with interpretation.

12	11	10	8	7	6	5	4	3	2	1
⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
										⌘
		⌘	⌘			⌘		⌘	⌘	⌘
⌘			⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
					⌘					
⌘	⌘			⌘	⌘	⌘	⌘	⌘	⌘	⌘
					⌘		⌘			
⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
⌘								⌘	⌘	
⌘				⌘		⌘		⌘	⌘	⌘
⌘			⌘	⌘		⌘		⌘	⌘	⌘
⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘

Figure 11.1 Manually created paleographic table, recording “typical” representatives for each letter in the alphabet.

The envisioned objective of our future research is an automatically derived paleographic table, accompanied by its algorithmic analysis. In this study, we will concentrate on a challenging intermediate goal of obtaining the main building block of such a table, i.e. the letters’ shape prior.

For consistency purposes, the following terminology is used throughout this section. By “letters” we designate the members of the alphabet, e.g. “*alep*”, “*bet*”, etc. Their realizations by the writer are the particular characters, e.g. an inscription may contain several “*bet*” characters. A “letter shape prior”, or in short “letter prior” represents a typical way of depicting a given letter, both empirically observed and estimated.

11.2 Prior Art

Ostensibly, the task of estimating the letter prior seems to be relatively straightforward, requiring a registration of the character images, their accumulation and subsequent thresholding. However, in reality, this undertaking turns out to be surprisingly difficult. Indeed, elastic image registration is an NP-complete problem (Keysers and Unger 2003). Moreover, multiple template alignment estimation was also shown to be NP-complete (Kopeck and Lomelin 1997). It may be that such difficulties have also limited (Riklin-Raviv et al. 2008) to a pair of images, in a variational shape-prior determination framework. Thus, the existing solutions of mutual registration problem are of heuristic nature, and tend to balance between the computational costs and the quality of the result.

The study (Kopeck and Lomelin 1997) proposed a sophisticated Aligned Template Estimation (ATE) framework, in which overlapping glyphs templates were searched in a page image. The authors used a two-phase iterative training algorithm, encompassing an alignment of pre-existing transcriptions given initial guess (existing transcriptions), as well as an ATE stage. The ATE step was implemented via a likelihood maximization procedure. The technique was designed for typewritten characters. Its results were reasonable given sufficiently large data and number of iterations. Nevertheless, some artifacts were present in the resulting “priors”, due to the method's “unawareness” of the different character properties, and inexact segmentation boundaries. The research (Bern and Goldberg 2000) proposed a variation on the theme of super-resolution within a single image, also in the context of printed text. Given a relatively clean binarized document image, the letters were registered, then iteratively clustered, taking phenomena such as touching letters into account. A Bayesian calculation yielded a prior, which was utilized for image de-noising purposes. The

results of this algorithm also exhibited certain artifacts, due to the exceedingly fine-grained clustering, and mistaking noisy characters for distinct glyphs.

For handwriting, several papers included prior estimation as an intermediate step in handwriting synthesis (i.e. a simulation of a particular handwriting style given a few writing examples). As opposed to the relatively fixed typewritten characters of previous works, now a more challenging cursive writing, with its high variance, was considered. As a relaxation, the inputs in these cases were clean and thinned writing examples. In (Devroye and McDougall 1995), after a segmentation achieved by a Hidden Markov Model, a curve control point interpolation was performed. The article (Wang et al. 2002) extracted priors in addition to a "tri-unit" technique (akin to the tri-grams of Speech Recognition). This was used in order to identify different types of "contact" strokes between various characters. The shape prior creation was composed of control point extraction (Gabor filters leading to B-splines approximation), affine registration and shape prior parameter estimation stages, with impressive results.

The paper (Edwards and Forsyth 2006) derived shape prior in the complicated world of historical documents (12th century manuscript). The authors initiated the priors with hand cut examples. The page image was then segmented into words and characters; each word possessed several possible segmentations (represented by a graph). For each word, the different possible segmentations were searched within a pre-existing dictionary (in the target language) by comparing the word image with the candidate word image derived from shape priors. The high confidence matches were accepted, and then the shape priors were updated. If necessary, new shape priors (possibly more than one for a single character) were created. The process was then repeated. A similar statistical language model was also utilized in (Cutter et al. 2010, 2011), where

candidate words are checked vs. an English corpus. Words (token) co-occurrence statistics was used in order to correctly identify problematic characters.

The issue of error vs. compression rate (i.e. the number of shape priors vs. their recognition accuracy) in the setting of historic documents was the topic of (Pletschacher 2008, 2009). It provided an empirical evidence for a relation between the optimal “safe threshold”, the maximum intra-cluster distance and the compaction ratio, with the “safe threshold” only permitting compaction of 20%.

A noteworthy modern variational approach in historic setting was presented in (Bar-Yosef 2008, 2009). Given a set of character edges, a confidence map (shape prior) was created for each character individually via a Gradient Vector Flow. Subsequently, the confidence map could be fitted back into the document image, utilizing the Active Contour method, in order to achieve high-quality segmentation. (Panagopoulos et al. 2009) utilized estimated “ideal” or “Platonic” prototypes for each letter of historical inscriptions for the purpose of writer identification analysis.

11.3 The Proposed Algorithm

This research, first published in (Shaus and Turkel 2017b), deals with ancient Hebrew ostraca. Many of the ostraca were not composed by professional scribes (see Sections 8 and 9), and therefore the variability of the handwriting is very high. The inscriptions are quite short (typically containing 30-100 characters), and their state of preservation is poor. These characteristics of the writing medium influenced the design of our algorithm. Contrary to prior art, only small amounts of characters for each type of letter are present for each ostrakon. Moreover, the inscriptions are highly degraded (with blurred and erased characters, as well as cracks and stains easily mistaken for

characters). Hence, upon implementing our algorithm, we preferred robust statistical estimators such as median and medoid (a representative object, whose dissimilarity to other objects in the population is minimal) over the commonly used mean, which is easily susceptible to noise (for another use of medoids and medians, see Section 5).

We assume grayscale images of the ostraca (e.g. acquired by methods described in Section 3, Faigenbaum et al. 2012). We also pre-suppose imperfect black and white facsimiles, registered to the grayscale ostraca images. The facsimiles are utilized for initial segmentation purposes, in a manner similar to that described in Section 4, i.e. the registered facsimiles provide us with initial indication regarding the position and the type of inscriptions' characters within the ostraca images. The algorithm utilizes the cropped (dilated and padded) character images; chooses a medoid image via simple registration procedure; registers all the other character images to the medoid image; calculates the initial prior via median calculation per each pixel coordinate; thresholds the prior via modification of Otsu's algorithm (Otsu 1979), and if desired, smoothes the result.

The detailed steps of our algorithm, for a given inscription and letter, are:

1. Cropping character images:

- 1.1. The characters' convex hulls (of width w_i and height h_i ($i = 1, \dots, K$)) are found at the facsimile level.
- 1.2. The convex hulls are dilated by $PAD \cdot \max\{w_i, h_i\}$ pixels (assuming 4-connectivity), with respect to a pre-determined parameter PAD (in our setting, $PAD = 0.1$, i.e. at least 10% addition on each side of the character – determined empirically).

1.3. The locations of the dilated convex hulls in the facsimile image are used in order to crop rectangular images $S_i(m,n):[1,M_i] \times [1,N_i] \rightarrow [0,255]$ of the characters from the grayscale ostraca images. Pixels corresponding to the dilated convex hulls assume the grayscale values of the inscription image, while other pixels assume the padding value of 255.

2. Padding character images:

2.1. The maximal dimensions of the character images are calculated:

$$M = \max \{M_i\}, N = \max \{N_i\}.$$

2.2. These dimensions are utilized in order to create padded character images of common size. The padding (by 255) is applied symmetrically on the opposite sides of S_i , resulting in padded images $P_i(m,n):[1,M] \times [1,N] \rightarrow [0,255]$, of the same size.

3. Initial characters' registration:

3.1. For each $i = 1, \dots, K$, and for each $j = 1, \dots, K$ s. t. $i \neq j$, a normalized cross-correlation fit (Pratt 1974) ρ_{ij} is calculated between P_i and S_j , keeping the shifts yielding the best fit for future use (on step 3.4).

3.2. The (not necessarily symmetrical) distances d_{ij} are calculated:

$$d_{ij} = \sqrt{(1 - \rho_{ij})/2} \text{ (see Van Dongen and Enright 2012 for further details).}$$

3.3. A medoid index $l = \arg \min_i \left(\sum_j (d_{ij}) \right)$ and an initial registered image $R_l = P_l$ are established.

3.4. For all $i = 1, \dots, K$, s. t. $i \neq l$, the $S_i(m,n)$ images are translated according to their optimal shift (calculated at stage 3.1) with respect to R_l , in order to obtain registered images R_i ; their padding value is 255.

4. Letter prior initialization:

The initial prior L_{mit} is calculated via median for each pixel coordinate, over all the registered character images: $L_{mit}(m, n) = \underset{i=1, \dots, K}{\text{median}}\{R_i(m, n)\} \therefore$

5. Letter prior thresholding:

A thresholded prior image L_{thr} is calculated via $L_{thr} = Otsu^*(L_{mit})$, where $Otsu^*$ is an adaptation of Otsu's algorithm (Otsu 1979) ignoring the histogram value of 255 (i.e. the padding values of steps 1.3, 2.2 and 3.4, which might skew the statistics).

6. Letter prior smoothing:

A smoothed prior image L_{sm} is calculated via $L_{sm} = MorphCV(L_{thr}, REG)$, where $MorphCV$ is a morphological solution to Chan-Vese (Chan and Vese 2001) segmentation framework, introduced and analyzed in Section 10, and REG is an optional regularization (smoothing) parameter, controlling the median filter radius within $MorphCV$.

7. Optional letter prior calculation loop:

The estimated prior L_{sm} can now be plugged-in at step 3.4, with all the s_i optimally fitted to L_{sm} instead of the medoid P_l . The resulting collection may then be refined (via the median, as in step 4), the outcome thresholded by $Otsu^*$ (as in step 5), and its result smoothed via $MorphCV$ (as in step 6). The loop can be either stopped at this stage, or repeated until convergence.

11.4 Results

Different configurations of our method were tested on the relatively large Arad 1, Arad 2 and Arad 24b (verso side) ostraca (Aharoni 1981). The 8-bit grayscale images

of the inscriptions were approximately of the same resolution, with a typical character size of 30,000-60,000 pixels (width and height varying depending on the character). Registered facsimiles, colored according to letter types, were also utilized; see Fig. 11.2-11.4 for images of ostraca and their facsimiles.

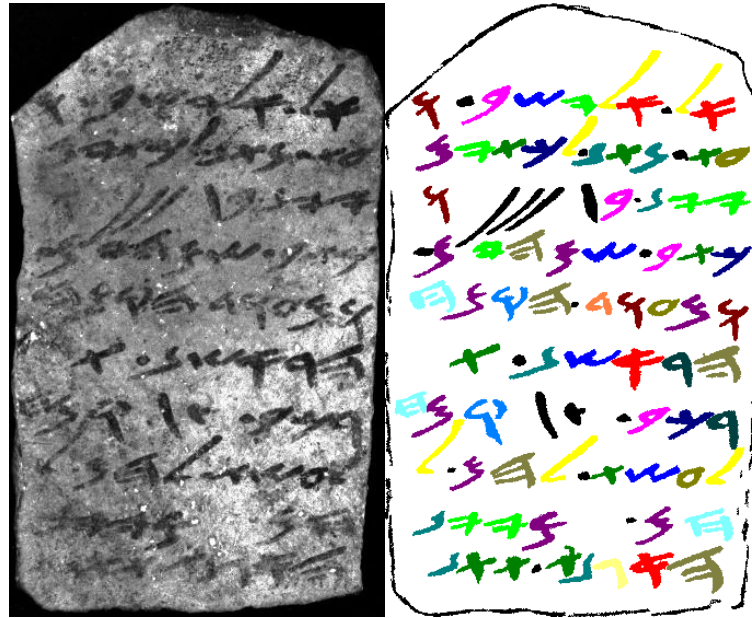


Figure 11.2 Arad 1 - an ostracon image (left) and its corresponding facsimile (right). The various colors of the facsimile indicate different letter types.

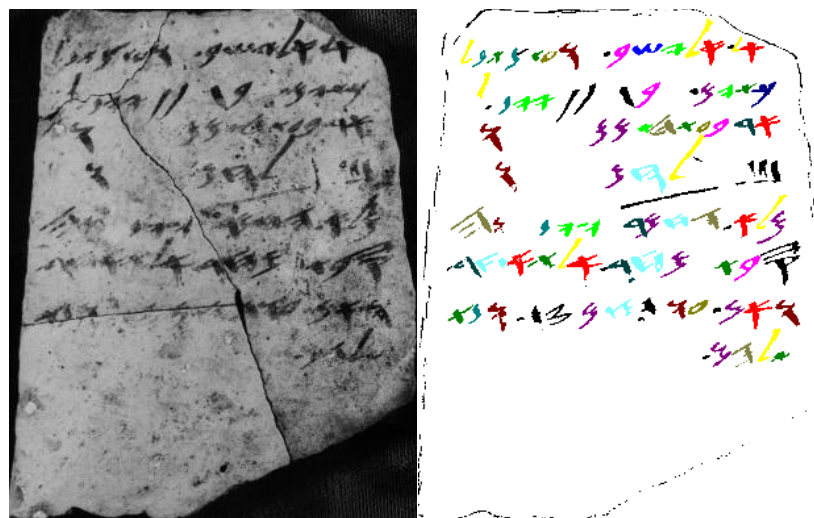


Figure 11.3 Arad 2 - an ostracon image (left) and its corresponding facsimile (right). The various colors of the facsimile indicate different letter types.

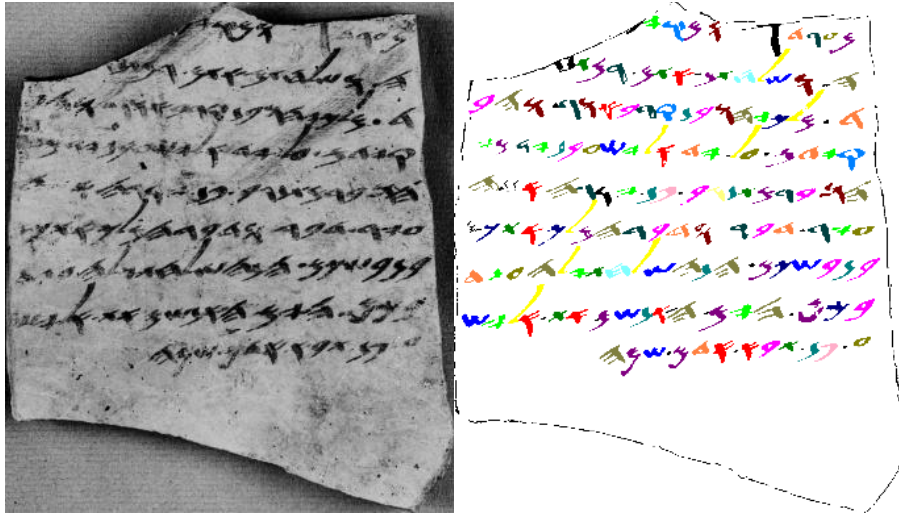


Figure 11.4 Arad 24b - an ostracon image (left) and its corresponding facsimile (right). The various colors of the facsimile indicate different letter types.

This size of the ostracon images was reduced by half (on each side) in some of the experiments, in order to test the performance of the algorithm in such setting. In total, 310 characters were utilized. Several representative examples of the algorithm's steps and its outcomes are provided in the following figures.

Fig. 11.5 shows an illustration of the algorithm's flow on a letter "yod" from Arad 24b. On the top row, a refinement of the letter (based on information from 14 characters), and an estimation of two consequent priors is shown, with no attempt at regularization (smoothing). On the bottom row, three consecutive priors are regularized by an algorithm from Section 10, with the regularization parameter set to $REG = 5$. Similarly, Fig. 11.6 shows the steps for a regularized computation of "mem" from Arad 2 (based on 10 characters).



Figure 11.5 An example of the algorithm’s flow for “*yod*” letter, Arad 24b. Top: a median-based initialization of a prior (utilizing information from 14 characters), and an estimation of two consequent priors, with no attempt at regularization (smoothing). Bottom: three consecutive priors are regularized by an algorithm introduced in Section 10, with median radius set to 5.



Figure 11.6 Steps for a regularized prior computation of “*mem*” from Arad 2 (based on 10 characters).

Fig. 11.7 provides a computation of a prior for the letter “*ayin*” from Arad 1 ostracon, in both full and partial resolution. It can be observed that in this case, “less is more”.



Figure 11.7 The letter “*ayin*” from Arad 1 (based on 3 characters). Top: computation of letter prior for full resolution imagery, regularization with radius=5. Bottom: computation of letter prior for partial resolution (halved in each axis), with no regularization, radius=5 and radius=10.

As visual observations of the results are subjective in nature, and since neither ancient nor modern writing specimens possess any kind of ground truth for letters’

priors, we settled on an experimental methodology akin to the one presented in Section 6. Each and every facsimile *character* of each and every ostracon was treated (in its turn) as “artificial” ground truth for a *letter’s prior*. Subsequently, “synthetic” character instances were obtained by adding incrementally increasing levels of disturbances to this prior. The resulting synthetic characters were utilized in order to estimate a prior. Finally, this estimation was compared to the “ground truth”, in order to deduce the precision and recall. Some details on the settings of various experiments, including the Gaussian noise levels, as well as the number of character instances involved, and the total number of experiments, are provided in Table 11.1.

Table 11.1 Experiments’ settings

Experiment	Settings		
	<i>Gaussian noise levels</i>	<i>Number of instances for each prior</i>	<i>Total number of experiments</i>
#1	Standard deviation of 200 gray values (out of 255).	2, 4, 6, 8, 10	1550
#2	Standard deviations of 50, 100, 150, 200 and 250 gray values (out of 255).	5	1550

In total, **3100** experiments were performed. The whole series of experiments took 586.2 seconds on Intel Core M-5Y10c 0.8GhZ, with 8 GB of memory on a single thread with no parallel computing. The results of experiment #1 for different ostraca can be seen at Tables 11.2-11.4 and Fig. 11.8. They indicate the robustness of the algorithm with respect to the number of characters, with excellent results for at least 4 characters.

Table 11.2 Results of experiment #1 for Arad 1 ostracon

Gaussian noise levels	Results for each scenario		
	Number of character instances for each prior	Average precision	Average recall
std = 200 gray values (out of 255)	2	94.55%	88.13%
	4	98.55%	98.17%
	6	99.07%	98.88%
	8	99.18%	99.02%
	10	99.19%	99.07%

Table 11.3 Results of experiment #1 for Arad 2 ostracon

Gaussian noise levels	Results for each scenario		
	Number of character instances for each prior	Average precision	Average recall
std = 200 gray values (out of 255)	2	90.28%	89.00%
	4	97.64%	97.71%
	6	98.46%	98.39%
	8	98.63%	98.50%
	10	98.66%	98.52%

Table 11.4 Results of experiment #1 for Arad 24b ostracon

Gaussian noise levels	Results for each scenario		
	Number of character instances for each prior	Average precision	Average recall
std = 200 gray values (out of 255)	2	87.23%	89.34%
	4	97.44%	97.82%
	6	98.73%	98.64%
	8	99.04%	98.87%
	10	99.14%	98.96%

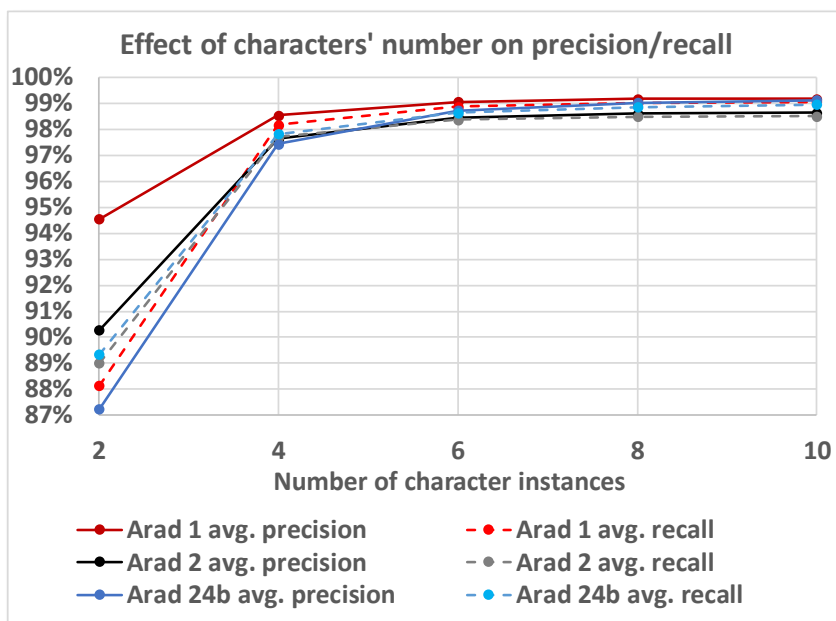


Figure 11.8 Results of experiment #1 for different ostraca.

The results of experiment #2 for different ostraca can be seen at Tables 11.5-11.7 and Fig. 11.9. The results indicate only a minor influence of the amount of noise on the average precision and recall.

Table 11.5 Results of experiment #2 for Arad 1 ostrakon

Number of character instances for each prior	Results for each scenario		
	Gaussian noise level (std)	Average precision	Average recall
5	50	99.05%	99.03%
	100	99.11%	99.05%
	150	99.27%	98.92%
	200	99.01%	98.10%
	250	97.83%	95.98%

Table 11.6 Results of experiment #2 for Arad 2 ostracon

Number of character instances for each prior	Results for each scenario		
	Gaussian noise level (std)	Average precision	Average recall
5	50	98.43%	98.42%
	100	98.53%	98.45%
	150	98.76%	98.35%
	200	98.45%	97.52%
	250	96.81%	95.43%

Table 11.7 Results of experiment #2 for Arad 24b ostracon

Number of character instances for each prior	Results for each scenario		
	Gaussian noise level (std)	Average precision	Average recall
5	50	99.09%	99.05%
	100	99.14%	99.05%
	150	99.19%	98.71%
	200	98.62%	97.56%
	250	96.81%	95.27%

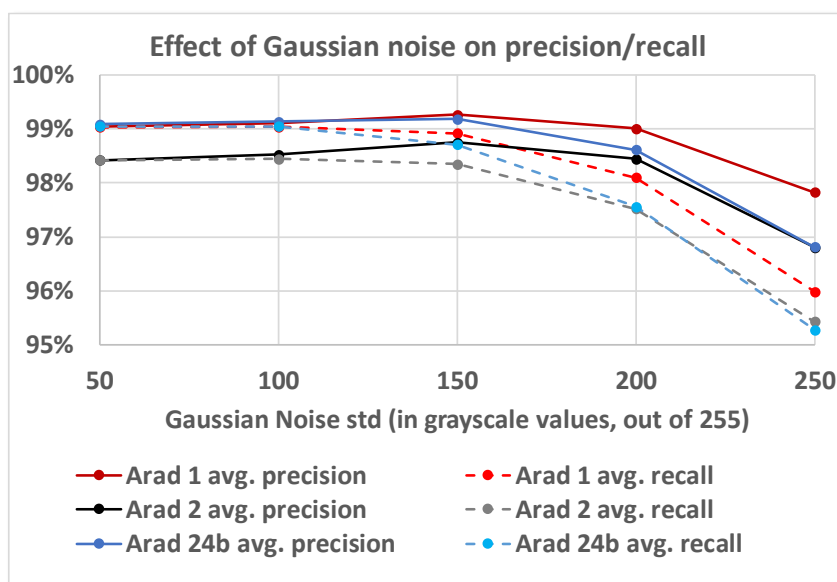


Figure 11.9 Results of experiment #2 for different ostraca.

11.5 Summary

The results of the experiments indicate the potential of our technique, particularly in the context of degraded historical characters. The algorithm is straightforward to implement, and is very fast. The dependence of our method on the number of characters is limited, and the results are only moderately affected by the accumulated noise.

12. Conclusions and Future Research Directions

This thesis presents diverse results pertaining to various documents' analysis issues, including image acquisition, binarization (either manual or automatic) and its quality assessment, writers' differentiation, image segmentation, and letters' shape prior estimation. All these methods are grounded upon real-world archaeological data and concrete empirical results, usable in their own right.

Among the main results of the thesis are:

- **Section 2:** Creation of a quality evaluation mechanism for manually created facsimiles. In addition, we establish that manual facsimiles, including the ones created by professional epigraphers, possess a non-negligible degree of subjectivity.
- **Section 3:** Creation of a quality evaluation mechanism for registered images of the same inscription, e.g. channels of multispectral images. The new Potential Contrast metric possesses several beneficial properties. One of them allows us to account, *analytically*, for all possible grayscale transformations of the image, upon evaluating its quality.
- **Section 4:** Introduction of a new binarization algorithm, harvesting useful data from the existing imperfect facsimiles of the inscriptions.
- **Section 5:** Sparse coding methods are adapted to the bi-color case, in order to improve the quality of imperfect binarizations of noisy documents.
- **Section 6:** Several quality measures, measuring the quality of binarizations vis-à-vis the original image (bypassing the currently common Ground Truth-based techniques), are introduced and tested.
- **Section 7:** Several metrics, measuring the intrinsic quality of individual binarized characters, are introduced. We tested various ways to combine the different

measures, with tree-based models which are shown to be the most advantageous in our case.

- **Section 8:** A framework consisting of several steps is introduced, in order to deal with a separation of writers of different inscriptions. The novelties of this approach include a new technique for features' combination; an independent experimentation on individual characters' level via randomized test; a p-value calculation of each experiment; a p-value combination between different *characters* via Fisher's method; a dichotomy of "separation" vs. "agnostic" results; and an independent results' verification via random graphs. The outcome of the algorithm on Arad corpus leads to historical implications.
- **Section 9:** An alternative framework to deal with the separation of writers of different inscriptions is introduced. The novelties of this approach include: a utilization of low-level binary pixel patterns' histogram per each character, potentially leading to hundreds of exceedingly general features (items in the histogram) instead of a few "tailored" ones in Section 8; a utilization of Kolmogorov-Smirnov a-parametric statistical test in order to deduce the p-value; a p-value combination between different *characters and features* (instead of only characters) via Fisher's method, leading to improved and more significant results.
- **Section 10:** A detailed analysis of the classical Chan-Vese segmentation algorithm reveals it is actually equivalent to an easily solvable Otsu binarization criteria, followed by a simple morphological median filter. This leads to the creation of a very fast segmentation and smoothing framework.
- **Section 11:** A letters' prior calculation technique is introduced, incorporating not only registration-based character image averaging, but also segmentation and smoothing via the newly introduced algorithm.

Several research directions, not exhausted in this thesis, are worth pursuing:

- The binarization procedure described in this thesis consists of two steps. First, a binarization is created, via a registration-based scheme (Section 4). Next, the binarization is improved, based on a clean and sparse dictionary (Section 5). In our view, a single-step binarization, possibly deriving the bi-color dictionary from the noisy grayscale data itself (i.e. “denoising” on dictionary level) can also be considered. This will require an approach different from the current ones, which commonly impose linear conditions between dictionary members upon dictionary construction, and their linear combinations upon image denoising. Such procedures are incompatible with strictly binary (i.e. black and white) data.
- The authors’ identification algorithms presented in Sections 8 and 9 consider two states of affairs – either a separation is established, and the authors of the two inscriptions under comparison are declared to be different, or there is not enough supporting evidence for such an assertion and the algorithm remains agnostic. Although this kind of reasoning is in no way rare in statistics, one may wonder whether a “same writer” conclusion can also be reached, possibly via some other computational mechanism. Indeed, this may be achievable, either by empirical estimation of various characters’ parameters, in case enough samples are present, or via equivalence testing (Rogers et al. 1993), which requires assumptions regarding the distributions in question.
- Letters’ shape prior estimation (described in Section 11) can benefit from a multiplicity of priors. In other words, even a single author can have several “ideal prototypes” for *alep*, *bet* etc. Such an “fine-grained” assumption is indeed made by (Bar-Yosef et al. 2007-9), with satisfactory outcomes.

- A further improvement to the shape prior algorithm, is an addition of a recursive step, “feeding” the priors back into the document, by using variational registration methods taking the prior into account. Such methods for general images were used in (Riklin-Raviv et al. 2007; Cohen et al. 2012). Properly registered priors can be used for a superior segmentation of the images, producing ever-improved priors, till convergence is achieved.
- An existence of priors for each author, corpus and period, can also assist in the creation of an OCR handwrite recognition tool for the First Temple period Hebrew, which was sidestepped in this thesis. This may involve a carefully supervised priors’ refinement procedure, akin to (Van Oosten and Schomaker 2014). Alternatively, it can benefit from dynamic labeling variational techniques such as the one presented in (Cremers et al. 2003).
- Finally, the employment of the proposed techniques to other historical (alphabetical, e.g. Greek and Latin, but also cuneiform and hieroglyphical) and modern inscriptions can be performed.

These challenging developments may constitute a worthwhile continuation of the current study.

13. References

- [Aharon et al. 2006] M. Aharon, M. Elad, A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation”, *IEEE Transactions on Signal Processing* 54.11, 4311-4322, 2006.
- [Aharoni 1981] Y. Aharoni, *Arad Inscriptions*. Israel Exploration Society, 1981.
- [Ahituv 2008] S. Ahituv, *Echoes from the Past: Hebrew and Cognate Inscriptions from the Biblical Period*, Carta, Jerusalem, 2008.
- [Ahonen et al. 2006] T. Ahonen, A. Hadid, M. Pietikäinen, “Face description with local binary patterns: Application to face recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.12, 2037-2041, 2006.
- [Aiolli and Giollo 2011] F. Aiolli, M. Giollo, “A study on the writer identification task for paleographic document analysis,” *Proceedings of the 11th IASTED International Conference on Artificial Intelligence and Applications (AIA 2011)*, 2011.
- [Akiyama et al. 1998] T. Akiyama, N. Miyamoto, M. Oguro, K. Ogura, “Faxed document image restoration method based on local pixel patterns”, *Proceedings of Photonics West '98 Electronic Imaging Symposium*, 253-262, 1998.
- [Albertz 2003] R. Albertz, *Israel in Exile: The History and Literature of the Sixth Century B.C.E.*, Society of Biblical Literature, Atlanta, 2003.
- [Al-Maadeed et al. 2016] S. Al-Maadeed, A. Hassaine, A. Bouridane, M. A. Tahir, “Novel geometric features for off-line writer identification”, *Pattern Analysis and Applications* 19, 699-708, 2016.
- [Álvarez et al. 2010] L. Álvarez, L. Baumela, P. Henríquez, P. Márquez-Neila, “Morphological snakes”, *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, 2197-2202, 2010
- [Ancient Hebrew 2016] http://www-nuclear.tau.ac.il/~eip/ostraca/DataSets/Arad_Ancient_Hebrew.zip
- [Anh et al. 2013] N. T. L. Anh, S.-H. Kim, H.-J Yang, “Color image segmentation using a morphological gradient-based active contour model”, *International Journal of Innovative Computing, Information and Control* 9.11, 4471-4484, 2013.
- [Armon 2012] <http://www.mathworks.com/matlabcentral/fileexchange/35038-descriptor-for-shapes-and-letters--feature-extraction>
- [Barkay 1992] G. Barkay, “The priestly benediction on silver plaques from Ketef Hinnom in Jerusalem”, *Tel Aviv* 19.2, 139-192, 1992.
- [Barkay et al. 2003] G. Barkay, M. J. Lundberg, A. G. Vaughn, B. Zuckerman, K. Zuckerman, “The challenges of Ketef Hinnom: Using advanced technologies to reclaim the earliest biblical texts and their context”, *Near Eastern Archaeology* 66.4, 162-171, 2003.

- [Barkay et al. 2004] G. Barkay, A. G. Vaughn, M. J. Lundberg, B. Zuckerman, “The amulets from Ketef Hinnom: A new edition and evaluation”, *Bulletin of the American Schools of Oriental Research* 334, 41-71, 2004.
- [Barney Smith 2010] E. H. Barney Smith. “An analysis of binarization ground truthing”, *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS 2010)*, 27-33, 2010.
- [Barney Smith and An 2012] E. H. Barney Smith, C. An, “Effect of ‘ground truth’ on image binarization”, *Proceedings of the 10th IAPR Workshop on Document Analysis Systems (DAS 2012)*, 250-254, 2012.
- [Bar-Yosef et al. 2007] I. Bar-Yosef, I. Beckman, K. Kedem, I. Dinstein, “Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents”, *International Journal of Document Analysis and Recognition* 9.2, 89-99, 2007.
- [Bar-Yosef et al. 2008] I. Bar-Yosef, A. Mokeichev, K. Kedem, I. Dinstein, “Global and local shape prior for variational segmentation of degraded historical characters”, *Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition (ICFHR 2008)*, 198-203, 2008.
- [Bar-Yosef et al. 2009] I. Bar Yosef, A. Mokeichev, K. Kedem, I. Dinstein, “Adaptive shape prior for recognition and variational segmentation of degraded historical characters”, *Pattern Recognition* 42, 3348-3354, 2009.
- [Beit-Arieh 2007] I. Beit-Arieh, *Horvat 'Uza and Horvat Radum: Two Fortresses in the Biblical Negev*, Tel Aviv: Emery and Claire Yass Publications in Archaeology, 2007.
- [Beit-Arieh and Freud 2015] I. Beit-Arieh, L. Freud, *Tel Malhata: A Central City in the Biblical Negev, Volume I*, Tel Aviv: Emery and Claire Yass Publications in Archaeology, 2015.
- [Ben Messaoud et al. 2011] I. Ben Messaoud, H. El Abed, H. Amiri, V. Märgner, “A design of a preprocessing framework for large database of historical documents”, *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (HIP 2011)*, 177-183, 2011.
- [Ben Messaoud et al. 2012] I. Ben Messaoud, H. Amiri, H. El Abed, V. Märgner, “Region based local binarization approach for handwritten ancient documents”, *Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012)*, 633-638, 2012.
- [Benjamini and Hochberg 1995] Y. Benjamini, Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing”, *Journal of the Royal Statistical Society, Series B* 57.1, 289-300, 1995.
- [Bern and Goldberg 2000] M. Bern, D. Goldberg, “Scanner-model-based document image improvement”, *Proceedings of the 2000 International Conference on Image Processing (ICIP 2000)*, 582-585, 2000.

- [Bernsen 1986] J. Bernsen, “Dynamic thresholding of grey-level images”, Proceedings of the Eighth International Conference on Pattern Recognition (ICPR 1986), 1251–1255, 1986.
- [Biller et al. 2013] O. Biller, A. Asi, K. Kedem, J. El-Sana, I. Dinstein, “WebGT: An interactive web-based system for historical document ground truth generation”, Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), 305-308, 2013.
- [Breuel 2001] T. M. Breuel, “Segmentation of handprinted letter strings using a dynamic programming algorithm”, Proceedings of the Sixth International Conference on Document Analysis and Recognition, (ICDAR 2001), 821-826, 2001.
- [Brown et al. 1988] R. M. Brown, T. H. Fay, C. U. Walker, “Handprinted symbol recognition system”, Pattern Recognition 21.2, 91-118, 1988.
- [Brown et al. 2008] B. J. Brown, C. Toler-Franklin, D. Nehab, M. Burns, D. Dobkin, A. Vlachopoulos, C. Dumas, S. Rusinkiewicz, T. Weyrich, “A system for high-volume acquisition and matching of fresco fragments: Reassembling Theran wall paintings”, ACM Transactions on Graphics (TOG), Proceedings of ACM SIGGRAPH 2008, 27.3, 84, 2008.
- [Brown et al. 2012] E. S. Brown, T. F. Chan, X. Bresson, “Completely convex formulation of the Chan-Vese image segmentation model”, International Journal of Computer Vision 98.1, 103-121, 2012.
- [Bulacu and Schomaker 2007] M. Bulacu, L. Schomaker, “Automatic handwriting identification on medieval documents”, Proceedings of the 14th International Conference on Image Analysis and Processing (ICIAP 2007), 279-284, 2007.
- [Bylinskii et al. 2016] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, “MIT saliency benchmark” website, <http://saliency.mit.edu/>
- [Canny 1986] J. Canny, “A computational approach to edge detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence 8.6, 679-698, 1986.
- [Casey and Lecolinet 1996] R. G. Casey, E. Lecolinet, “A survey of methods and strategies in character segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence 18.7, 690-706, 1996.
- [Catté et al. 1995] F. Catté, F. Dibos, G. Koepfler, “A morphological scheme for mean curvature motion and applications to anisotropic diffusion and motion of level sets”, SIAM Journal on Numerical Analysis 32.6, 1895-1909, 1995.
- [Chan and Vese 2001] T. F. Chan, L. Vese, “Active contours without edges”, IEEE Transactions on Image Processing 10.2, 266-277, 2001.
- [Chan et al. 2000] T. F. Chan, B. Sandberg Yezriev, L. Vese, “Active contours without edges for vector-valued images”, Journal of Visual Communication and Image Representation 11.2, 130-141, 2000.

- [Clausner et al. 2011] C. Clausner, S. Pletschacher, A. Antonacopoulos, “Aletheia - An advanced document layout and text ground-truthing system for production environments”, Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), 48-52, 2011.
- [Cohen et al. 2012] R. Cohen, K. Kedem, I. Dinstein, J. El-Sana, “Occluded character restoration using active contour with shape priors”, Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), 497-502, 2012.
- [Corder and Foreman 2014] G. W. Corder, D. I. Foreman, Nonparametric Statistics: A Step-by-Step Approach, Wiley, Hoboken NJ, 2014.
- [Cremers et al. 2003] D. Cremers, N. Sochen, C. Schnörr, “Towards recognition-based variational segmentation using shape priors and dynamic labeling”, Proceedings of the International Conference on Scale-Space Theories in Computer Vision, 388-400, 2003.
- [Cutter et al. 2010] M. P. Cutter, J. van Beusekom, F. Shafait, T. M Breuel, “Unsupervised font reconstruction based on token co-occurrence”, Proceedings of the 10th ACM Symposium on Document engineering (DocEng 2010), 143-150, 2010.
- [Cutter et al. 2011] M. P. Cutter, J. van Beusekom, F. Shafait, T. M Breuel, “Font group identification using reconstructed fonts”, Proceedings of SPIE 7874, Document Recognition and Retrieval XVIII (DRR XVIII), 78740N-1-8, 2011.
- [Davis et al. 1997] G. Davis, S. Mallat, M. Avellaneda, “Adaptive greedy approximations”, Constructive Approximation 13, 57–98, 1997.
- [De Hoon et al. 2004] M. J. L. de Hoon, S. Imoto, J. Nolan, S. Miyano, “Open source clustering software”, Bioinformatics 20.9, 1453-1454, 2004, <http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm#pycluster>
- [Dead Sea Scrolls 2016] Dead Sea Scrolls Digital Project, the Israel Museum, Jerusalem website, <http://dss.collections.imj.org.il>
- [Devroye and McDougall 1995] L. Devroye, M. McDougall, “Random fonts for the simulation of handwriting”, Electronic Publishing 8.4, 281-294, 1995.
- [Dinstein and Shapira 1982] I. Dinstein, Y. Shapira, “Ancient Hebraic handwriting identification with run-length histograms”, IEEE Transactions on Systems, Man, and Cybernetics 12.3, 405-409, 1982.
- [Droettboom et al. 2012] M. Droettboom, I. Fujinaga, K. MacMillan, G. S. Chouhury, T. DiLauro, M. Patton, T. Anderson, “Using the Gamera framework for the recognition of cultural heritage materials”, Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, 11-17, 2002, <http://gamera.informatik.hsnr.de>.
- [Dunn 1961] O. J. Dunn, “Multiple comparisons among means”, Journal of the American Statistical Association 56.293, 52-64, 1961.
- [Edwards and Forsyth 2006] J. Edwards, D. Forsyth, “Searching for character models”, Advances in Neural Information Processing Systems 18 (NIPS 2005), 331-338, 2006.

[Engan et al. 1999] K. Engan, B. D. Rao, K. Kreutz-Delgado, “Frame design using focus with method of optimal directions (MOD)”, Proceedings of Norwegian Signal Processing Symposium, 65-69, 1999.

[Epshtein et al. 2010] B. Epshtein, E. Ofek, Y. Wexler, “Detecting text in natural scenes with stroke width transform”, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), 2963-2970, 2010.

[Faigenbaum et al. 2012] S. Faigenbaum, B. Sober, **A. Shaus**, M. Moinester, E. Piasetzky, G. Bearman, M. Cordonsky, I. Finkelstein, “Multispectral images of ostraca: Acquisition and analysis”, Journal of Archaeological Science 39.12, 3581–3590, 2012.

[Faigenbaum et al. 2014] S. Faigenbaum, B. Sober, I. Finkelstein, M. Moinester, E. Piasetzky, **A. Shaus**, M. Cordonsky, “Multispectral imaging of two Hieratic inscriptions from Qubur el-Walaydah,” Ägypten und Levante XXIV, 349-53, 2014

[Faigenbaum et al. 2015] S. Faigenbaum, B. Sober, M. Moinester, E. Piasetzky, G. Bearman, “Multispectral imaging of Tel Malhata ostraca”, In: I. Beit-Arieh, L. Freud, eds. Tel Malhata: A Central City in the Biblical Negev, Volume I. Tel Aviv: Emery and Claire Yass Publications in Archaeology, 510-513, 2015.

[Faigenbaum, Shaus, Sober et al. 2013] S. Faigenbaum, **A. Shaus**, B. Sober, E. Turkel, E. Piasetzky, “Evaluating glyph binarizations based on their properties”, Proceedings of the 2013 ACM symposium on Document engineering (DocEng 2013), 127-130, 2013.

[Faigenbaum-Golovin et al. 2015] S. Faigenbaum-Golovin, C. A. Rollston, E. Piasetzky, B. Sober, I. Finkelstein, “The Ophel (Jerusalem) ostracon in light of new multispectral images”, Semitica 57, 113-37, 2015.

[Faigenbaum-Golovin et al. 2017] S. Faigenbaum-Golovin, A. Mendel-Geberovich, **A. Shaus**, B. Sober, M. Cordonsky, D. Levin, M. Moinester, B. Sass, E. Turkel, E. Piasetzky, I. Finkelstein, “Multispectral imaging reveals biblical-period inscription unnoticed for half a century”, PLOS ONE 12.6, e0178400, 2017.

[Faigenbaum-Golovin, Shaus, Sober et al. 2015] S. Faigenbaum-Golovin, **A. Shaus**, B. Sober, I. Finkelstein, D. Levin, M. Moinester, E. Piasetzky, E. Turkel, “Computerized paleographic investigation of Hebrew Iron Age ostraca”, Radiocarbon 57.2, 317-325, 2015.

[Faigenbaum-Golovin, Shaus, Sober et al. 2016] S. Faigenbaum-Golovin, **A. Shaus**, B. Sober, D. Levin, N. Na’aman, B. Sass, E. Turkel, E. Piasetzky, I. Finkelstein, “Algorithmic handwriting analysis of Judah’s military correspondence sheds light on composition of biblical texts”, Proceedings of the National Academy of Sciences 113.17, 4664-4669, 2016.

[Faigenbaum-Golovin, Shaus, Sober et al. 2017] S. Faigenbaum-Golovin, **A. Shaus**, B. Sober, A. Mendel-Geberovich, E. Piasetzky, I. Finkelstein, “Shedding light on Iron Age Hebrew ostraca via modern imaging and computational technologies”, TAU Archaeology 3, 12, 2017.

- [Fecker et al. 2014a] D. Fecker, A. Asi, W. Pantke, V. Märgner, J. El-Sana, T. Fingscheidt, “Document writer analysis with rejection for historic Arabic manuscripts”, Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014), 743-748, 2014.
- [Fecker et al. 2014b] D. Fecker, A. Asi, V. Märgner, J. El-Sana, and T. Fingscheidt, “Writer identification for historical Arabic documents”, Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014), 3050-3055, 2014.
- [Fiel et al. 2014] S. Fiel, F. Hollaus, M. Gau, R. Sablatnig, “Writer identification on historical Glagolitic documents”, Proceedings of SPIE 9021, Document Recognition and Retrieval XXI, 902102, 2014.
- [Finkelstein et al. 2012] I. Finkelstein, S. Ben Dor Evian, E. Boaretto, D. Cabanes, M. T. Cabanes, A. Eliyahu-Behar, S. Faigenbaum, Y. Gadot, D. Langgut, M. Martin, M. Meiri, D. Namdar, L. Sapir-Hen, R. Shahack-Gross, **A. Shaus**, B. Sober, M. Toffolo, N. Yahalom-Mack, L. Zapassky, S. Weiner, “Reconstructing ancient Israel: Integrating macro- and micro-archaeology”, Hebrew Bible and Ancient Israel 1.1, 133-150, 2012.
- [Finkelstein et al. 2015] I. Finkelstein, S. Weiner, E. Boaretto, “Preface—The Iron Age in Israel: The exact and life sciences perspectives”, Radiocarbon 57.2, 197–206, 2015.
- [Fischer et al. 2010] A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, M. Stolz, “Ground truth creation for handwriting recognition in historical documents”, Proceedings of the 9th IAPR Workshop on Document Analysis Systems (DAS 2010), 3-10, 2010.
- [Fisher 1925] R. A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh, 1925.
- [Fox et al. 2013a] V. L. Fox, M. Milanova, S. Al-Ali, “A hybrid morphological active contour for natural images”, International Journal of Computer Science, Engineering and Applications 3.4, 1-13, 2013.
- [Fox et al. 2013b] V. L. Fox, M. Milanova, S. Al-Ali, “A morphological multiphase active contour for vascular segmentation”, International Journal on Bioinformatics & Biosciences 3.3, 1-12, 2013.
- [Freund and Schapire 1997] Y. Freund, R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting”, Journal of Computer and System Sciences 55.1, 119–139, 1997.
- [Gatos et al. 2004] B. Gatos, I. Pratikakis, S. J. Perantonis, “An adaptive binarization technique for low quality historical documents”, Lecture Notes in Computer Science 3163, 102-113, 2004.
- [Gatos et al. 2006] B. Gatos, I. Pratikakis, S. Perantonis, “Adaptive degraded document image binarization”, Pattern Recognition 39, 317–327, 2006.

- [Gatos et al. 2009] 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), 1375-1382, 2009.
- [Gatos et al. 2011] B. Gatos, K. Ntirogiannis, I. Pratikakis, “DIBCO 2009: Document image binarization contest,” International Journal on Document Analysis and Recognition 14.1, 35-44, 2011.
- [Gersho and Gray 1991] A. Gersho, R. M. Gray, Vector Quantization and Signal Compression, Norwell, MA, Kluwer Academic, 1991.
- [Gilboa et al. 2004] A. Gilboa, A. Karasik, I. Sharon, U. Smilansky, “Towards computerized typology and classification of ceramics”, Journal of Archaeological Science 31.6, 681-694, 2004.
- [Griechisch et al. 2014] E. Griechisch, M. I. Malik, M. Liwicki, “Online signature verification based on Kolmogorov-Smirnov distribution distance”, Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014), 738-742, 2014.
- [Grossman 2010] M. L. Grossman (ed.), Rediscovering the Dead Sea Scrolls, Eerdmans, 2010.
- [He et al. 2005] 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), 538-542, 2005.
- [Herzog 2002] Z. Herzog, “The fortress mound at Tel Arad: An interim report”, Tel Aviv 29.1, 3-109, 2002.
- [Hunt et al. 2001] L. Hunt, M. J. Lundberg, B. Zuckerman, “Eyewitness to the past: Reclaiming ancient inscriptions with modern technologies through USC’s West Semitic Research and InscriptiFact projects”, Biblos 50.1, 79-100, 2001.
- [Hunter 2007] J. D. Hunter, “Matplotlib: A 2D graphics environment”, Computing in Science & Engineering 9, 90-95, 2007, <https://matplotlib.org>, version 2.0.0.
- [Jain and Dubes 1981] A. K. Jain, R. C. Dubes, Algorithms for Clustering Data, Prentice-Hall, 1981.
- [Jalba and Roerdink 2009] A. C. Jalba, J. B. T. M. Roerdink, “An efficient morphological active surface model for volumetric image segmentation”, Proceedings of the International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing (ISMM 2009), 193-204, 2009.
- [Jones et al. 2001] E. Jones, E. Oliphant, P. Peterson et al., “SciPy: Open Source Scientific Tools for Python”, 2001-, <http://www.scipy.org>, version 0.19.0.
- [Kamel and Zhao 1993] M. Kamel, A. Zhao, “Extraction of binary character/graphics images from grayscale document images”, CVGIP: Graphical Models and Image Processing 55.3, 203-217, 1993.

- [Kapur et al. 1985] J. N. Kapur, P. K. Sahoo, A. K. C. Wong, “A new method for gray-level picture thresholding using the entropy of the histogram”, *CVGIP: Computer Vision, Graphics, and Image Processing* 29.3, 273-285, 1985.
- [Kaufman and Rousseeuw 1987] L. Kaufman, P. J. Rousseeuw, “Clustering by means of medoids”, in: Y. Dodge ed., *Statistical Data Analysis Based on the L1-Norm and Related Methods*, North-Holland, Birkhäuser Basel, 405–416, 1987.
- [Kendall 1938] M. Kendall, “A new measure of rank correlation”, *Biometrika* 30. 1/2, 81–93, 1938.
- [Keyzers and Unger 2003] D. Keyzers, W. Unger, “Elastic image matching is NP-complete”, *Pattern Recognition Letters* 24.1, 445-453, 2003.
- [Kishore et al. 2015] P. V. V. Kishore, C. R. Prasad, “Train rolling stock segmentation with morphological differential gradient active contours”, *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI 2015)*, 1174-1178, 2015
- [Kittler and Illingworth 1986] J. Kittler, J. Illingworth, “Minimum error thresholding”, *Pattern Recognition* 19.1, 41-47, 1986.
- [Kopec and Lomelin 1997] G. E. Kopec, M. Lomelin, “Supervised template estimation for document image decoding”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.12, 1313-1324, 1997.
- [Kreutz-Delgado and Rao 2000] K. Kreutz-Delgado, B. D. Rao, “FOCUSS-based dictionary learning algorithms”, *Wavelet Applications in Signal and Image Process. VIII*, pp. 4119-53, 2000.
- [Lai and Kuo 2000] Y. K. Lai, C. C. J. Kuo, “A Haar wavelet approach to compressed image quality measurement”, *Journal of Visual Communication and Image Representation* 11.1, 17-40, 2000.
- [Lavee 2013] T. Lavee, *Computer Analysis of the Dead Sea Scroll Manuscripts*, MSc Thesis, Tel Aviv University, 2013.
- [Lee and Chen 1992] H. J. Lee, B. Chen, “Recognition of handwritten Chinese characters via short line segments”, *Pattern Recognition* 25.5, 543-552, 1992.
- [Lemaire 1977] A. Lemaire, *Inscriptions Hébraïques, Vol. 1: Les Ostraca, Littératures Anciennes du Proche-Orient* 9, Paris: Cerf, 230-231, 1977.
- [Lemaire 1981] A. Lemaire, *Les Écoles et la Formation de la Bible dans l’ancien Israël*, OBO 39, Fribourg and Göttingen, 1981.
- [Lesage et al. 2005] S. Lesage, R. Gribonval, F. Bimbot, L. Benaroya, “Learning unions of orthonormal bases with thresholded singular value decomposition”, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’05)* 5, 293-296, 2005.

- [Leu 1992] J. G. Leu, “Image contrast enhancement based on the intensities of edge pixels”, *CVGIP: Graphical Models and Image Processing* 54.6, 497-506, 1992.
- [Lewicki and Olshausen 1999] M. S. Lewicki, B. A. Olshausen, “A probabilistic framework for the adaptation and comparison of image codes”, *Journal of the Optical Society of America A* 16, 1587–1601, 1999.
- [Li et al. 2009] X. Li, D. Tao, X. Gao, W. Lu, “A natural image quality evaluation metric”, *Signal Processing* 89.4, 548-555, 2009.
- [Lipschits and Vanderhooft 2011] O. Lipschits, D. S. Vanderhooft, *The Yehud Stamp Impressions: A Corpus of Inscribed Impressions from the Persian and Hellenistic Periods in Judah*, Eisenbrauns, Winona Lake, 2011.
- [Lipschits et al. 2008] O. Lipschits, I. Koch, **A. Shaus**, S. Guil, “The enigma of the biblical bath and the system of liquid volume measurement during the First Temple period”, *Ugarit-Forschungen* 42, 453-478, 2018.
- [Liu and Peng 2012] S. Liu, Y. Peng, “A local region-based Chan-Vese model for image segmentation”, *Pattern Recognition* 45.7, 2769-2779, 2012.
- [Lowe 2004] D. G. Lowe, “Distinctive image features from scale-invariant keypoints”, *International Journal of Computer Vision* 60.2, 91-110, 2004.
- [Lu et al. 2010] S. Lu, B. Su, C. L. Tan, “Document image binarization using background estimation and stroke edge”, *International Journal on Document Analysis and Recognition* 13.4, 303–314, 2010.
- [Mazar and Netzer 1986] A. Mazar, E. Netzer, “On the Israelite fortress at Arad”, *Bulletin of the American Schools of Oriental Research* 263, 87-91, 1986.
- [McGillivray et al. 2009] C. McGillivray, C. Hale, E. H. Barney Smith, “Edge noise in document images”, *Proceedings of the Third Workshop on Analytics for Noisy Unstructured Text Data (AND 2009)*, 17-24, 2009.
- [Mendel-Geberovich et al. 2017] A. Mendel-Geberovich, **A. Shaus**, S. Faigenbaum-Golovin, B. Sober, M. Cordonsky, E. Piasetzky, I. Finkelstein, “A brand new old inscription: Arad ostracon 16 rediscovered via multispectral imaging”, *Bulletin of the American Schools of Oriental Research (BASOR)* 378, 113-125, 2017.
- [Mendel-Geberovich et al. forthcoming] A. Mendel-Geberovich, S. Faigenbaum-Golovin, **A. Shaus**, B. Sober, M. Cordonsky, E. Piasetzky, I. Finkelstein, I. Milevski, “A renewed reading of Hebrew ostraca from cave A-2 at Ramat Beit Shemesh (Nahal Yarmut), based on multispectral imaging”, *Vetus Testamentum*, forthcoming.
- [Michelson 1927] A. A. Michelson, *Studies in Optics*, University of Chicago Press, 1927
- [Modern Hebrew 2016] http://www-nuclear.tau.ac.il/~eip/ostraca/DataSets/Modern_Hebrew.zip

- [Mumford and Shah 1989] D. Mumford, J. Shah, “Optimal approximation by piecewise smooth functions and associated variational problems”, *Communications on Pure and Applied Mathematics* 42, 577-685, 1989.
- [Na’aman 2002] N. Na’aman, *The Past that Shapes the Present: The Creation of Biblical Historiography in the Late First Temple Period and After the Downfall*, Yeriot, Jerusalem, 2002 (Hebrew).
- [Na’aman 2003] N. Na’aman, “Ostrakon 40 from Arad reconsidered”, in: C. G. Hertog, U. Hübner, S. Münger, eds., *Saxa Loquentur. Studien zur Archäologie Palästinas/Israels. Festschrift für Volkmar Fritz zum 65 Geburtstag*, AOAT 302, Münster, 199–204, 2003.
- [Na’aman 2010] N. Na’aman, “Textual and historical notes on the Eliashib archive from Arad”, *Tel Aviv* 38.1, 83-93, 2011.
- [Naveh 1960] J. Naveh, “A Hebrew letter from the seventh century B.C.”, *Israel Exploration Journal* 10.3, 129-139, 1960.
- [Négrate 1992] A. L. Négrate, A. Beghdadi, H. Dupoisot, “An image enhancement technique and its evaluation through bimodality analysis”, *CVGIP: Graphical Models and Image Processing* 54.1, 13-22, 1992.
- [Niblack 1986] W. Niblack, *An Introduction to Digital Image Processing*, Prentice-Hall, 115–116, 1986.
- [Nicolaou et al. 2014] A. Nicolaou, F. Slimane, V. Märgner, M. Liwicki, “Local binary patterns for Arabic optical font recognition”, *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS 2014)*, 76-80, 2014.
- [Ntirogiannis et al. 2008] K. Ntirogiannis, B. Gatos, I. Pratikakis, “An objective evaluation methodology for document image binarization techniques”, *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems (DAS 2008)*, 217-224, 2008.
- [Ntirogiannis et al. 2012] K. Ntirogiannis, B. Gatos, I. Pratikakis, “Performance evaluation methodology for historical document image binarization”, *IEEE Transactions on Image Processing* 22.2, 595-609, 2012.
- [Ntirogiannis et al. 2014] K. Ntirogiannis, B. Gatos, I. Pratikakis, “H-DIBCO 2014 – Handwritten document image binarization competition”, *Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014)*, 809-813, 2014.
- [Ojala et al. 1996] T. Ojala, M. Pietikäinen, D. Harwood, “A comparative study of texture measures with classification based on featured distributions”, *Pattern Recognition* 29.1, 51-59, 1996.
- [Oliveira et al. 2013] R. B. Oliveira, J. M. R. S. Tavares, N. Marranghello, A. S. Pereira, “An approach to edge detection in images of skin lesions by Chan-Vese model”, *Proceedings of the 8th Doctoral Symposium in Informatics Engineering*, 2013.

- [Olshausen and Field 1996] B. A. Olshausen, D. J. Field, “Natural image statistics and efficient coding”, *Network: Computation in Neural Systems* 7.2, 333–339, 1996.
- [Opelt 2006] A. Opelt, A. Pinz, M. Fussenegger, P. Auer, “Generic object recognition with boosting”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28.3, 416-431, 2006; http://www.emt.tugraz.at/~pinz/data/GRAZ_02
- [Osher and Sethian 1988] S. Osher, J. A. Sethian, “Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulation”, *Journal of Computational Physics* 79, 12-49, 1988.
- [Otsu 1979] N. Otsu, “A threshold selection method from gray-level histograms”, *IEEE Transactions on Systems, Man, and Cybernetics* 9.1, 62–66, 1979.
- [Pan et al. 2006] Y. Pan, J. D. Birdwell, S. M. Djouadi, “Efficient implementation of the Chan-Vese models without solving PDEs”, *IEEE 8th Workshop on Multimedia Signal Processing*, 350-354, 2006.
- [Panagopoulos et al. 2009] M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, S. Tracy, “Automatic writer identification of ancient Greek inscriptions”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.8, 1404-1414, 2009.
- [Pantke et al. 2013] W. Pantke, V. Märgner, D. Fecker, T. Fingscheidt, A. Asi, O. Biller, J. El-Sana, R. Saabni, M. Yehia, “HADARA—A software system for semi-automatic processing of historical handwritten Arabic documents”, *Proceedings of the Archiving Conference 2013*, 161-166, 2013.
- [Papaodysseus et al. 2014] C. Papaodysseus, P. Rousopoulos, F. Giannopoulos, S. Zannos, D. Arabadjis, M. Panagopoulos, E. Kalfa, C. Blackwell, S. Tracy, “Identifying the writer of ancient inscriptions and byzantine codices. A novel approach,” *Computer Vision and Image Understanding* 121, 57-73, 2014.
- [Paredes and Kavallieratou 2010] R. Paredes, E. Kavallieratou, “ICFHR 2010 contest: Quantitative evaluation of binarization algorithms”, *Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010)*, 733-736, 2010.
- [Pavel et al. 1987] M. Pavel, G. Sperling, T. Riedl, A. Vanderbeek, “Limits of visual communication: The effect of signal-to-noise ratio in the intelligibility of American sign language”, *Journal of the Optical Society of America A* 4.12, 2355-2365, 1987.
- [Pedregosa et al. 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, “Scikit-learn: Machine learning in Python”, *Journal of Machine Learning Research* 12, 2825-2830, 2011, <http://scikit-learn.org>, version 0.18.1.
- [Peli 1990] E. Peli, “Contrast in complex images”, *Journal of the Optical Society of America A* 7.10, 2032-2040, 1990.
- [PIL 2009] PIL library, version 1.1.6, <http://www.pythonware.com/products/pil>, 2009.

- [Pillow 2010] Pillow library, version 4.1.0, <https://github.com/python-pillow>, 2010-.
- [Pletschacher 2008] S. Pletschacher, “Representation of digitized documents using document specific alphabets and fonts”, Proceedings of the Archiving Conference 2008, 198-202, 2008.
- [Pletschacher 2009] S. Pletschacher, “A self-adaptive method for extraction of document-specific alphabets”, Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009), 656-660, 2009.
- [Potikha 2011] L. Potikha, Computerized Paleography Exploration of Historical Manuscripts, MSc Thesis, Tel Aviv University, 2011.
- [Pratikakis et al. 2010] I. Pratikakis, B. Gatos, K. Ntirogiannis, “H-DIBCO 2010 – Handwritten document image binarization competition”, Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010), 727-732, 2010.
- [Pratikakis et al. 2011] I. Pratikakis, B. Gatos, K. Ntirogiannis, “ICDAR 2011 document image binarization contest (DIBCO 2011)”, Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), 1506-1510, 2011.
- [Pratikakis et al. 2012] I. Pratikakis, B. Gatos, K. Ntirogiannis, “H-DIBCO 2012 – Handwritten document image binarization competition”, Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), 817-822, 2012
- [Pratikakis et al. 2013] I. Pratikakis, B. Gatos, K. Ntirogiannis, “ICDAR 2013 document image binarization contest (DIBCO 2013)”, Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), 1471-1476, 2013.
- [Pratt 1974] W. K. Pratt, “Correlation techniques of image registration”, IEEE Trans. on Aerospace and Electronic Systems 3, 353-358, 1974.
- [Python 2010] Python programming language, version 2.7, <https://www.python.org>, 2010
- [R Core Team 2012] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org>
- [Ratnakar 1998] V. Ratnakar, “RAPP, lossless image compression with runs of adaptive pixel patterns”, Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems & Computers, Vol 2, 1251–1255, 1998.
- [Reisner et al. 1924] G. A. Reisner, C. Fischer, D. Lyon, Harvard Excavations at Samaria, 1908–1910, Harvard University, Cambridge, 1924.

- [Riklin-Raviv et al. 2007] T. Riklin-Raviv, N. Kiryati, N. Sochen, “Prior-based segmentation and shape registration in the presence of perspective distortion”, *International Journal of Computer Vision* 72.3, 309-328, 2007.
- [Riklin-Raviv et al. 2008] T. Riklin-Raviv, N. Sochen, N. Kiryati, “Shape-based mutual segmentation”, *International Journal of Computer Vision* 79.3, 231-245, 2008.
- [Ripley 2016] B. Ripley, “tree: Classification and Regression Trees”, R package version 1.0-37, 2016.
- [Rogers et al. 1993] J. L. Rogers, K. I. Howard, J.T. Vessey, “Using significance tests to evaluate equivalence between two experimental groups”, *Psychological Bulletin* 113, 553-565, 1993.
- [Rollston 1999] C. A. Rollston, *The Script of Hebrew Ostraca of the Iron Age: 8th-6th Centuries BCE*, PhD Thesis, Johns Hopkins University, Baltimore, 1999.
- [Rollston 2006] C. A. Rollston, “Scribal education in ancient Israel: The Old Hebrew epigraphic evidence”, *Bulletin of the American Schools of Oriental Research* 344, 47–74, 2006.
- [Rollston 2010] C. A. Rollston, *Writing and Literacy in the World of Ancient Israel: Epigraphic Evidence from the Iron Age*, Society of Biblical Literature, Atlanta, 2010
- [Saund et al. 2009] E. Saund, J. Lind, P. Sarker, “PixLabeler: User interface for pixel-level labeling of elements in document images”, *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*, 646–650, 2009.
- [Sauvola and Pietikainen 2000] J. Sauvola, M. Pietikainen, “Adaptive document image binarization”, *Pattern Recognition* 33, 225–236, 2000.
- [Schmid 2012] K. Schmid, *The Old Testament, A Literary History*, Fortress Press, Minneapolis, 2012.
- [Schomaker 2007] L. R. B. Schomaker, “Writer identification and verification,” in N. Ratha, V. Govindaraju, eds., *Advances in Biometrics: Sensors, Systems and Algorithms*, Springer-Verlag, London, 247-264, 2007.
- [Schomaker 2016] L. Schomaker, “Design considerations for a large-scale image-based text search engine in historical manuscript collections”, *it-Information Technology* 58.2, 80-88, 2016, <http://application02.target.rug.nl/monk/demo.html>.
- [Schomaker et al. 2007] L. Schomaker, K. Franke, M. Bulacu, “Using codebooks of fragmented connected-component contours in forensic and historic writer identification”, *Pattern Recognition Letters* 28.6, 719-727, 2007.
- [SciPy 2001] O. E. Jones, P. Peterson P, et al., *SciPy: Open Source Scientific Tools for Python*, version 0.19.0, <http://www.scipy.org>, 2001-.
- [Sexton et al. 2000] A. Sexton, A. Todman, K. Woodward, “Font recognition using shape-based quad-tree and kd-tree decomposition”, *Proceedings of the 3rd International*

Conference on Computer Vision, Pattern Recognition and Image Processing, 212-215, 2000.

[Sezgin and Sankur 2004] M. Sezgin, B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation”, *Journal of Electronic Imaging* 13.1, 146-165, 2004.

[Shaus and Turkel 2016] **A. Shaus**, E. Turkel, “Chan-Vese revisited: Relation to Otsu’s method and a parameter-free non-PDE solution via morphological framework”, *Proceedings of the 12th International Symposium on Visual Computing (ISVC 2016)*, *Advances in Visual Computing, Lecture Notes in Computer Science 10072*, Vol. I, 203-212, 2016.

[Shaus and Turkel 2017a] **A. Shaus**, E. Turkel, “Writer identification in modern and historical documents via binary pixel patterns, Kolmogorov-Smirnov test and Fisher’s method”, *Proceedings of the IS&T International Symposium on Electronic Imaging 2017, 22nd Human Vision and Electronic Imaging Conference (HVEI 2017)*, 203-211; *Journal of Imaging Science and Technology* 61.1, 0104041–0104049, 2017.

[Shaus and Turkel 2017b] **A. Shaus**, E. Turkel, “Towards letter shape prior and paleographic tables estimation in Hebrew First Temple period ostraca”, *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing (HIP 2017)*, 13-18, 2017.

[Shaus et al. 2010] **A. Shaus**, I. Finkelstein, E. Piasezky, “Bypassing the eye of the beholder: Automated ostraca facsimile evaluation”, *Maarav* 17.1, 7-20, 2010.

[Shaus et al. 2012a] **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)*, 170-174, 2012.

[Shaus et al. 2012b] **A. Shaus**, E. Turkel, E. Piasezky, “Binarization of First Temple period inscriptions - Performance of existing algorithms and a new registration based scheme”, *Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012)*, 645-650, 2012.

[Shaus et al. 2013] **A. Shaus**, B. Sober, E. Turkel, E. Piasezky, “Improving binarization via sparse methods”, *Proceedings of the 16th International Graphonomics Society Conference (IGS 2013)*, 163-166, 2013.

[Shaus et al. 2016a] **A. Shaus**, B. Sober, S. Faigenbaum-Golovin, A. Mendel-Geberovich, E. Piasezky, E. Turkel, “Facsimile creation: Review of algorithmic approaches”, in: I. Finkelstein, C. Robin, T. Römer eds., *Alphabets, Texts and Artefacts in the Ancient Near East, Studies Presented to Benjamin Sass*, 474-488, 2016.

[Shaus et al. 2016b] **A. Shaus**, B. Sober, E. Turkel, E. Piasezky, “Beyond the ground truth: Alternative quality measures of document binarizations”, *Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016)*, 495-500, 2016.

- [Shaus et al. 2017a] **A. Shaus**, S. Faigenbaum-Golovin, B. Sober, E. Turkel, E. Piasezky, “Potential contrast - A new image quality measure”, Proceedings of the IS&T International Symposium on Electronic Imaging 2017, Image Quality and System Performance XIV Conference (IQSP 2017), 52-58, 2017.
- [Shaus et al. 2017b] **A. Shaus**, B. Sober, S. Faigenbaum-Golovin, A. Mendel-Geberovich, D. Levin, E. Piasezky, E. Turkel, “Statistical inference in archaeology: Are we confident?”, in: O. Lipschits, Y. Gadot, M. J. Adams eds., Rethinking Israel: Studies in the History and Archaeology of Ancient Israel in Honor of Israel Finkelstein, Eisenbrauns, Winona Lake, 389-401, 2017.
- [Shio 1989] A. Shio, “An automatic thresholding algorithm based on an illumination-independent contrast measure”, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 1989), 632-637, 1989.
- [Sivic and Zisserman 2003] J. Sivic, A. Zisserman, “Video Google: A text retrieval approach to object matching in videos”, Proceedings of the 9th International Conference on Computer Vision, 1470-1477, 2003.
- [Sober and Levin 2017] B. Sober, D. Levin, “Computer aided restoration of handwritten character strokes”, Computer-Aided Design 89, 12-24, 2017.
- [Sober et al. 2014] B. Sober, S. Faigenbaum, I. Beit-Arieh, I. Finkelstein, M. Moinester, E. Piasezky, **A. Shaus**, “Enhancement of ostraca reading: Three test cases of multispectral imaging”, Palestine Exploration Quarterly 146.3, 185-197, 2014.
- [Solem et al. 2006] J. E. Solem, N. C. Overgaard, A. Heyden, “Initialization techniques for segmentation with the Chan-Vese model”, Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006), 171-174, 2006.
- [Sreeraj and Idicula 2011] M. Sreeraj and S. M. Idicula, “A survey on writer identification schemes,” International Journal of Computer Applications, no. 26, vol. 2, 23-33, 2011.
- [Stathis et al. 2008a] P. Stathis, E. Kavallieratou, N. Papamarkos, “An evaluation survey of binarization algorithms on historical documents”, Proceedings of the 19th International Conference on Pattern Recognition (ICPR 2008), 1-4, 2008.
- [Stathis et al. 2008b] P. Stathis, E. Kavallieratou, N. Papamarkos, “An evaluation technique for binarization algorithms”, Journal of Universal Computer Science 14.18, 3011-3030, 2008.
- [Tahmasbi 2014] <http://www.mathworks.com/matlabcentral/fileexchange/38900-zernike-moments>
- [Tahmasbi et al. 2011] A. Tahmasbi, F. Saki, S. B. Shokouhi, “Classification of benign and malignant masses based on Zernike moments”, Computers in biology and medicine 41.8, 726-735, 2011.
- [Torczyner et al. 1938] H. Torczyner et al. Lachish I: The Lachish Letters. London, 1938.

- [Tree 2011] Tree model, R version 2.12.2. <http://www.r-project.org>, 2011.
- [Trier and Jain 1995] Ø. D. Trier, A. K. Jain, “Goal-directed evaluation of binarization methods”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17.12, 1191-1201, 1995.
- [Trier and Taxt 1995] Ø. D. Trier, T. Taxt, “Evaluation of binarization methods for document images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17, 312–315, 1995.
- [Trier et al. 1996] Ø. D. Trier, A. K. Jain, T. Taxt, “Feature extraction methods for character recognition - a survey”, *Pattern Recognition* 29.4, 641-662, 1996.
- [Ussishkin 1988] D. Ussishkin, “The date of the Judean shrine at Arad”, *Israel Exploration Journal* 38, 142-157, 1988.
- [Van der Walt et al. 2011] S. van der Walt, S. C. Colbert, G. Varoquaux, “The NumPy array: A structure for efficient numerical computation”, *Computing in Science & Engineering*, 13, 22-30, 2011, www.numpy.org, version 1.12.1.
- [Van der Walt et al. 2014] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu and the scikit-image contributors, “scikit-image: Image processing in Python”, *PeerJ* 2:e453, 2014, <http://scikit-image.org>, version 0.13.0.
- [Van der Zant et al. 2009] T. van der Zant, L. Schomaker, S. Zinger, H. van Schie, “Where are the search engines for handwritten documents?”, *Interdisciplinary Science Reviews* 34.2-3, 224-235, 2009.
- [Van Dongen and Enright 2012] S. Van Dongen, A. J. Enright, “Metric distances derived from cosine similarity and Pearson and Spearman correlations”, *arXiv:1208.3145*, 2012.
- [Van Oosten and Schomaker 2014] J.-P. van Oosten, L. Schomaker, “Separability versus prototypicality in handwritten word-image retrieval”, *Pattern Recognition* 47, 1031-1038, 2014.
- [Vese and Chan 2002] L. Vese, T. F. Chan, “A multiphase level set framework for image segmentation using the Mumford and Shah model”, *International Journal of Computer Vision* 50.3, 271-293, 2002.
- [Wang et al. 2002] J. Wang, C. Wu, Y. Q. Xu, H. Y. Shum, L. Ji, “Learning-based cursive handwriting synthesis”, *Proceedings of the 8th International Workshop on Frontiers of Handwriting Recognition (IWFHR 2002)*, 157-162, 2002.
- [Wang et al. 2010] X. F. Wang, D. F. Huang, H. Xu, “An efficient local Chan–Vese model for image segmentation”, *Pattern Recognition* 43, 603-618, 2010.
- [Welk et al. 2011] M. Welk, M. Breuß, O. Vogel, “Morphological amoebas are self-snakes”, *Journal of Mathematical Imaging and Vision* 39, 87-99, 2011.

[White and Rohrer 1983] M. White, G. D. Rohrer, “Image thresholding for optical character recognition and other applications requiring character image extraction”, IBM Journal of Research and Development 27.4, 400-411, 1983.

[Wolf et al. 2010] L. Wolf, N. Dershowitz, L. Potikha, T. German, R. Shweka, Y. Choueka, “Automatic palaeographic exploration of Genizah manuscripts”, In: F. Fischer, C. Fritze, G. Vogeler eds., *Codicology and Palaeography in the Digital Age 2*, 157-159, 2010.

[Xia et al. 2007] R. Xia, W. Liu, J. Zhao, L. Li, ‘An optimal initialization technique for improving the segmentation performance of Chan-Vese model”, Proceedings of the IEEE International Conference on Automation and Logistics, 411-415, 2007.

[Xu and Wang 2008] H. Xu, X. F. Wang, “Automated segmentation using a fast implementation of the Chan-Vese models”, Proceedings of the 4th International Conference on Intelligent Computing (ICIC 2008), Lecture Notes in Computer Science, Vol. 5227, 1135-1141, 2008.

[Zitová and Flusser 2003] B. Zitová, J. Flusser, “Image registration methods: A survey”, *Image and Vision Computing* 21, 977–1000, 2003.

תקציר

המחקר המתואר בתיזה עוסק בשיטות מתמטיות לניתוח אוסטרקונים, כתובות עבריות מימי בית שני (מאה 8-6 לפנה"ס). טקסטים אלו, אשר נכתבו בדיו על גבי שברי חרסים, נוצרו בממלכות ישראל ויהודה, והינם בין הממצאים הכתובים הבלעדיים בני התקופה ששרדו עד ימינו. בשל כך, האוסטרקונים חשובים למחקרים ארכיאולוגיים, היסטוריים, אפיגרפיים, פילולוגיים ובלשניים, ומהווים נדבך קריטי בחקר המקרא המודרני. זאת למרות תוכנם אשר עוסק לרוב בענייני דיומא כגון אספקת מזון, רישום מיסוי, והעברת פקודות צבאיות תמציתיות.

מומחים לכתבי יד עתיקים מבצעים חלק נכבד ממשימות הניתוח בשיטות ידניות. אלו גוזלות מאמצים וזמן רבים ולא אחת מובילות למסקנות שנויות במחלוקת, שמערבות תיעוד עם פרשנות. לעומת זאת, לדעתנו, הכתובות מימי בית ראשון, ששיני הזמן נגסו בהן ללא רחם, מהוות קרקע פורייה לפיתוח שיטות "חסינות רעש" בתחומים דוגמת רכישת ועיבוד תמונה, ראיה ממוחשבת, ולמידה חישובית. מטרת המחקר הנוכחי, אפוא, הינה **יצירת מכלול של כלים ממוחשבים לניתוח חומרים אפיגרפיים מימי בית ראשון**. במסגרת זאת, התיזה עוסקת בנושאים הבאים:

- הערכת איכות של פקסימיליות ידניות (פרק 2)
- הערכת איכות הצילומים של אוסטרקונים (פרק 3)
- יצירת בינאריזציות (תמונות שחור-לבן) של כתובות (פרק 4)
- שיפור בינאריזציות נתונות באמצעות מודלים "דלילים" (פרק 5)
- הערכת איכות של בינאריזציות (פרק 6)
- הערכת איכות של אותיות מובחנות בתוך בינאריזציות (פרק 7)
- הפרדת כותבים של כתובות (פרקים 8 ו-9; מוצעות שתי שיטות שונות להפרדה)
- סגמנטציה מהירה של תמונות (פרק 10; מוצע שיפור לאלגוריתם הקלאסי של Chan-Vese)
- חישוב אבטיפוס (prior) עבור אותיות (פרק 11)

כל השיטות עברו אישוש ניסיוני על מידע ארכיאולוגי וגררו שורה של פרסומים.

מוקדש לאחייני עידו גופר ולדודי
איליה שאוס, זכרונם לברכה.



הפקולטה למדעים מדויקים ע"ש ריימונד וברלי סאקלר

בית הספר למדעי המתמטיקה

המחלקה למתמטיקה שימושית

שיטות בראיה ממוחשבת ולמידה חישובית לצורך ניתוח כתובות מימי בית ראשון

חיבור לשם קבלת התואר דוקטור לפילוסופיה

ע"י

אריה שאוס

העבודה התבצעה בהנחייתו של פרופ' אלי טורקל

ובהנחיה משנית של פרופ' ישראל פינקלשטיין

הוגש לסנאט של אוניברסיטת תל אביב

דצמבר 2017